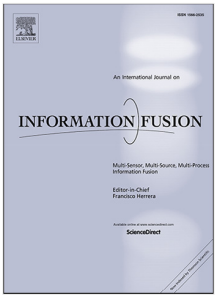
# Journal Pre-proof




A comprehensive survey of Vision-Language Models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets

Sufyan Danish, Abolghasem Sadeghi-Niaraki, Samee Ullah Khan, L. Minh Dang, Lilia Tightiz, Hyeonjoon Moon

PII: S1566-2535(25)00695-5
DOI: https://doi.org/10.1016/j.inffus.2025.103623
Reference: INFFUS 103623

To appear in: *Information Fusion*

Received date : 25 March 2025
Revised date : 4 August 2025
Accepted date : 7 August 2025

Please cite this article as: S. Danish, A. Sadeghi-Niaraki, S.U. Khan et al., A comprehensive survey of Vision-Language Models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets, *Information Fusion* (2025), doi: https://doi.org/10.1016/j.inffus.2025.103623.

# A Comprehensive Survey of Vision-Language Models: Pretrained Models, Fine-Tuning, Prompt Engineering, Adapters, and Benchmark Datasets

Sufyan Danish[a], Abolghasem Sadeghi-Niaraki[c], Samee Ullah Khan[b], L. Minh Dang[c,d,e], Lilia Tightiz[f] and Hyeonjoon Moon[a,*]

[a]Department of Computer Science and Engineering, Sejong University, Seoul, 05006, , South Korea
[b]Advanced Research and Innovation Center (ARIC), Khalifa University of Science and Technology, Abu Dhabi, 127788, Abu Dhabi, UAE
[c]The Institute of Research and Development, Duy Tan University, Da Nang, 550000, Viet Nam
[d]Faculty of Information Technology, Duy Tan University, Da Nang, 550000, Viet Nam
[e]Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Seoul, 05006, Republic of Korea
[f]School of Computing, Gachon University, 1342 Seongnamdaero, Seongnam-si, Gyeonggi-do,, 13120, Republic of Korea

ARTICLE INFO

*Keywords*:
Vision Language Model
Pre-trained Model
Prompt Engineering
Fine Tuning
Adapter
dataset
Multi-language Model Survey

ABSTRACT

Vision Language Models (VLMs) have significantly advanced multimodal tasks like image captioning, visual question answering, and multimodal retrieval. This survey presents a systematic review of 115 published papers from 2018 to 2025. It focuses key VLM components including fine-tuning strategies, prompt engineering techniques, pre-trained models, adapter modules, and benchmarking datasets. For each component, we present taxonomies and summarize comparative findings across standard VLM benchmarks. The survey emphasizes the role of lightweight, parameter-efficient adaptation methods in reducing computational overhead while maintaining strong task performance, particularly in real-world deployment contexts. It further examines the strengths and limitations of prompt-based learning, dataset-specific tuning strategies, and architectural trade-offs. Finally, the paper identifies open challenges in scalability, generalization, and bias, and explores emerging research directions including symbolic reasoning, multilingual adaptation, and energy-efficient VLM design. To our best knowledge, this is the first comprehensive survey to integrate these critical components into a single, cohesive survey paper, intended to serve as a foundational resource for researchers and practitioners striving to optimize VLMs for diverse real-world scenarios. The highlights of the review are available at GitHub directory.

## 1. Introduction

Artificial Intelligence (AI) has made significant developments in recent years, with impressive advancements in domains such as Natural Language Processing (NLP) and Computer Vision (CV). The integration of these two domains into Vision Language Model (VLMs) represents one of the most significant breakthroughs in enabling machines to process and generate multimodal information that involves both vision and language. These models aim to connect the gap between the visual world and textual understanding, allowing for a more refined and comprehensive approach to tasks that involve both modalities. The success of VLMs is mainly driven by the utilization of large scale pretrained models and SOTA deep learning methodologies, such as transformers which have contributed to considerable advancements in both accuracy and generalization in various tasks [1].

hmoon@sejong.ac.kr (H. Moon)
ORCID(s):

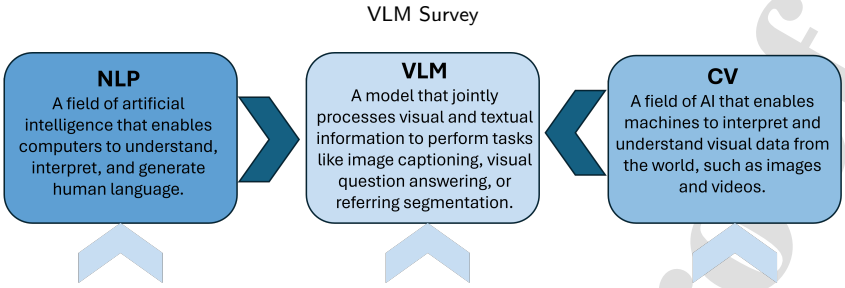



**Figure 1:** Conceptual model of the paper, where VLMs unify NLP and CV by jointly interpreting text and images to enable tasks such as image captioning, visual question answering, and image/video retrieval.

VLMs have a wide range of practical applications, including image captioning, Visual Question Answering (VQA), multimodal retrieval, visual reasoning, and other related domains [2, 3, 4]. To enhance the sophistication and effectiveness of these applications, it is essential to develop efficient techniques for fine-tuning and prompt engineering. These methods are crucial for developing pretrained models that respond more to particular tasks or fields, improving their efficiency while maintaining computational efficiency [5, 6]. However, despite the impressive capabilities of VLM, its deployment still faces challenges, especially in fine-tuning and task-specific adaptation. Fine-tuning large-scale models remains a significant challenge, particularly due to the substantial computational resources required and the difficulty in maintaining generalization across diverse domains [7]. Techniques such as prompt engineering, adapter-based methods, and the use of specialized datasets have emerged as effective approaches to address these challenges, enabling more precise refinement of model applications for specific use cases [6, 8].

The primary motivation for conducting this survey lies in the need for a comprehensive review of optimization techniques and their application to VLMs. While substantial progress has been made in enhancing the foundational architectures of multimodal models [9], this paper specifically focuses on the methods that enable these models to be effectively adapted and fine-tuned for real-world applications. In particular, we examine fine-tuning strategies, prompt engineering, adapter-based approaches, and the role of datasets as critical components in boosting the performance and versatility of VLMs. **Figure 1** provides a graphical representation of the VLM concept.

### 1.1. Background and Motivation

The field of VLMs has witnessed groundbreaking advancements in recent years, fueled by progress in deep learning and the growing availability of large-scale multimodal datasets. Early models, such as image-captioning systems, focused on aligning image features with textual descriptions [10]. However, more sophisticated architectures such as LLaVA [11] and PaliGemma 2 [12] have significantly transformed the landscape. These models leverage transformers and other advanced neural architectures to construct robust joint representations of vision and language, enabling them to perform complex tasks like zero-shot learning and cross-model retrieval with remarkable effectiveness.

Apart from the impressive success of these pre-trained models and their efficient application to particular tasks often demands substantial computational resources and expertise. Fine-tuning, which involves adjustment of parameters within a pre-trained model using smaller datasets to task specific, is highly effective but resource intensive [13]. Techniques such as Low-Rank Adaptation (LoRA) and bit-fitting have been developed to tackle this challenge, thus increasing the efficiency of fine-tuning [14]. Task-specific prompt engineering has demonstrated significant efficacy in optimizing VLMs, enhancing performance in tasks like VQA and image captioning through the strategic guidance of model outputs through meticulously designed [8]. Adapter-based methodologies further streamline fine-tuning by adjusting smaller model components or incorporating task specific modules, avoiding

extensive retraining. Datasets play a vital role in VLM development. The variety and caliber of datasets not only exert influence over the training progress but also affect the evaluation and benchmarking of the models themselves. Datasets such as MMVP-VLM [15], GEOBench-VLM [16], MammoVLM [17] and MMT-Bench [18] have played a crucial role in extending the limits of what VLM can accomplish. However, challenges related to dataset bias, annotation quality, and adaptation to domain-specific remain important challenges that need to be resolve in order to improve model robustness and fairness [19].

In summary, there is a pressing need for a thorough understanding of the techniques that enable VLMs to be effectively fine-tuned and optimized for specific tasks. This survey addresses that need by providing a comprehensive review of fine-tuning strategies, prompt engineering, adapter-based methods, pretrained VLM architectures, and datasets. It offers an in-depth analysis of State-of-The-Art (SOTA) methodologies while identifying existing challenges, critical gaps, and emerging research directions that hold the potential to drive further advancements in the field.

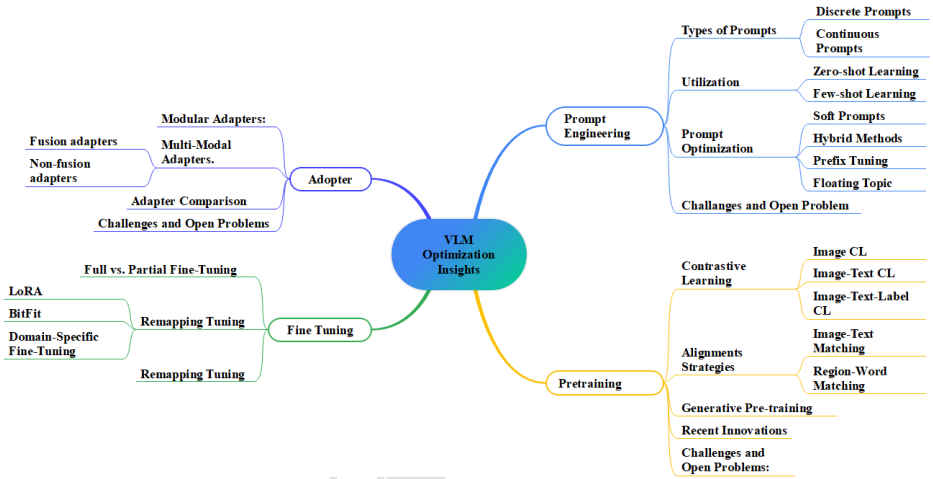



**Figure 2:** Conceptual model of the paper.

## 1.2. Scope and Objectives

The scope and objectives of this survey encompass a comprehensive exploration of optimization techniques applied in VLMs, focusing on their design, adaptation, and performance enhancement. The survey outlines on the following core concepts:

1. Various fine-tuning methodologies are examined that adapt pretrained VLMs to specific downstream tasks, including full fine-tuning, parameter-efficient tuning, and task-aware finetuning mechanisms.
2. This survey analyzes the role of prompt formulation in zero-shot and few-shot learning scenarios, reviewing manual, automatic, and soft prompting techniques that enable efficient adaptation without modifying the model's core parameters.
3. A lightweight approaches such as adapters, prefix tuning, and LoRA, are investigate which aim to minimize computational complexity and storage costs while preserving competitive performance across multiple tasks.

4. A structured overview is provide that influential pretrained VLMs, highlighting their architectural designs, training objectives, and core capabilities that serve as a foundation for downstream applications.
5. The review covers a range of large-scale and task-specific datasets used for pretraining and evaluation, including image-text, video-text, and instruction-tuned corpora. The influence of dataset quality, diversity, and alignment on model generalization is also discussed.

The survey has three main objectives:

1. To critically evaluate the foundation models employed in the optimization of VLMs and their influence on improving performance. It involves a systematic review of key optimization techniques such as fine-tuning, prompt engineering, and adapter-based learning. The survey aims to assess how these approaches contribute align with the large language model to enhancing the adaptability, task-specific performance, and overall generalization capabilities of VLMs across diverse multimodal benchmarks.
2. To analyze the challenges encountered in fine-tuning, prompt engineering, adapter methods, pretraining, and dataset utilization, and to provide potential solutions. A core focus of this survey is to identify the technical and practical limitations faced during the optimization of VLMs. It includes issues such as overfitting during fine-tuning, prompt sensitivity in few-shot scenarios, scalability constraints of adapters, and biases or insufficiency in training datasets. For each challenge, the survey discusses SOTA solutions and strategies proposed in recent literature to mitigate these barriers.
3. To suggest future research directions by exploring innovative methodologies to enhance the efficiency and generalization of VLMs. Building on the analysis of existing techniques and challenges, the survey highlights emerging trends and proposes avenues for future research. These include low-resource adaptation methods, dynamic prompt tuning, multimodal data augmentation, and the integration of VLMs with real-world deployment considerations such as edge efficiency and cross-lingual generalization.

This survey aims to provide a comprehensive review of these essential components, highlighting their contributions to the SOTA in VLM research. By systematically reviewing and synthesizing key advancements in these areas, the paper will present achievements and identify promising new directions for further exploration, thereby guiding future research and technological development in Vision-Language Models. **Figure 2** illustrates the conceptual model presented in the paper.

### 1.3. Main Contributions

The goal of this survey is to provide concrete contributions to understanding and optimizing Vision Language Models in several aspects, specifically fine-tuning prompt engineering, adapters, pretraining objectives, pre-trained models and datasets. To the best of our knowledge, no prior work has integrated all these aspects into a single, unified and comprehensive survey. The major contributions of this paper are as follows:

1. We provides a comprehensive review of fine-tuning techniques within the context of VLMs. It delves into the associated computational and temporal challenges, evaluating strategies to enhance efficiency and effectiveness in fine-tuning processes. This analysis aims to identify best practices for optimizing performance while minimizing resource demands.
2. We investigate the emerging area of prompt engineering a crucial strategy for leveraging pretrained models in novel applications with minimal additional training. The survey evaluates the effectiveness of prompt engineering across various scenarios, presenting a systematic framework for its application and highlighting innovative use cases in VLMs.

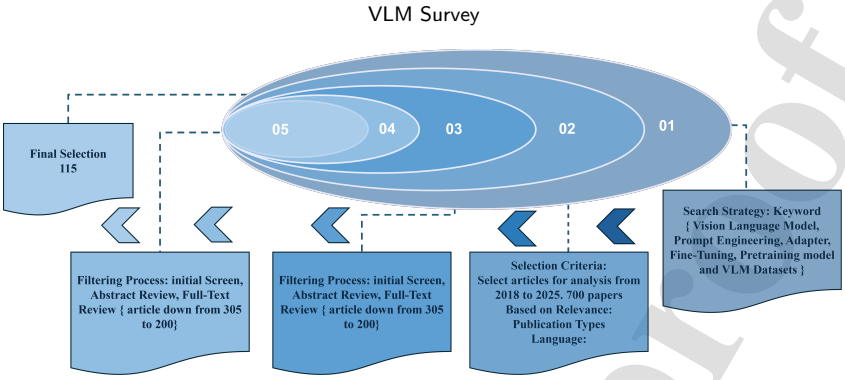



**Figure 3:** We selected papers based on their relevance to vision-language models, focusing on key areas such as pretrained architectures, fine-tuning, prompt engineering, adapters, and benchmarks. Priority was given to technically significant and widely cited works from top-tier conferences and journals.

3. The use of adapters is assessed as a flexible solution for customizing pretrained models to novel tasks. This study systematically reviews diverse adapter variants, discusses their integration approaches and evaluates their influence on the performance of VLMs efficacy, contributing to a deeper understanding of adapter-based model customization.
4. Examines the innovations and ongoing development of pretrained models in the field of Vision-Language Models. It covers how these models are developed, continually updated and optimized providing insights into their crucial role and potential for future enhancements.
5. We assess the role of multimodal datasets in the development and performance of VLMs, delving into their crucial characteristics, inherent challenges, and prospective pathways for dataset refinement including mitigating annotation reliance and augmenting diversity.

### 1.4. Conceptual Overview

To explain the fundamental concepts covered in this survey paper, we present a conceptual overview of the major components involved in optimizing VLMs.

**Fine-Tuning:** Fine-tuning constitutes adapting a pretrained VLM to specialized tasks through additional training on limited, application-specific datasets. This approach adjusts the model parameters to optimize performance for target applications while utilizing generalized knowledge acquired during pre-training.

**Pre-trained Models:** Pre-trained VLMs are foundational architectures trained on extensive multimodal datasets (e.g., image-text pairs) to develop generalized cross-modal representations. These models establish versatile computational frameworks that allow efficient knowledge transfer for specialized applications without demanding extensive annotated data.

**Prompt Engineering:** Prompt engineering involves the strategic formulation of input structures and designing that incorporates textual queries, contextual descriptions or multimodal instructions to optimize Vision-Language Model performance for specific objectives. This technique utilizes model pre-acquired knowledge by carefully crafting inputs to generate accurate and task relevant outputs.

**Adapters:** Adapters represent modular, task-specific architectural components integrated into pre-trained VLMs. These components facilitate efficient adaptation to novel applications through supplementary trainable layers that focus on target objectives while preserving the model's foundational competencies, all without comprehensive modification of core parameters.

**VLM Datasets:** VLM datasets comprise paired multimodal data that include image caption or video-text datasets that provide aligned visual and textual information. These datasets are crucial for both

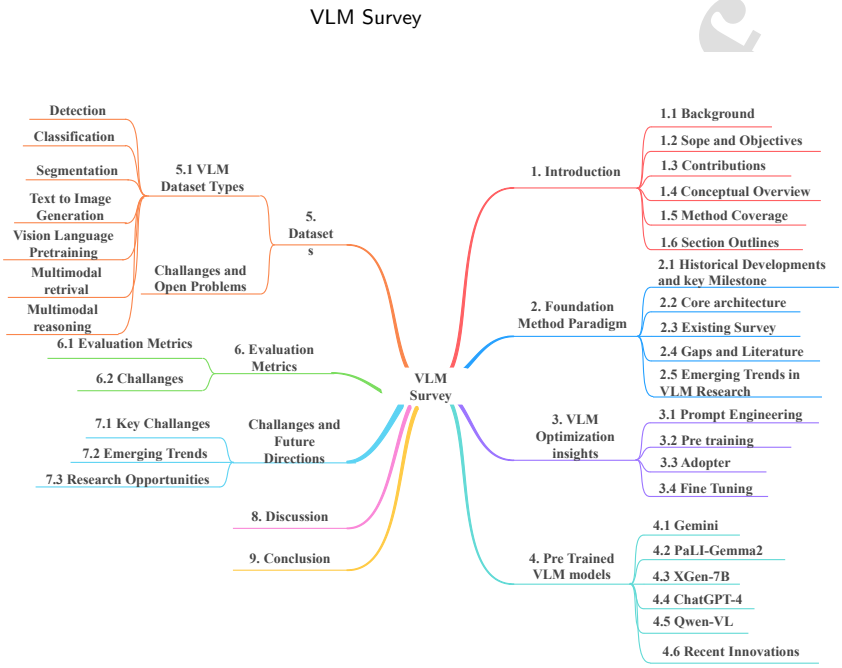



**Figure 4:** Schematic overview of the survey design highlighting optimization strategies in VLM. The figure categorizes key components covered in the paper, including fine-tuning techniques, prompt engineering, adapter and pretraining model. Each component is critically analyzed to provide comprehensive insights into current trends, challenges, and future research directions in VLM optimization

pretraining and fine tuning Vision Language Models, enabling the models to capture the complex relationships between visual and linguistic modalities across a diverse range of domains.

## 1.5. Method Coverage

The approach used in the selection of papers for this survey encompassed a series of successive phases done systematically in order to the utilization of high quality and relevance research articles mainly from leading scholarly publications. Our procedure was designed to provide a thorough and unbiased literature review on Vision-Language Models with a particular focus on fine-tuning methods, pertaining models, prompt engineering, adapter techniques and VLM benchmark dataset as demonstrated in Figure 5. The selection process is detailed below:

### *1.5.1. Search Strategy*

We conducted an extensive search in several well-established and recognized electronic databases. The main terminologies employed in the search included Vision Language Model, Prompt Engineering, Adapter, Fine-Tuning, Pretraining Model and VLM Datasets. These terms were meticulously chosen according to the relevance to main features of the survey and their relation to recent developments in the domain. The following databases were prioritized for the search:

- Science Direct
- Springer
- Scopus

- Web of Science

Alongside these academic databases, we also conducted supplementary research through Google Scholar to obtain papers that may not have been acquired in the above-mentioned databases. Given the rapid evolution inherent in VLM research, we incorporated preprints from arXiv, contingent upon their acceptance for publication in peer-reviewed conferences or journals. The **Figure 5** shows a graphical representation of the academic databases applied in the survey paper.

#### *1.5.2. Selection Criteria*

advancements

- Relevance: Many papers directly related to VLM fine tuning, prompt engineering, adapter methods pretrained models and benchmark datasets were prioritized. Such studies that were merely indirectly related or exhibited a limited scope were systematically excluded.
- Publication Type: We considered peer-reviewed journal articles, conference proceedings and well-recognized preprints from arXiv.
- Language: With the language, only English language publications were considered in the process of selection due to the linguistic predominance within the research society and the emphasis on global trends.

#### *1.5.3. Filtering Process*

The filtering process was conducted in several stages:

- Initial Screening: The first step was to scan through the titles of the papers, abstract and keywords for the presence of the specified keywords. At this stage, repeated entries and articles that did not meet the relevance criteria were identified and eliminated. This step narrowed it down from 600 articles to just 305 articles at the first selection criterion alone.
- Abstract Review: The rest of the papers were subjected to an abstract level analysis and critique accordingly. Studies that were found irrelevant or presented with minor detail were omitted at this level. Following this review, the number of articles was further reduced to 200.
- Full-Text Review: The next step was to perform a more screening and systematic analysis of the full texts of the remaining articles. The main objectives of this review were to establish the extent and depth of the methodologies described, as well as the relevance and significance of the findings. Some overlaid studies that did not offer significant information regarding fine tuning, prompt engineering and adapter methods were not included.

#### *1.5.4. Quality Assurance*

In order to ascertain the quality of the selected articles, the review process incorporated the subsequent quality assessments:

- Peer Review: Every paper included in the final set was either peer reviewed or published in reputed journals or conferences.
- Reputation of Conference/Journal: To ensure pertinent and highly impactful research, only international conferences and the best-tier journals were taken into account.
- Reproducibility and Transparency: Great importance was accorded to the studies that offered enough information on how the experiments were conducted, datasets used when assessing the models and the structures of the models to help develop other models.

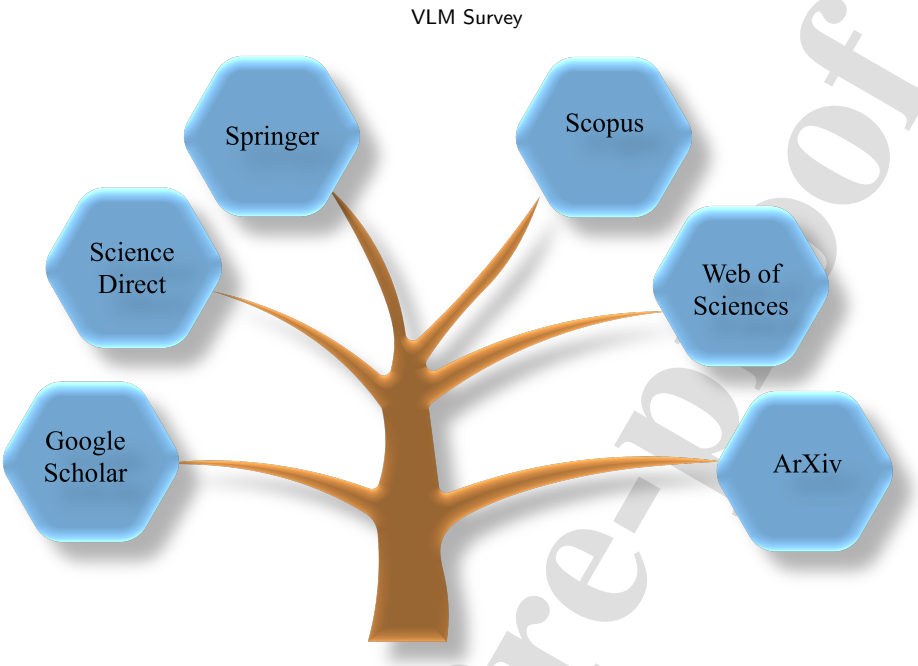



**Figure 5:** Illustration of the search engine and databank system used for systematic paper selection. This process involves querying multiple academic databases using predefined keywords related to VLMs, followed by filtering based on relevance, publication venue, and methodological contribution. The resulting databank serves as the foundation for building the comprehensive literature review presented in this survey.

#### *1.5.5. Final Selection*

Finally, the total number of papers selected in the filtering process was 115 papers for the survey. These papers were deemed as the newest, most focused, diverse, and impactful work with regard to VLM fine-tuning, prompt engineering, adapter methods, pre-trained models, and datasets. Thus, the selected studies were evaluated in terms of their methodological approaches, contributions and perspectives to present the state of the art in the domain. By virtue of this systematic and multistage approach, the studies on which this survey is based were all relevant and of high quality as described in **Figure 3**.

### 1.6. Section Outlines

The structure of this survey is as follows: Section 2 discusses Foundation Model Paradigms which include the development trajectory of the VLM and existing survey. Section 3 delves into fine-tuning, prompt engineering, pre-trained models, and adapters, highlighting their methodologies, applications challenges and open research probelem. Section 4 outlines the latest VLM pretrained model while Section 5 discusses datasets and its various types. Section 6 presents evaluation matrices of Vision-Language Models and benchmarks, emphasizing challenges, Emerging trends and Research Opportunities. Section 7 explores open challenges and future research directions. Section 8 outlines discussion and while the Section 9 concludes of the survey paper. The whole paper contained is explained in the **Figure 4**.

## 2. Foundation Model Paradigms

Vision-Language Models [20] are the next generation of machine learning models developed to understand and simultaneously process visual and textual information. These models aim to address the gap between the visual understanding capabilities of CV and the linguistic abilities of NLP. Through the establishment of unified representations for both images and textual content, Vision Language Models enable seamless multimodal understanding, thereby enhancing the ability of machines to interpret and produce information in tasks that encompass both modalities [21].

**Significance of Vision-Language Models** Vision Language Models represent a transformative step in AI by tackling intricate tasks that require a deep understanding of the correlations between visual and textual information. Their significance is highlighted in tasks such as image captioning [22], which involves the automated generation of precise and semantically rich textual descriptions corresponding to images. Similarly, VQA [23] which derives contextually grounded responses to textual queries through object identification and relational reasoning within visual scenes. Another essential application is multimodal retrieval, that substantially enhances information accessibility by allowing cross-modal searches (e.g., image retrieval via textual queries or conversely) across heterogeneous data formats [24], Furthermore, visual reasoning [25] extracts logical interpretations from visual scenarios such as entity relationship analysis or outcome prediction, thereby supporting decision-making capabilities in complex visual contexts. Beyond these foundational functions, these models demonstrate indispensable utility in specialized domains: In healthcare diagnostics [26], they enhance precision through interpretative analysis of medical imagery synergized with clinical narratives. Similarly, autonomous systems employ them for environmental navigation and contextually informed determinations through multimodal data integration. a Moreover, geospatial intelligence applications facilitate examination of satellite imagery contextualized with spatiotemporal metadata for land-use analytics and terrain assessment [27]. Consequently, given their capacity for synergistic multimodal fusion, Vision-Language Models have emerged as a foundational substrate for addressing challenges in fields requiring integrated multimodal intelligence.

### 2.1. Historical Development and Key Milestones

The trajectory of Vision-Language Models demonstrates accelerated development, transitioning from rudimentary architectures to complex frameworks. Initial multimodal frameworks such as proposed by J.Mao at.el, [28], are simple models linking images with text, typically employing manual-crafted features, which demonstrated limited capacity for capturing complex cross-modal relationships. A critical milestone was the introduction of frameworks like Visual Semantic Embedding (VSE) [29], which employed neural networks to project images and texts into a unified representational space, substantially improving cross-modal retrieval efficacy. These developments established the foundation for end-to-end trainable architectures, including Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models that improved capabilities in image captioning and visual question answering [30]. The true innovation in Vision-Language Models came with the development of large-scale pre-trained models, specifically CLIP (Contrastive Language-Image Pretraining) [5], which used a contrastive learning approach, and achieved remarkable zero-shot generalization across diverse downstream applications using multimodal datasets. Its ability for task transfer without application-specific refinement represented a substantial conceptual progression. Subsequent innovations like Flamingo [9] further refined VLM capabilities through parameter efficient designs, achieving competitive performance across diverse domains via modular integration of pretrained representations. These methodological advances, amplified by models including Vary [31], BRAVE [32], CLIP-FSSC [33], and BLIP (Bootstrapping Language-Image Pretraining) [34], have collectively enhanced functional capacities across multimodal tasks, establishing SOTA VLMs as prevailing methodological paradigms within vision language models.

**Impact of Large-Scale Pre-Trained Models** The rise of large-scale pre-trained models has fundamentally transformed the landscape of Vision-Language Models. These models, trained on billions of image-text pairs, exhibit unprecedented capabilities in zero-shot and few-shot learning, eliminating the need for extensive labeled data for new tasks. Key contributions of pre-trained Vision-Language Models include:

- Scalability: The ability to handle a wide range of downstream tasks with minimal task-specific adjustments [35, 36].
- Generalization: Strong zero-shot [37] and few-shot [38] capabilities allow models to perform well on unseen datasets and tasks.
- Efficiency: Pre-training reduces the dependency on task-specific data and computational resources, paving the way for scalable deployment across domains [39, 40].

The developmental trajectory of Vision-Language Models highlights a shift from narrow, task-specific systems to generalized, scalable architectures capable of addressing complex multimodal challenges. These advancements underline the importance of foundational models in modern AI research and their growing impact across diverse application areas.

## 2.2. Core Architectures

The architectures driving Vision-Language Models have evolved to incorporate several innovations, most notably the Transformer architecture, which has revolutionized both natural language processing and computer vision. The Transformer model [41] uses self-attention mechanisms to allow models to weigh the importance of different parts of input data, whether visual or textual, and capture long-range dependencies. This architecture has become the foundation for many state-of-the-art Vision-Language Models, providing significant improvements in both accuracy and computational efficiency.

One of the most important advancements has been the development of Vision Transformers (ViTs) [42], which adapt the transformer model originally designed for NLP tasks to vision tasks. ViTs have proven highly effective, particularly in scenarios where large datasets are available, and have surpassed traditional Convolutional Neural Networks (CNNs) in several vision benchmarks. In addition to transformers, many Vision-Language Models utilize encoder-decoder architectures, which have proven effective for tasks like image captioning and VQA. In these models, the encoder processes the visual or textual input, and the decoder generates the output (e.g., a caption or an answer to a question). This architecture is essential for tasks where one modality (e.g., an image) needs to be translated into another modality (e.g., text) [5].

Furthermore, hybrid architectures have emerged as another important development. These models combine CNNs (for processing visual data) with transformers (for textual data), allowing each architecture to leverage its strengths. This hybridization facilitates better multimodal learning and increases the robustness of models across various tasks [43]. Critical research areas in VLM architectures also include cross-modal alignment and multimodal fusion. Cross-modal alignment ensures that the visual and textual representations exist in a shared space, facilitating easier integration of both modalities [44]. Meanwhile, multimodal fusion methods aim to combine the features from both modalities into a unified representation, improving performance on tasks like image captioning and VQA [45, 46]. The role of representation learning in Vision-Language Models has also been pivotal. By learning a shared representation for both vision and language, these models can generalize better across different tasks, increasing their robustness and enabling them to excel in complex multimodal applications [47, 48]. The **Figure 6** represent the general hierarchic of the VLM.

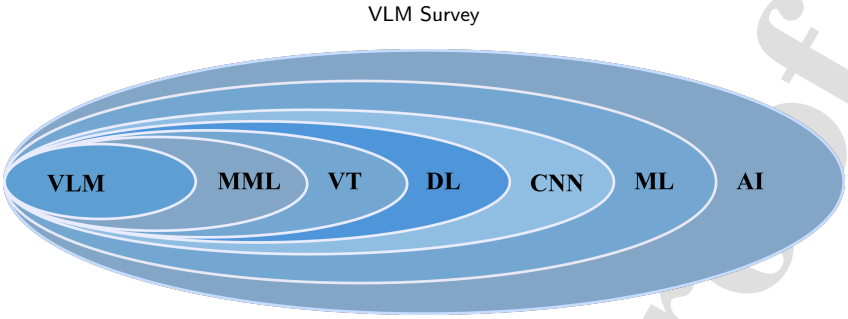



**Figure 6:** General hierarchical framework of vision-language models

## 2.3. Comparison with Existing Survey

The landscape of Vision-Language Models has been extensively explored in recent literature, with numerous surveys focusing on various aspects of model development, adaptation, and application. However, existing surveys often address specific components of Vision-Language Models in isolation, providing valuable but fragmented insights. Several surveys excel in detailing pre-trained models, fine-tuning techniques, and prompt engineering, which are central to the methodology of our survey, A VLM Survey: Fine-Tuning, Pre-trained Models, Prompt Engineering, Adapter, and Benchmarking Data. Yet, many of these works overlook the critical integration of these techniques, which is essential for optimizing model performance and ensuring scalability in real-world applications.

For instance, A Survey [49], emphasize pre-trained models and pre-training techniques. While these works contribute valuable insights into model architecture, they often provide limited coverage of fine-tuning and adapter techniques, focusing more on theoretical aspects than on practical applications. Similarly a Survey [50], present important real-world insights into vision-language interaction in autonomous driving but lack coverage on fine-tuning, prompt engineering, and adapter methods, which are essential for scalable model deployment. Surveys such as Zhou et al. [51] focus on geospatial applications of Vision-Language Models, highlighting the use of pre-trained models in this domain. However, the focus on this narrow application limits the survey's scope, as it does not address fine-tuning, prompt engineering, or adapter techniques, which are crucial for broader model adaptability.

| Paper/Year | Strengths | Limitations/Weaknesses |
|---|---|---|
| 2022 [52] | • In-depth review of VLM tasks and representation learning.<br>• Covers large models and their uses. | • Lacks detailed coverage of fine-tuning and adapter techniques.<br>• Limited focus on datasets. |
| 2022 [49] | • Thorough discussion of pre-trained models.<br>• Covers architectures and pre-training techniques. | • Limited coverage of fine-tuning or adapter techniques.<br>• Focus on pre-trained models may not be as practical for real-world scenarios. |
| 2022 [58] | • Provides insights into emerging multimodal research trends.<br>• Focus on integrating vision and language. | • Does not focus on fine-tuning or pre-training techniques.<br>• Limited discussion of pre-trained models. |
| 2023 [54] | • Detailed coverage of prompt engineering techniques.<br>• Explores different types of prompts used for Vision-Language Models. | • Does not cover fine-tuning or adapter-based methods.<br>• Limited practical applications in real-world settings. |

| Paper/Year | Strengths | Limitations/Weaknesses |
|---|---|---|
| 2024 [55] | • Highlights trust-based approaches for Vision-Language Models.<br>• Provides insights into vulnerabilities of Vision-Language Models. | • Does not focus on fine-tuning or adapter techniques.<br>• Limited discussion on pre-training techniques. |
| 2024 [56] | • Thorough analysis of OOD detection in Vision-Language Models.<br>• Addresses real-world deployment challenges. | • Does not focus on pre-training techniques, fine-tuning, or adapter methods.<br>• Narrow focus on one task (OOD detection). |
| 2024 [21] | • Comprehensive overview of Vision-Language Models.<br>• Discusses pre-trained models and pre-training techniques. | • Limited discussion on fine-tuning techniques.<br>• Does not cover adapter-based methods. |
| 2024 [57] | • Focus on low-shot learning and pre-trained models.<br>• Strong theoretical background using representer theorem. | • Does not address fine-tuning techniques.<br>• Limited exploration of adapter-based methods. |
| 2024 [51] | • Focus on geospatial applications of Vision-Language Models.<br>• Pre-trained models used in geospatial contexts. | • Does not discuss fine-tuning or adapter techniques.<br>• Narrow focus on geospatial applications. |
| 2024 [50] | • Focus on autonomous driving and vision-language interaction.<br>• Real-world problem-solving focus. | • Narrow domain focus.<br>• Does not focus on fine-tuning, prompt engineering, or adapter methods. |
| 2024 [53] | • Comprehensive overview of pre-trained models.<br>• Covers different pre-training techniques. | • Does not discuss fine-tuning or adapter techniques in depth.<br>• Limited application-based examples. |
| 2024 [7] | • Focus on vision tasks like image captioning and VQA.<br>• Describes the use of pre-trained models in vision tasks. | • Limited discussion on datasets and evaluation metrics.<br>• Does not address prompt engineering or adapter methods. |
| 2024 [62] | • Focus on efficient fine-tuning methods.<br>• Covers adapter-based tuning and prompt-based fine-tuning. | • Limited discussion of pre-training techniques.<br>• Does not fully address datasets or model evaluation. |
| 2025 [59] | • Detailed coverage of efficiency techniques suitable for deployment in edge and resource-constrained devices.<br>• Discusses performance-memory trade-offs. | • Lacks comprehensive coverage of prompt engineering, adapters, and dataset benchmarking.<br>• Narrow focus on edge devices. |
| 2025 [60] | • Extensive insights into synthetic data generation and integration with VLMs, strong emphasis on dataset construction and benchmarking.<br>• First comprehensive survey on the intersection of VLMs and synthetic data. | • Does not explore prompt engineering and adapter methodologies.<br>• Limited focus primarily to synthetic data generation. |
| 2025 [61] | • Detailed coverage of VQA tasks, robust evaluation metrics and benchmarking focused specifically on VQA tasks. | • Limited to visual question answering context, lacks comprehensive coverage of prompt engineering and adapter methods. |
| Our Survey | • Comprehensive coverage of the latest advancements in VLM methodologies, including: | • Fine-tuning technique |

**Table 1**
Technique-based comparison of the VLMs composed of various attributes including adapter, dataset, fine-tuning, pre-trained.

| Year/Paper | Category | FT | PE | Pre-TT | Adapter | Pre-TM | Dataset |
|---|---|---|---|---|---|---|---|
| 2022 [52] | General VL Tasks and Methods | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ |
| 2022 [53] | VLM pre-Training | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ |
| 2023 [54] | Prompt Engineering Techniques | ✗ | ✓ | ✓ | ✗ | ✓ | ✓ |
| 2024 [3] | Medical VLM | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ |
| 2024 [7] | VLPM Architecture | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ |
| 2024 [21] | Methodologies and Future Directions | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ |
| 2024 [50] | Autonomous Driving | ✗ | ✗ | ✓ | ✓ | ✓ | ✓ |
| 2024 [51] | Geospatial VLM | ✗ | ✗ | ✓ | ✗ | ✓ | ✓ |
| 2024 [55] | Trustworthiness of Vision-Language Models | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ |
| 2024 [56] | OOD Detection | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ |
| 2024 [57] | Low-shot VLM Adaptation | ✓ | ✗ | ✗ | ✗ | ✓ | ✓ |
| 2024 [58] | Emerging Trends in VL Research | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ |
| 2025 [59] | Edge Devices | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ |
| 2025 [60] | Synthetic Data & VLMs | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ |
| 2025 [61] | VQA | ✓ | ✗ | ✓ | ✗ | ✓ | ✓ |
| our survey paper | Comprehensive Overview | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Likewise, Li et al. [52], provide an in-depth review of VLM tasks and representation learning, but do not thoroughly explore fine-tuning or adapter methods, and their discussion on datasets is also limited. A survey presented by Vatsa et al. [55] concentrate predominantly on VLM trustworthiness, yielding valuable vulnerability insights while omitting consideration of fine-tuning methodologies, adapter techniques, and providing insufficient examination of pre-training paradigms. Similarly, Miyai et al. [56] delivers rigorous out-of-distribution (OOD) detection analysis which is critical for real-world robustness. yet neglect fine-tuning procedures, pre-training strategies, and adapter mechanisms, leaving significant gaps in understanding how Vision Language Models can be adapted for diverse tasks.

Similarly, Ghosh et al. [21] furnish an extensive pre-trained model taxonomy but afford limited analysis of fine-tuning approaches and do not cover adapter-based methodologies. Ding et al. [57] establish a robust theoretical foundation for low-shot learning within pre-trained frameworks, notwithstanding their superficial exploration of fine-tuning and adapter techniques, hindering their practical applicability in real-world VLM adaptation.

Conversely, Xing et al. [62] emphasize parameter-efficient adaptation (encompassing adapter-based and prompt-tuning methods) to enhance training efficiency and resource optimization. Notwithstanding these contributions, their work insufficiently addresses pre-training fundamentals and insufficiently covers dataset and evaluation protocols, which are essential considerations for effective VLM deployment.

Several recent surveys have advanced the field of Vision-Language Models. Shinde et al. [59] provided valuable insights into efficiency techniques tailored for edge device deployment. However, their work does not

thoroughly explore adapter-based methods or prompt engineering. Similarly, Mohammadkhani et al. [60] focused on synthetic data generation and its application in VLMs, offering useful perspectives on dataset creation and benchmarking, but they overlooked adapters and prompt engineering techniques. Meanwhile, Huynh et al. [61] concentrated solely on Visual Question Answering (VQA). Collectively, these surveys remain limited in scope, lacking coverage of broader methodological components such as prompt design and adapter integration.

In contrast, our survey aims to address these gaps by presenting the first unified framework that systematically integrates the full VLM pipeline, covering pre-training strategies, fine-tuning techniques, prompt engineering, adapter-based tuning (including LoRA, BitFit, Houlsby, and Compacter), and empirical benchmarking. Beyond simply listing techniques, we introduce structured taxonomies and evaluation tables that clarify the applicability, advantages, and limitations of various adapter types and prompt strategies. We also provide a detailed dataset audit that includes multilingual coverage, modality types, annotation bias, and domain generalization—areas largely neglected in prior works. Importantly, our discussion covers multiple application domains including medical imaging, autonomous driving, document understanding, robotics, and general-purpose vision-language tasks, thereby demonstrating the cross-domain applicability of our framework. We further highlight practical deployment considerations such as parameter-efficient tuning, scalability, latency, and modularity, which are critical for industrial adoption but often overlooked in prior surveys. This comprehensive approach makes our survey the first to integrate the full life cycle of Vision-Language Models, from pre-trained models and datasets to fine-tuning, task-specific prompts, and scalable adapters. Ultimately, our survey sets a new benchmark in the field, enabling researchers and practitioners to optimize Vision-Language Models for both academic research and industrial applications. **Table 1** outlines the domains and optimization techniques targeted by each respective survey, while **Table 2** illustrates the strengths and weaknesses of the existing survey papers.

## 2.4. Gaps in the Literature

Despite the wealth of insights provided by existing surveys and research in the Vision-Language Model domain, several critical gaps remain in the literature. These gaps represent promising areas for future research, particularly in the following aspects:

### *2.4.1. Fine-Tuning Methods*

Although significant progress has been made in fine-tuning (FT) pre-trained models for vision-language tasks, computationally efficient fine-tuning techniques remain unexplored. Current approaches, especially those involving large-scale models, often demand extensive computational resources, making them less accessible for researchers with limited infrastructure. The development of low-resource fine-tuning strategies and transfer learning methods is particularly scarce, representing an important avenue for further investigation [63]. Future work should focus on minimizing the computational cost while maintaining the performance of Vision-Language Models across various applications.

### *2.4.2. Pre-trained Models*

Pre-trained models such as CLIP [64], Flamingo [9], and ViLBERT [65] have become foundational for many vision-language applications. However, there is limited exploration of optimization strategies tailored to specific tasks. Although these models are pre-trained on massive datasets, their adaptability to diverse task-specific requirements remains insufficiently addressed. Additionally, domain-specific pre-training, which could enhance the models' performance on specialized tasks, has not been extensively studied [66]. Further research is needed to investigate the potential of specialized pre-training techniques for domain-specific applications, enhancing their generalization ability.

### *2.4.3. Prompt Engineering*

Prompt engineering (PE) has gained significant attention in recent literature, yet many surveys have yet to examine how different prompt design strategies impact model performance across a wide range of domains. While general and task-specific prompts have been discussed, there remains a gap in understanding how these prompts can be fine-tuned for particular use cases. The lack of in-depth studies on prompt adaptation for specific domains presents a promising direction for future research [67]. Studies should explore task-specific prompt design and its influence on fine-tuning Vision-Language Models for optimal performance.

#### *2.4.4. Adapters*

Adapter-based methods have emerged as a promising solution for efficient model adaptation without requiring full fine-tuning. However, there is still a need for a more thorough understanding of the limitations of these methods, particularly in scaling them to larger models. Research into identifying the most effective adapter architectures and their impact on task performance remains limited. Further exploration into how adapters can be utilized for cross-task generalization and large-scale model deployment is essential [68].

#### *2.4.5. VLM Datasets*

The development of high-quality, multimodal datasets continues to be a significant challenge. While datasets such as MS COCO [69] and Visual Genome [70] have been widely used, they remain limited in their diversity and scope. These datasets lack sufficient variety in content and real-world contexts, which affects the generalizability of Vision-Language Models. There is a pressing need for more diverse, multilingual, and cross-domain datasets that can foster better generalization and support real-world applications of Vision-Language Models [71]. The creation of richer, more varied datasets is crucial for advancing the field.

### 2.5. Emerging Trends in VLM Research

Recent trends in VLM research reveal several promising research trajectories. Notably, the integration of cross-modal transformer architectures and proliferation of multitask learning methodologies are attracting considerable attention. These frameworks synergistically integrate heterogeneous learning paradigms to augment model efficacy across diverse applications. Furthermore, advancements in self-supervised learning facilitate enhanced vision-language alignment by enabling VLMs to acquire representations with minimal dependence on annotated data, thereby improving both scalability and operational efficiency. Concurrently, sophisticated multimodal fusion techniques continue to progress, significantly refining intermodal interactions to elevate overall performance in complex reasoning tasks. As the discipline progresses, hybrid architectures integrating vision language learning paradigms are expected to assume increasingly pivotal roles. An emerging area of interest is model efficiency, specifically reducing computational overhead during fine-tuning, training, and inference. Approaches such as sparse transformers, distillation, and knowledge transfer are being explored to improve the accessibility and scalability of Vision Language Models. These approaches aim to optimize large scale VLMs for resource-constrained environments while expanding practical deployment potential across domains [72].

## 3. VLM Optimization Insight

As Vision-Language Models continue to expand in both size and complexity, their systematic optimization emerges as a critical demand for enhancing efficacy across diverse multimodal applications. Strategic optimization methodologies not only enhance models capacity for cross-domain generalization and specialization but also ensure computationally efficient deployment in practical application [73]. Core techniques, including prompt engineering, pre-training paradigms, adapter mechanisms, and fine-tuning methods, constitute foundational elements for maximizing VLM performance. Subsequent sections present a comprehensive analysis of principal optimization strategies, specifically prompt engineering, adapter, pre-training methodologies, and fine-tuning approaches. Each technique contributes substantially to advancing task-specific proficiency, computational resource efficiency, and scalable implementation frameworks.

### 3.1. Prompt Engineering

Prompt engineering has emerged as a critical optimization approach for Vision Language Models, enabling the model ability to perform complex multimodal tasks. By strategically employing natural language instructions, prompts direct VLMs to perform diverse operations with minimal task-specific training requirements. This approach is particularly valuable in scenarios where labeled data is limited, allowing Vision-Language Models to generalize across a range of domains. Effective prompt engineering not only optimizes model performance but also enhances computational efficiency and task adaptability [74]. This section explores the various types of prompts, applications in zero-shot [75] and few-shot learning [76], and optimization techniques used to improve prompt-based performance. **Figure 7** presents the research development of the prompt engineering from 2021 to 2025.

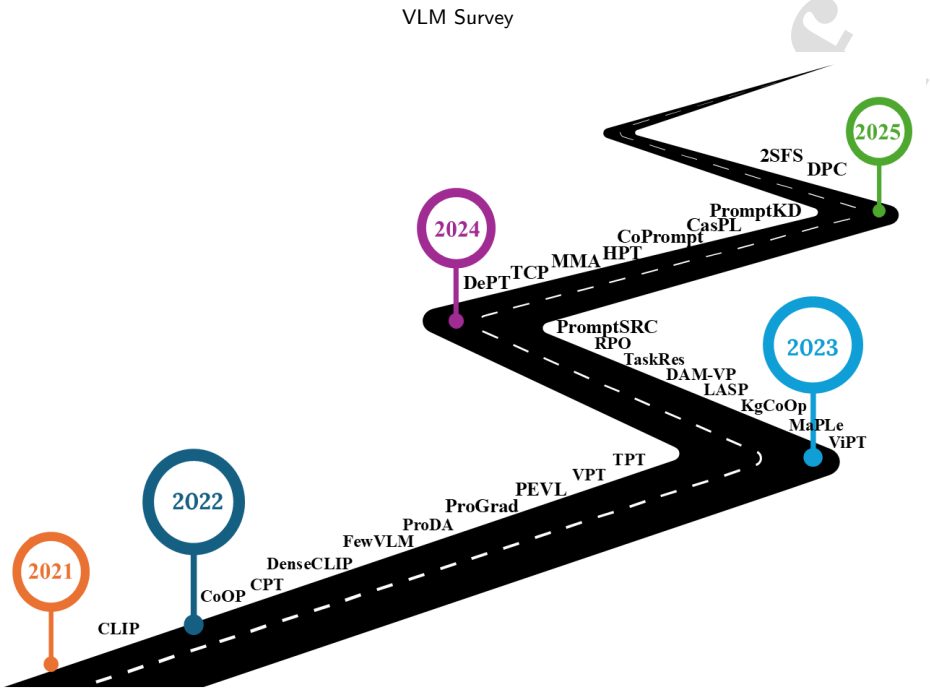



**Figure 7:** Research progress in prompt engineering techniques from 2021 to 2025.

### *3.1.1. Types of Prompts:*

Prompts in Vision Language Models can be classified into two primary types: discrete prompts and continuous prompts. Each type serves distinct purposes depending on the nature of the task and the level of model flexibility required [77].

- **Discrete Prompts:** Discrete prompts as also called Hard Prompts [78] are pre-defined, human-readable text inputs used to instruct the model in how to approach a given task. These prompts are straightforward to implement and interpret, making them a popular choice for tasks with clear, well-defined objectives. For example, a typical discrete prompt might be phrased as, "What is the object in the image?" [5]. While effective for tasks that require basic recognition or classification, discrete prompts can be less flexible in handling more complex, context-sensitive tasks, limiting their utility in certain multimodal applications [79]. Several recent studies have explored the application and impact of discrete prompts in the VLM domain. Notable work includes [78] proposed knowledge-aware prompts that incorporate semantic class descriptions as discrete inputs while [80, 81] conducted a comprehensive review of prompt engineering methods, highlighting the role of discrete prompts in VLMs. [74] explored black-box optimizers to generate discrete, human-readable prompts. Additional notable contributions include the works of [82], [83], [84], [85], [86] and [87], each offering unique perspectives on the design, optimization, and interpretability of discrete prompts in VLMs.
- **Continuous Prompts:**Continuous prompts [88] employ trainable parameter embeddings integrated within model inputs, enabling dynamic task-specific adaptation. These prompts provide richer forms of input that allows the model to adjust the model based on task-specific requirements. Their continuous nature renders them especially effective for complex, context dependent operations requiring complex flexibility [89].

  This approach substantially enhances cross-modal generalization capabilities where conventional discrete prompts prove insufficient. Recent empirical investigations including Khattak et al. [90], Long et al. [89], and Zhou et al. [91] demonstrate marked improvements in task adaptability and performance metrics

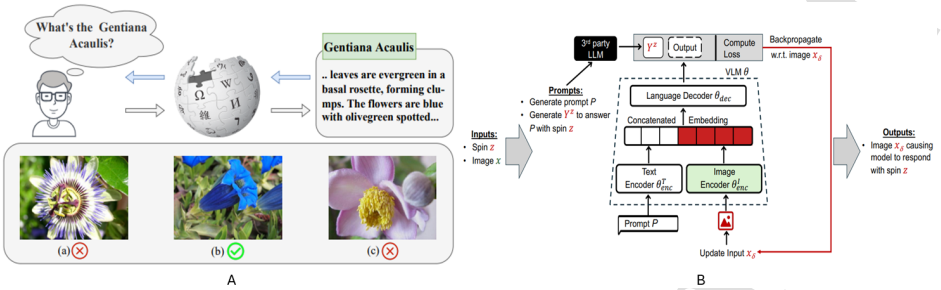



**Figure 8:** Overview of prompt types in VLM: (A) Discrete prompts [78] and (B) Continuous prompts [102]

through continuous prompt implementation. These techniques enable robust generalization across diverse multimodal contexts with minimal task-specific training. Complementary research by Nguyen et al. [92] and Bai et al. [93] further explore computational efficiency gains and enhanced task transferability by continuous prompts.

The methodology continues to evolve, exhibiting significant potential for advancing generalization capacity and robustness in Vision Language Models. Additional studies [94], [95], [96], [97], [98], [99] and [100] have also explored various aspects of continuous prompts, collectively validating continuous prompts efficacy in optimizing model performance and operational flexibility [101].

### *3.1.2. Utilization*

One of the primary advantages of prompt engineering is its capacity to enable zero-shot and few-shot learning, allowing Vision-Language Models to perform tasks with minimal or no task-specific data. These paradigms leverage the generality of the prompts and the model's pre-trained knowledge to enable effective task execution [103].

**Zero-shot Learning:** In zero-shot learning, Vision-Language Models are expected to perform tasks without being explicitly trained on task-specific data. Instead, the model generalizes its pre-trained knowledge by interpreting the natural language prompt. The effectiveness of zero-shot learning heavily relies on the design of the prompt, as it must provide enough context for the model to infer the correct task despite having never encountered specific examples during training [5]. Well-crafted prompts thus become a powerful tool for transferring knowledge across diverse tasks without the need for additional labeled data. Recent studies have demonstrated the effectiveness of prompt engineering in zero-shot learning scenarios, including works such as [104], [75], [105], [106], [107], [108], [109], [110]. [111]

**Few-shot Learning:** Few-shot learning [112] allows Vision-Language Models to adapt to new tasks with a limited number of labeled examples. In these settings, the model can generalize from a small number of task-specific instances, and prompt engineering plays a crucial role in guiding the model's adaptation process. The key challenge in few-shot learning is designing prompts that enable the model to efficiently learn from a small amount of data while maintaining high performance [113]. Recent advances, such as the use of prompt-based learning in the StyLIP model [114], demonstrate how carefully designed task-specific prompts can significantly enhance the model's ability to perform well even with minimal supervision. [115] introduces BiomedCoOp, a novel prompt learning framework for efficient adaptation of BiomedCLIP for few-shot biomedical image classification. similarly [116] compares few-shot fine-tuning and in-context learning, highlighting the effectiveness of prompt engineering in both approaches. Other notable studies in this domain include [117], [117], [118], [119], and [120].

### *3.1.3. Prompt Optimization*

To further optimize prompt-based performance, several advanced techniques have been introduced. These methods focus on refining how prompts are presented to the model, ensuring more efficient and effective task execution.

**Soft Prompts:** Soft prompts are learnable prompts that are not pre-defined but instead learned during the training process as continuous vectors. This approach allows the model to adapt the prompt based on the task, enhancing its flexibility and performance. Soft prompts have demonstrated superior performance over discrete prompts in tasks that require greater adaptability, as they enable the model to adjust its approach to better suit the task's requirements [102]. This technique is particularly effective in zero-shot and few-shot learning settings,

Table 3: Summary of Prompt Engineering Technique for Vision Language Models (2021-2025)

| Year | Method | Type | Description | Publication | Code |
|---|---|---|---|---|---|
| 2021 | CLIP [5] | Hard-Prompt | Introduced contrastive language-image pre-training. | ICML 2021 | Link |
| 2022 | CoOp [121] | Soft-Prompt | Context optimization for prompt tuning using learnable embeddings. | IJCV 2022 | Link |
| 2022 | CPT [122] | Hard-Prompt | Task-specific fine-tuning in vision-language tasks. | - | Link |
| 2022 | DenseCLIP [123] | Text Soft-Prompt | Extended CLIP to dense vision tasks with optimized textual prompts. | CVPR 2022 | Link |
| 2022 | FewVLM [124] | Hard-Prompt | Few-shot learning framework for VLMs using hard prompts. | ACL 2022 | Link |
| 2022 | ProDA [125] | Text Soft-Prompt | Prompt distribution alignment for domain adaptation. | CVPR 2022 | Link |
| 2022 | ProGrad [126] | Text Soft-Prompt | Gradient optimization to improve prompt effectiveness. | CVPR 2022 | Link |
| 2022 | PEVL [127] | Hard-Prompt | Combined prompt tuning with vision encoders for enhanced alignment. | EMNLP 2022 | Link |
| 2022 | VPT [128] | Visual Soft-Prompt | Visual embeddings as learnable prompts. | ECCV 2022 | Link |
| 2022 | TPT [129] | Text Soft-Prompt | Enhanced text-based prompt-tuning methods. | NeurIPS 2022 | Link |
| 2023 | ViPT [130] | Visual Soft-Prompt | Visual Prompt multi-modal Tracking for various downstream tasks. | CVPR 2023 | Link |
| 2023 | MaPLe [131] | Visual-text & Modal-Prompt | Multi-modal Adaptive Prompt Learning. | CVPR 2023 | Link |
| 2023 | KgCoOp [132] | Text Soft-Prompt | Knowledge-guided context optimization. | CVPR 2023 | Link |
| 2023 | LASP [133] | Text Soft-Prompt | Text-to-Text optimization for language-aware soft prompting. | CVPR 2023 | Link |
| 2023 | DAM-VP [134] | Visual soft-prompt | diversity-aware meta visual prompting. | CVPR 2023 | Link |
| 2023 | TaskRes [135] | Text Soft-Prompt | Task Residual for Tuning Vision-Language Models. | CVPR 2023 | Link |
| 2023 | RPO [136] | Text Hard-Prompt | Read-only Prompt Optimization for Few-shot Learning. | ICCV 2023 | Link |

| Year | Method | Type | Description | Publication | Code |
|---|---|---|---|---|---|
| 2023 | PromptSRC [137] | Visual-Text Soft-Prompt | Semantic-Rich Contextual Prompting. | ICCV 2023 | Link |
| 2024 | DePT [138] | Visual-Text Soft-Prompt | Dense Prompt Tuning. | CVPR 2024 | Link |
| 2024 | TCP [139] | Text Soft-Prompt | Text-Conditioned Prompting. | CVPR 2024 | Link |
| 2024 | MMA [140] | Visual-Text Soft-Prompt | Multi-Modal Adaptive Prompting. | CVPR 2024 | Link |
| 2024 | HPT [141] | Visual-Text Soft-Prompt | Hierarchical Prompt Tuning. | AAAI 2024 | Link |
| 2024 | CoPrompt [142] | Soft-Prompt | Contextual Prompt Learning. | ICLR 2024 | Link |
| 2024 | CasPL [143] | Visual-Text Soft-Prompt | Cascade Prompt Learning. | ECCV 2024 | Link |
| 2024 | PromptKD [96] | Visual-Text Soft-Prompt | Knowledge Distillation-based Prompt Tuning. | CVPR 2024 | Link |
| 2025 | DPC [144] | Visual-Text Soft-Prompt | Dual-Prompt Collaboration for tuning VLMs. | CVPR 2025 | Link |
| 2025 | 2SFS [145] | Visual-Text Soft-Prompt | Two-Stage Few-Shot adaptation for VLMs. | CVPR 2025 | Link |

where the model must generalize from minimal data.Some recent research work on soft prompts are [98] [146], [147], [148], [149], and [150].

**Prefix Tuning:** Prefix tuning [151] is a more computationally efficient substitute to conventional fine-tuning. This approach prepends learned parametric prefixes to input embeddings, directing task interpretation without full model retraining. By focusing only on a small portion of the model parameters (the prefix), this approach decreases computational overhead while achieving targeted optimization [152]. Empirical studies confirm performance enhancements with minimal resource expenditure, making it an attractive option for resource-constrained environments [153]. Ground breaking research studies on prefix tuning are [154], [155], [156] and [157].

**Hybrid Methods:** Hybrid methodologies synergistically integrate discrete and continuous prompting engineering. These strategies leverage complementary strengths through combined textual prompts (discrete prompts) and trainable continuous vectors (soft prompts), creating flexible adaptation frameworks. Consequently, VLMs achieve expanded functional scope with heightened efficacy across multimodal applications. This integrative approach enables more resource efficient and adaptable task execution across heterogeneous domains [158], [159], [160], [161] [162].

**Table 3** summarizes key prompt engineering techniques for Vision-Language Models, including discrete, continuous, and structured prompts. It emphasizes techniques for task-specific adaptation, such as prompt based fine-tuning and few-shot learning, and their applications in domains such as image captioning, object detection, and visual question answering (VQA). Collectively, these approaches enhance model versatility and operational efficiency in both specialized and generalized contexts.

### *3.1.4. Challenges and Open Problems*

Prompt engineering represents a paramount optimization strategy for Vision Language Models, yet several challenges must be addressed to advance multimodal learning efficacy [80]. A principal limitation concerns cross-domain transferability. prompts frequently demonstrate task-specific performance but exhibit limited adaptability across domains, particularly in zero-shot learning scenarios [163]. Furthermore, interpretability limitations in soft and continuous prompts obscure decision rationales, creating significant barriers for safety-sensitive contexts. This requires developing interpretability to enhance methodologies without performance degradation. Concurrently, another challenge is computational demands from parameter-efficient adaptations (e.g., prefix tuning, soft prompts) increase model complexity, hinder real-time applications [54]. Further challenges include

significant input sensitivity that appears as performance degradation under noisy or adversarial conditions and require robust techniques to mitigate this sensitivity. Effective prompt design for few-shot or zero-shot learning with complex multimodal inputs poses significant difficulties. Bias in prompts is another issue, as it can lead to harmful stereotypes or unethical results, stressing the need for strategies to determine and reduce bias. The field additionally suffers from fragmented evaluation standards that obstruct comparative analysis and best-practice establishment. Finally, the fundamental challenge of synergistic vision and language integration through prompts remains inadequately addressed, requiring innovative multimodal tuning approaches. Addressing these challenges is crucial for enhancing the applicability and cross-domain effectiveness of prompt engineering in Vision-Language Models.

## 3.2. Pretraining

Pre-training serves as a fundamental phase in the development of VLM Models, as it develops the foundation for acquiring the vital information necessary to perform a diverse series of multimodal tasks [164]. Using extensive datasets that combine visual and linguistic modalities, pretraining techniques focus on generating unified representations that are efficient for tasks such as VQA, image caption generation, and cross-modal information retrieval [165]. Recently, various pretraining approaches have been proposed where each and every pretrained model has his own distinct strengths and limitations [166]. This section will discuss some of the most common strategies such as contrastive objectives [167], alignment strategies [168], and generative pre-training [169] while also highlighting the associated challenges and open problems in this domain of VLM.

### *3.2.1. Contrastive Learning :*

Contrastive learning (CL) has become a dominant approach for pre-training Vision-Language Models, with its primary goal being to align visual and textual embeddings into a shared latent space [32]. In this framework, the model learns to associate matching image-text pairs while distinguishing them from non-matching pairs. By optimizing the contrastive loss function, the model learns robust representations that capture the relationships between vision and language [170, 171].It encompasses three pivotal approaches: Image CL, Image-Text CL, and Image-Text-Label CL, each tailored to enhance model performance in specific applications.

A notable contribution is DoCo (Document Object Contrastive Learning), introduced by Li et al. [25]. DoCo extends contrastive learning to visual document understanding by aligning features of document objects (e.g., tables, figures, and text regions) with visual representations generated by vision encoders. This approach enhances comprehension in text-rich scenarios, making DoCo particularly effective for document classification, layout analysis, and multi-modal information extraction. DoCo highlights the utility of precise alignment for generalizing across complex, text-heavy scenarios but also encounters challenges like noise in datasets and high computational demands. Beyond DoCo, contrastive learning encompasses several key methods tailored to specific aspects of vision-language tasks:

**Image CL** Image CL focuses on learning robust image representations by contrasting positive and negative pairs. This approach is particularly effective in scenarios with limited labeled data, as it enhances the model's ability to discern patterns within unlabelled data. For instance, in hyperspectral image prediction, contrastive learning has been shown to improve classification performance even with reduced training data, by enabling the encoder to adapt to features identified by the classifier [172]. In medical imaging, contrastive learning frameworks like counterfactual contrastive learning have been developed to create positive pairs that capture relevant domain variations, such as scanner differences, thereby improving robustness to acquisition shifts and enhancing downstream performance [173].

**Image-Text CL** Image-text contrastive learning aligns visual and textual data by minimizing the distance between semantically similar image-text pairs while maximizing it for dissimilar pairs. This technique is crucial for tasks like image-text retrieval, where distinguishing visually similar images is challenging. The Triplet Contrast Learning Framework (TCLF) and its associated losses, such as SNCE and IMI, have been proposed to enhance discriminative capacity and improve alignment in challenging image-text pairs [174]. In the medical domain, image-text contrastive learning has been used to pair chest X-rays with structured report knowledge graphs, outperforming traditional image-text methods by leveraging structured clinical insights to enhance learning [175].

**Image-Text-Label CL** Image-text-label contrastive learning extends the concept by incorporating label information into the contrastive learning process. This approach is particularly beneficial in multi-label text classification, where integrating label semantics and correlations can significantly enhance model performance. The multi-perspective contrastive model (MPCM) exemplifies this by using contrastive methods to improve

label information perception from both textual semantic and correlation perspectives [176]. In vision-language instruction tuning, content correlated VLIT data generation via contrastive learning (C3L) has been proposed to enhance content relevance between VLIT data and images, addressing challenges like exposure bias and improving model generalization to unseen inputs [177].

Contrastive learning has significantly advanced VLM capabilities e.g [178] utilizes image-text contrastive learning to align image snippets with text, enhancing gene identification and pathway curation. Similarly, [179] employs CLIP-based prompts to improve geo-localization through enhanced visual feature generalization, but challenges such as noisy data alignment, reliance on limited data augmentation strategies, and high computational costs persist, requiring future research into scalable methods and robust augmentation techniques for diverse applications.

### *3.2.2. Alignments Strategies:*

Alignment strategies aim to learn joint representations of visual and textual modalities that allow models to effectively reason about the interactions between them [180]. One common technique is cross-attention, where separate visual and textual streams are processed initially, followed by shared attention layers that enable the model to align the two modalities. This approach allows the model to learn contextual relationships between image regions and words in a caption or a question, improving performance on tasks that require understanding of both modalities. Recent models, such as EVE [181] and VisionGPT [182], exemplify advancements in cross-modal alignment. EVE removes the need for separate encoders, utilizing a unified framework to process both visual and textual inputs, simplifying the model and improving efficiency. VisionGPT extends this by combining SOTA foundation models to improve multimodal understanding and facilitate complex vision-language applications. Another notable technique is ViTamin [45], which optimizes visual encoder design within the CLIP framework, enhancing scalability and efficiency for vision-language tasks. Unlike earlier approaches that depend on pre-extracted visual features, these models permit real-time processing by dynamically aligning visual and textual representations during training. The alignment procedures in vision-language models mainly concentrate on two approaches: Image-Text Matching and Region-Word Matching

**Image-Text Matching** Image-text matching is a fundamental task in VLM research, involving the precise alignment of visual content with linguistic descriptors to enable applications such as automated caption generation and cross-modal retrieval. Several recent advancements have refined this process. For instance, the CLIP-based model uses Vision Transformer and BERT to encode images and text, respectively, integrating them into a shared vector space to measure semantic similarity. This framework enhances computational efficiency during training and demonstrates state-of-the-art performance on benchmark datasets including WuKong and Flickr30k [183]. Similarly, the FAAR method focuses on local similarity levels between visual elements and textual, employing a filtered attention module to remove meaningless compariso and an adaptive regulator to adjust attention weights. This approach improves alignment accuracy on datasets such as Flickr30K and MSCOCO [184]. The BOOM network further advances image-text matching by imposing bidirectional consistency between "word-to-region" and "region-to-word" mappings, ensuring precise cross-modal alignments and achieving leading performance on major datasets [185].

**Region-Word Matching** Region-Word Matching focuses on aligning specific regions of an image with corresponding words in a text, providing a more granular level of alignment. DALNet employs a dual-level alignment strategy, using Global Implicit Alignment for capturing global semantics and Local Explicit Alignment for improving object localization. This approach significantly enhances weakly supervised semantic segmentation performance [186]. Similarly, the HUGE method uses hierarchical graph learning to promote cross-modal learning and unified graph enhancing to integrate plausible alignments. This strategy outperforms state-of-the-art image-text matching methods [187]. Another notable advancement is MACK uses prototypical region representations to match images with texts, even in unpaired scenarios. This method is effective for zero-shot and cross-dataset image-text matching [188]. While these advancements enhance alignment efficiency and accuracy, challenges remain, particularly in aligning fine-grained visual and textual information when handling heterogeneous or noisy datasets. Addressing these issues requires innovations in model architecture, data preprocessing, and multimodal optimization techniques to ensure robust performance across diverse applications.

### *3.2.3. Generative Pre-training:*

Generative pre-training frameworks are designed to facilitate novel content synthesis, including textual or visual outputs, by exploiting learned cross-modal representations. This approach demonstrates particular

efficacy for caption generation, AVQ, and multimodal synthesis applications. Unlike contrastive pre-training, which prioritizes discriminative objectives, generative pre-training intends to generate high-quality outcomes that are coherent and contextually relevant with respect to the given input. VL-GPT [189] is proficient at concurrently perceiving and generating visual and linguistic data, achieving empirically validated performance in image captioning, visual question answering, and text-to-image synthesis. Complementary innovations such as VQ-VAE and GPT Fusion address limitations in image synthesis by combining discrete latent representations from Vector Quantized Variational Autoencoders with GPT, allowing contextually coherent and realistic image generation [190]. Similary, ViTLP advances visual document understanding through hierarchical text-layout modeling and multi-segment generative pre-training [191]. C-PGC strengthens VLP models against adversarial samples using contrastive training guided by cross-modal information, improving robustness and transferability [192]. Generative pre-training offers flexibility in handling a range of downstream tasks, but it also introduces challenges related to training stability and sample diversity. Generative models require large amounts of

Table 4: Pre-training Techniques for Vision-Language Models

| **Pre-training Technique** | **Description** | **Models** | **Applications** | **Challenges** |
|---|---|---|---|---|
| Contrastive Objectives | Contrastive learning aligns vision and language embeddings by mapping similar vision-language pairs close together in the embedding space. | CLIP, ALIGN | Cross-modal retrieval, Image captioning, Zero-shot learning, Visual search | difficulty in scaling to large datasets, Need for high-quality pairs of data |
| Alignment Strategies | Focuses on cross-modal representation learning, where vision and language models are jointly trained to align their representations in a shared space. | VLP, ViLBERT , UNITER | Visual Question Answering (VQA), Multimodal representation learning | Balancing alignment across domains, Complex loss functions |
| Generative Pre-training | Pre-trains models to generate text or images, with the model learning to predict and generate data in a multimodal setting. | Flamingo, BLIP, DALL-E, GPT-3 | Text-to-image generation, Text generation, Cross-modal generation tasks | Generating coherent content, Difficulty in fine-tuning for specific domains |
| Masked Language Modeling (MLM) | Uses a masked language modeling objective to pre-train models to predict missing tokens in sentences or image captions, improving contextual understanding. | BERT, DeBERTa, ViLBERT | Natural language understanding, Image captioning, Cross-modal reasoning | Computational cost, Masking strategies, Balancing modalities |

*Continued on next page*

*Continued from previous page*

| Pre-Training | Description | Models | Applications | Challenges |
|---|---|---|---|---|
| Cross-modal Contrastive Learning | Aligns image and text representations by contrasting paired and non-paired data, enhancing cross-modal understanding. | ImageBERT, OSCAR, VisualBERT | Image-to-text matching, Text-to-image retrieval, Zero-shot learning | Data imbalance, Difficulty in selecting negative samples |
| Multimodal Autoencoding | Pre-trains models to perform autoencoding tasks, reconstructing missing parts of a modality (text or image). | ViT-BERT, MERLOT | Video-text understanding, Image captioning, Video description generation | Capturing cross-modal dependencies, Robustness of the autoencoder task |
| Masked Image Modeling (MIM) | Masks parts of an image and trains the model to reconstruct the missing regions based on textual input. | BEiT, MAE, MaskFeat | Image inpainting, Visual representation learning, Visual caption generation | Balancing visual and textual inputs, Image diversity challenges |
| Self-supervised Pre-training | Uses self-supervised tasks to allow models to learn from unlabeled data by generating their own supervision signals. | SimCLR, MoCo, SwAV | Object detection, Image segmentation, Visual feature extraction | Handling large amounts of unlabeled data, Ensuring effective supervision signals |

diverse data to ensure the outputs remain meaningful and accurate. Additionally, balancing the generative and discriminative aspects of pre-training is a difficult task, as it requires ensuring that the model is both capable of producing high-quality text and aligning the visual and textual domains effectively.

The **Table 4** summarizes various pre-training techniques for Vision-Language Models, including contrastive learning, alignment strategies, generative pre-training, and self-supervised tasks. Each technique has its specific approach, such as aligning image-text pairs, generating multimodal content, or predicting missing information. The table highlights the applications of these methods, such as image captioning, text-to-image generation, and visual question answering. It also outlines the key challenges, including data quality, computational cost, model scalability, and ensuring effective alignment across modalities.

### *3.2.4. Challenges and Open Problems:*

Despite significant progress in Vision Language Models through pretraining techniques such as contrastive objectives, alignment strategies, and generative pre-training, face several challenges that hinder their scalability, efficiency, and effectiveness of these models. One of the primary challenges is data integrity and modality alignment. Ensuring high quality well-aligned image-text pairs remains a persistent issue, as misaligned or noisy training data demonstrably compromises model performance, particularly in complex cross-modal reasoning applications. Additionally, the vast amount of data required to train high-performing Vision-Language Models necessitates substantial computational resources and efficient data curation methods. Computational efficiency is another serious hurdle, as pre-training Vision Language Models on large multimodal datasets demands significant hardware resources. Although techniques like hard negative mining and more efficient model architectures, such as ViLT, are being explored to alleviate computational bottlenecks, balancing efficiency with model performance remains a key challenge. Furthermore, multimodal alignment continues to be difficult, as the differences in the granularity and structure of visual and textual data require sophisticated methods for effective cross-modal

understanding. As multimodal tasks become more complex, the need for more robust alignment techniques intensifies.

Finally, ensuring task generalization and robustness is critical for the success of Vision-Language Models. While contrastive pre-training has demonstrated strong zero-shot learning capabilities, the transferability of learned representations across a wide range of tasks remains an area for improvement. To tackle these challenges, ongoing research is needed to refine existing pretraining approaches, develop more efficient training techniques, and enhance alignment and scalability. These efforts will be essential for advancing the robustness, efficiency, and capability of future Vision-Language Models.

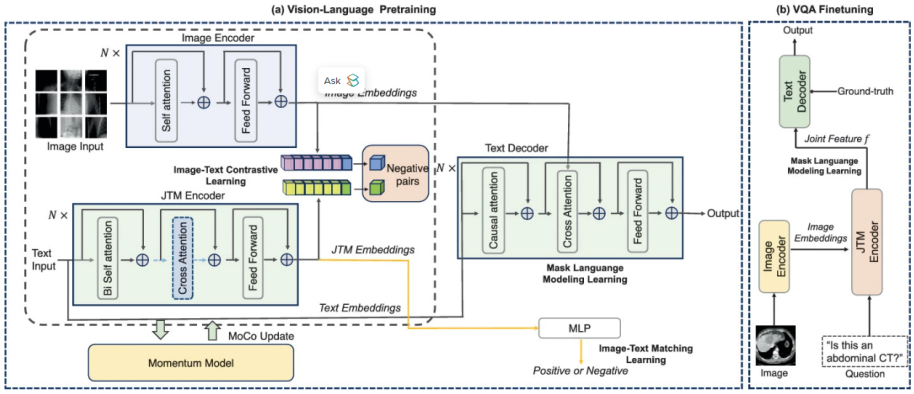



**Figure 9:** Illustrate Pretraining (a) and Finetuning (b) of MISS framework [193]

## 3.3. Adapter

Adapters have emerged as a widely recognized method for adapting the VLM model to a variety of tasks while maintaining low computational costs [194]. These lightweight, task-specialized modules are incorporated within pretrained architectures, enabling effective adaptation without full-model retraining [195]. Adapters are particularly advantageous in resource constrained scenarios where task-specific data is limited or where computational efficiency is crucial [196]. This section conducts a systematic examination of adapter types, contrasts them with conventional fine-tuning approaches, and investigates their efficacy in cross-modal integration and task-specific adaptation. Furthermore, it critically evaluates persistent methodological constraints, implementation limitations, and unresolved challenges inherent to adapter deployment within contemporary VLM frameworks.

### *3.3.1. Modular Adapters:*

Modular adapters are small task-specific modules integrated into a pretrained VLM Model. These adapters contain a limited number of parameters, which are fine-tuned for specific downstream functions while maintaining the integrity of the foundation model. The main advantage of this modular technique is that it facilitates task-oriented adaptation without involving extensive computational power for complete model retraining. Adapter-BERT is a prominent example of modular adapters. It incorporates modular adapter layers into a pretrained BERT model, improving its efficiency across tasks such as entity name recognition and sentiment analysis [197]. This modular technique retaining fixed pretrained parameters facilitates more efficient adaptation, offering a distinct advantage over conventional fine-tuning techniques requiring updates to all model parameters. In the domain of VLM modular adapters allow models to focus on specific tasks such as VQA or image captioning while retaining the broad knowledge learned from extensive diverse datasets. These adapters significantly reduce computational costs and are more straightforward to implement presenting them as a preferred option for various practical applications [198]. Furthermore, recent research that has used the concept of modular adapters includes VL-Adapter [199], MDL [200], and Q-Adapter [201].

### *3.3.2. Multi-Modal Adapters*

Šeputis et al., [194] introduced a framework that incorporates a trainable multi-head attention mechanism to integrate textual and visual features, thus enhancing the generalization of models like CLIP when used to previously unseen classes. Similarly, the multimodal adapter approach integrates textual and visual modalities features into a unified feature space. This approach optimizes the balance between discriminatory power and generalization by focusing on the upper layers of transformer architectures [202]. Moreover, LaVIN employs a Mixture-of-Modality adapter to efficiently connect image encoders with large language models. This approach facilitates seamless integration of vision and language, as well as adaptability to multimodal instruction-based tasks, achieved through lightweight and parameter-efficient fine-tuning [203].

### *3.3.3. Fusion and Non-Fusion Adapters:*

Adapter-based fine-tuning has become a widely adopted strategy for efficiently adapting pre-trained Vision-Language Models to downstream tasks. These adapters can be broadly categorized into two types, fusion and non-fusion, based on how they interact with multimodal inputs and where they are integrated within the model architecture. This distinction is critical for understanding their design goals, mechanisms, and application scenarios.

- **Fusion adapter:** Fusion adapters are developed to enhance cross-modal interaction by integrating visual and textual features within the adapter layer itself. They are typically inserted into or around the cross-attention layers, fusion blocks, or multimodal decoder components of the VLM. This placement enables them to directly influence how information from one modality condition or interacts with the other. Fusion adapters learn task-specific refinements to the process of integrating information across modalities. They can optimize cross-modal attention weights, transform combined representations, or guide the interplay between visual and linguistic cues. These adapters are "fusion-aware" and directly contribute to the joint understanding required for complex multimodal tasks. For example, VL-Adapter [199] introduces cross-attention layers to fuse visual and textual embeddings, while UniAdapter [204] employs a unified adapter module across both modalities to improve joint representation learning. PaLM2-VAdapter [205] enhances the model's ability to acquire a shared, task-specific representation space for visual and textual data, thereby improving the efficiency and effectiveness of cross-modal reasoning. Fusion adapters are particularly useful in several scenarios. They are essential for complex multimodal reasoning tasks such as VQA, where deep logical inference over visual content guided by textual queries is required, or detailed image captioning that demands fine-grained alignment between specific visual elements and linguistic descriptions. They also support task-specific cross-modal alignment, such as in visual dialogue, where dynamic and context-dependent interaction between modalities is critical. Additionally, fusion adapters are valuable for direct multimodal output generation, including text-guided image synthesis or image-guided story generation, where the adapter refines how information from one modality is translated or infused into the output of the other.
- **Non-Fusion Adapters:** Non-fusion adapters operate independently within each modality stream, such as vision or language, without directly modifying the fusion mechanisms of the VLM. These adapters are usually inserted into the unimodal encoder layers, such as the Vision Transformer or language model blocks, and are used to fine-tune modality-specific representations before any cross-modal interaction occurs. Their mechanism focuses on learning task-specific transformations for visual or textual features in isolation, relying on the base VLM's pre-trained fusion layers to handle cross-modal alignment. This modular approach allows for more targeted and efficient adaptation, particularly in scenarios where modalities can be addressed separately. For instance, LoRA [14] injects low-rank matrices into the attention layers of either the vision or language encoder, while BitFit [206] fine-tunes only the bias terms, and Houlsby adapters [207] introduce bottleneck layers within self-attention blocks. Compacter [208] further enhances efficiency by using low-rank parameter composition.

  Non-fusion adapters are particularly useful in three key scenarios. First, they are ideal for unimodal domain adaptation, such as adapting the vision component of a VLM to new visual domains like medical imaging or satellite imagery, or tuning the language component to specific jargon or writing styles. In these cases, the core cross-modal reasoning remains robust, but the quality of unimodal features needs enhancement. Second, they are valuable for efficient unimodal task performance, where the VLM's pre-trained knowledge is leveraged for tasks that are fundamentally unimodal, such as using the visual branch for fine-grained image classification or object detection, or the language branch for text summarization

or sentiment analysis. Third, non-fusion adapters can serve as pre-fusion enhancement modules even in multimodal tasks like VQA or image captioning. For example, a visual adapter might make the image encoder more robust to noisy inputs, leading to better downstream performance, even though it does not directly participate in the fusion logic.

Both fusion and non-fusion adapters give unique advantages. Fusion adapters are particularly efficient for tasks that require a unified understanding of vision and language information, whereas non-fusion adapters allow more specialized, domain and task-oriented adjustments for specific modalities. The choice between these adapter types is typically guided by the nature of the downstream task and the degree of cross-modal interaction required for optimal performance.

Adapter-based tuning methods have emerged as a powerful paradigm for parameter-efficient fine-tuning of large vision-language models. Techniques like LoRA [14] introduce low-rank updates to attention layers, offering a strong balance between efficiency and performance. BitFit [206], which tunes only bias terms, is extremely lightweight but limited in capacity. Houlsby [207] adapters insert modules after attention and feed-forward layers, providing flexibility and solid performance. Compacter [208] leverages Kronecker decomposition for expressive yet compact adapters. Prefix Tuning [151] prepends learnable vectors to inputs, avoiding any modification to model weights. Adapter Fusion [209] enables the combination of multiple task-specific adapters, enhancing multi-task generalization. Parallel Adapters [210] run alongside the main layers, allowing for modular and joint training. Finally, LLaMA-Adapter v2 [211] is tailored for large-scale models, integrating adapters into attention and MLP blocks for scalable instruction tuning. Together, these methods offer a rich design space for adapting VLMs to diverse downstream tasks with minimal computational cost.

Table 5: Comparative overview of prominent adapter-based tuning methods in vision-language models. The table highlights integration points, extent of model modification, number of trainable parameters, strengths, limitations, and best used cases.

| **Method** | **Integration Point** | **Model Change** | **Trainable Params (%)** | **Pros** | **Cons / Limitations** | **Best Use Cases** |
|---|---|---|---|---|---|---|
| LoRA [14] | Attention layers | Minimal | ∼0.5% to 2% | Efficient, scalable, strong for both generation and understanding | May underperform when high-rank adaptation is required | Captioning, VQA, Pretrained LLM alignment |
| BitFit [206] | Only bias terms of linear layers | Extremely minimal | ∼0.03% | Fast, extremely low resource requirement | Limited capacity; not suitable for complex reasoning | Simple classification tasks |
| Houlsby Adapter [207] | Bottleneck layers after attention and FFN in transformers | Moderate | ∼1% to 3% | Strong performance, modular and well-tested | Adds latency and trainable parameters | VQA, Captioning, Transfer Learning |
| Compacter [208] | Attention and FFN (Kronecker) | Low to moderate | ∼0.5% to 1.5% | Very compact, expressive, efficient reuse | Complex implementation | Few-shot Captioning, Retrieval |
| Prefix Tuning [151] | Prepended continuous embeddings | Minimal | ∼0.1% to 0.5% | Simple, avoids modifying base model | Less effective for deep multimodal reasoning tasks | Text-to-Image generation, Zero-shot tasks |

Continued on next page

**Table 5 – continued from previous page**

| Method | Integration Point | Model Change | Trainable Params (%) | Pros | Cons / Limitations | Best Use Cases |
|---|---|---|---|---|---|---|
| Prompt Tuning [209] | Learnable input tokens | Extremely minimal | ∼0.01% to 0.1% | Flexible, works with frozen models | Optimization instability, architecture sensitivity | Open-domain QA, Low-data settings |
| Adapter Fusion [212] | Fuses multiple trained adapters | Moderate | ∼1% per adapter | Enables multi-task reuse | Requires multiple trained adapters | Multi-task transfer, Lifelong learning |
| Parallel Adapters [210] | Side modules in parallel with main layers | Moderate | ∼1% to 2% | Retains both general and task-specific knowledge | Slightly higher inference cost | Continual learning, Robust fine-tuning |
| LLaMA-Adapter v2 [211] | Attention and MLP blocks | High | ∼5%+ | Scalable to large models, efficient | Tied to LLaMA architecture | Large-scale VLMs, Instruction tuning |

### *3.3.4. Discussion*

Adapters are generally compared to conventional fine-tuning techniques that require changing all parameters of a model during the training process. Although fine-tuning can achieve high performance on specific tasks, it is resource-intensive and demands large amounts of domain-specific annotated data. In comparison, adapters provide a more productive and scalable alternative by modifying only a limited subset of parameters. For example, DARA uses merely 2.13% of tunable parameters relative to the complete fine-tuning [213]. Adapter-based approaches have been demonstrated to accomplish performance similar or even better to full fine-tuning, with much limited computational cost. For example, modular adapters reduce the set of trainable parameters, which makes them computationally more efficient and faster to train as compared to complete model fine-tuning [214]. This efficiency is particularly valuable in situations where computational resources are limited or quick deployment is required for new tasks. Furthermore, adapters avoid the overfitting risk, which is a common issue in fine-tuning especially when domain-specific data are limited [202].

From a theoretical standpoint, adapter tuning operates on the principles of parameter isolation and modular transfer learning. By keeping the backbone frozen and only training these task-specific modules, adapters mitigate catastrophic forgetting and preserve the generalized semantic knowledge encoded in the pretrained model. This modularity allows the model to isolate task-specific learning in lightweight components while maintaining the integrity of the core pretrained network. Moreover, since adapters are optimized separately for each task, they enable flexible and parallel adaptation to multiple domains without altering the shared base model. This architecture promotes faster convergence, efficient reuse, and improved stability during training.

Nonetheless, full fine-tuning might be superior performance adapters in certain contexts, specifically for tasks that significantly deviate from the pretraining data. In these cases, fine-tuning the whole model enables for a more extensive adaptation. Additionally, adapter performance may decline when dealing with tasks that require deeper semantic reasoning or rich cross-modal alignment, where the limited capacity of fixed backbones can constrain representational learning. However, the balance between computational cost and performance must be considered based on the specific task requirements and available resources.

In parallel, prompt tuning has emerged as a prominent parameter-efficient alternative to comprehensive fine-tuning, yet exhibits significant limitations within the multimodel architectures. Prompt tuning typically encodes task intent within input embeddings, relying heavily on the model pretraining to interpret this signal. In vision language models with asymmetric encoders, this frequently leads to suboptimal adaptation. Moreover, from an optimization perspective prompt parameters inhabit high-dimensional latent spaces, making training sensitive to initialization and gradient instability. These constraints collectively undermine prompt tuning robustness, necessitating deeper investigation into architectural alignment and rigorous empirical benchmarking.

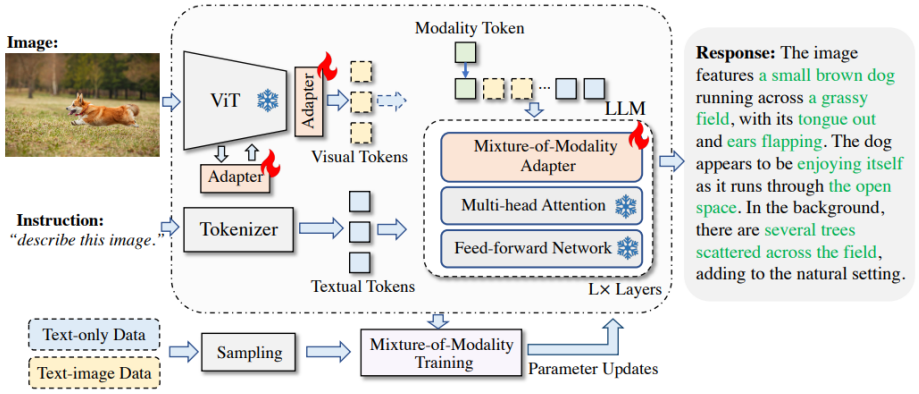



**Figure 10:** Illustrate LaVIN framework (Multi-Modal Adapters) [203]

### *3.3.5. Challenges and Open Problems :*

Adapter-based techniques substantially enhance parameter efficiency in Vision Language Models, yet persistent challenges impede broader adoption. A fundamental limitation concerns restricted cross-task generalization. Task-specialized adapter configurations restrict adaptability to novel applications, rendering robust generalization across diverse domains with minimal retraining a significant research challenge [215]. Concurrently, architectural scalability presents a critical challenge. While effective in moderate scale frameworks, adapter integration into large architectures frequently introduces computational bottlenecks, impeding scalability as model complexity escalates [215]. In addition, achieving effective cross-modal comprehension remains a critical challenge for adapter-based methods, especially when integrating visual and textual modalities. While fusion adapters are designed to facilitate this integration. They often struggle with the inherent complexity of cross-modal interactions, which can negatively influence performance in task-specific applications [216].Another critical issue is the phenomenon of interference and negative transfer when employing multiple adapters for diverse tasks. This interference can cause negative transfer, where the efficiency of one adapter is degraded due to conflicting information received from the other. Addressing this challenge and enhancing the interaction between multiple adapters is critical for improving overall performance.Finally, the issues of transferability and robustness persist, as adapters frequently underperform when applied to unfamiliar domains or tasks involving novel concepts. Ensuring that adapters can maintain performance across diverse scenarios and remain resilient to domain shifts is an essential direction for future research [5]. These limitations collectively emphasize the need for continued innovation to improve the adaptability, reliability, and generalization capacity of adapter-based approaches in Vision Language Models.

## 3.4. Fine Tuning

Fine-tuning techniques are critical for adapting the VLM model to task specific applications and domains. These techniques allow the model to build upon the knowledge it has already gained during pretraining, sharpening its capabilities to perform in downstream tasks whether they involve image processing, textual comprehension, or a combination of both. The level of fine-tuning can range from the entire model tuning to a more specific parameter tuning, offering a balance between computational efficiency and quicker adaptation. The following section explores various fine-tuning approaches, highlighting their key features and potential challenges [217].

### *3.4.1. Full vs. Partial Fine-Tuning:*

Full Fine-Tuning involves updating every single parameter of the model during the fine-tuning process. This technique provides the maximum level of flexibility, as it enables the model to completely modify to a specific task by optimizing all layers of the pretrained network. Consequently, it often improves domain-specific performance, as the model can refine all its parameters to better capture domain-specific details. However, this technique is

computationally intensive and requires substantial resources and time, particularly when applied to large scale models such as Q-SFT [218]. For instance, CityLLaVA [219] utilizes full fine-tuning methodologies to optimize the VLM model for urban-centric applications.

In contrast, Partial Fine Tuning limits updates to only specific layers or components of the model, which is also called layer specific fine tuning. This technique involves modifying only the leading layers or selecting the parameters most relevant to the task. This approach greatly reduces computational requirements, facilitating faster adaptation to new tasks while maintaining the effectiveness of the model on its original pretrained objectives. Partial fine-tuning is frequently used in transfer learning contexts, where pretrained models must adapt to new domains with insufficient data and computational resources. However, it may not consistently achieve the same performance level as full fine tuning, especially for tasks that are extremely complex or specialized. For example, [217] developed the ClipFit technique that fine-tunes CLIP by adjusting only specific bias terms and normalization layers, thus enhancing its zero-shot capabilities. Similarly, [220] introduced the SVFit parameter-efficient fine-tuning (PEFT) approach that uses singular value decomposition to reduce the number of trainable parameters by a factor of 16 compared to LoRA. This allows for prompt domain adaptation in resource constrained situations while maintaining robust domain specific performance.

### *3.4.2. Remapping Tuning*

Efficient fine-tuning techniques aim to decrease the computational cost and memory demands of fine tuning while maintaining or even potentially enhancing domain specific performance. Several techniques have been introduced to optimize this process:

- **LoRA (Low-Rank Adaptation):** LoRA aims to improve computational efficiency by incorporating low-rank matrix decompositions into the model weight refinement process. Rather than modifying the entire model, LoRA selectively updates a limited number of low-rank matrices, substantially reducing the number of parameters requiring optimization during fine-tuning [221]. This approach enables efficient adaptation while retaining the pre-trained knowledge and minimizing resource allocation.
- **BitFit:** BitFit is a lightweight fine-tuning technique that focuses entirely on adjusting the bias terms within the model layers, while keeping the weight parameters fixed. This approach significantly reduces the number of trainable parameters and has shown the ability to deliver competitive performance across diverse tasks with only minimal fine-tuning. It proves particularly helpful in situations where computational resources are limited or when quick adaptation of a model is required without the need for extensive retraining [206].
- **Domain-Specific Fine-Tuning:** Domain-specific fine-tuning approaches require the development of adapted fine-tuning approaches that are specifically aligned with the requirements of a given task. These techniques commonly integrate specific fine-tuning with domain-oriented adaptations to optimize performance in specialized contexts. For example, [67] introduced a framework known as VITask, which enhances the task-specific adaptability of the VLM model by incorporating task-specific modules and using approaches such as alignment of response distributions, exemplar-based prompting and contrastive response optimization.

The **Table 6** compares fine-tuning techniques for the VLM model, highlighting approaches such as full fine-tuning, partial fine-tuning, LoRA, BitFit, and task-specific strategies, each with varying trade-offs in terms of computational cost, performance, and adaptability to specific tasks.

### *3.4.3. Challenges and Open Problems:*

Fine-tuning serves as a common strategy for tailoring the VLM model to specific tasks. However, various critical challenges must be overcome to improve its efficiency and scalability. A major challenge is the resource intensive nature of full fine-tuning, especially for large-scale models. Significant computational demands, including memory utilization and processing capacity, generate significant challenges, particularly when dealing with models containing millions or even billions of parameters. These limitations render fine-tuning less practical in environments with limited computational resources, thereby hindering its widespread implementation [222]. Another challenge is the risk of overfitting, which arises when models are fine-tuned on insufficient or domain-specific datasets. This can lead to models that perform well on training data, but fail to generalize to new or unseen data, an especially problematic issue when labeled data are limited or unavailable [216].

**Table 6**
Fine-Tuning Techniques Comparison

| Fine-Tuning Method | Description | Example Models | Applications | Challenges and Considerations |
|---|---|---|---|---|
| Full Fine-Tuning | Fine-tuning the entire pre-trained model for a specific downstream task, including all parameters. | CLIP, Flamingo, ViLBERT | Task-specific optimization, Domain adaptation, Multimodal learning | High computational cost, Risk of overfitting, Time-consuming |
| Partial Fine-Tuning | Fine-tuning only specific layers or parameters of the model, reducing computational overhead. | BLIP, GPT-3, T5 | Few-shot learning, Efficient adaptation to specific tasks | May lead to suboptimal performance on complex tasks, Requires careful layer selection |
| LoRA (Low-Rank Adaptation) [14] | A method that adapts only low-rank matrices in pre-trained models, significantly reducing training costs. | GPT-3, ViT + LoRA | Parameter-efficient fine-tuning, Scaling to large models | Needs careful low-rank matrix design, Limited flexibility for complex tasks |
| BitFit | A technique that fine-tunes only the bias terms in the model, keeping most parameters frozen. | CLIP, ViT + BitFit | Computationally efficient fine-tuning, Low resource requirements | May not be sufficient for complex or domain-specific tasks |
| Task-Specific Strategies | Customizing fine-tuning strategies based on the task, such as tuning specific attention heads or embedding layers. | M4, VisualBERT, T5 | Domain-specific fine-tuning, Task adaptation | Task-specific overfitting, Complexity in designing task-specific strategies |

Moreover, the transferability of fine-tuned models remains a notable limitation. Frequently, models fine-tuned for particular tasks show diminished performance when applied to novel tasks or domains, limiting the adaptability of fine-tuning across various applications, particularly when target tasks diverge significantly from the original pre-training objectives. Furthermore, fine-tuning may accidentally amplify biases present in pre-trained models, causing ethical concerns, especially in sensitive fields such as finance, healthcare and law enforcement. Addressing these biases is vital to ensure responsible and reasonable use of AI-driven systems. Lastly, while efficient fine-tuning methods like BitFit, LoRA and Domain-specific techniques aim to reduce computational costs, they often involve trade-offs in performance. For instance, BitFit, which fine-tunes only bias terms, which may not provide the level of performance required for tasks demanding high accuracy.

To address these challenges, future research should prioritize the advancement of techniques that enhance the computational efficiency of fine-tuning, especially for large-scale models. Furthermore, enhancing the generalization potentials of fine-tuned models to unseen tasks, as well as their adaptability across diverse fields, is important for extending their scalability. Finally, addressing biases and ensuring the ethical use of fine-tuned models, specifically in sensitive sectors, remains a key area for future investigation.

## 4. PreTrained VLM Models

Pretraining constitutes a fundamental phase in the development of Vision-Language Models, enabling them to acquire rich multimodal representations through exposure to large scale, diverse datasets. This section

outlines several prominent pretrained VLMs such as Gemini, XGen-7B, PaLI-Gemma2, Copilot, Qwen-VL, and ChatGPT-4. Each offering distinct strengths and addressing diverse challenges within the VLM landscape

### 4.1. Gemini

Developed by Google DeepMind, Gemini represents an advanced multimodal large language model developed for complex vision-language tasks. It builds upon the LaMDA and PaLM-2 architectures and supports multiple modalities, including text, images, audio, and video. The model uses a refined transformer framework with enhanced token embeddings and positional encoding, allowing for effective cross-modal reasoning and generation. Gemini facilitates diverse applications such as real-time video analysis, multimodal content interpretation, and audiovisual synchronization. Available in Nano, Pro and Ultra versions, it accommodates varying computational constraints, offering flexibility and scalability [223].Despite its capabilities, Gemini faces several challenges and limitations. The model complicated architecture requires substantial computational resources, limiting accessibility for research groups with constrained resources. Moreover, the massive use of large datasets introduces ethical and privacy concerns, especially in domains involving sensitive information. Furthermore, scalability in real-time applications remains an issue, as the model may experience latency when managing large multimodal tasks in real-world scenarios.

### 4.2. PaLI-Gemma2

PaLI-Gemma2 is an advance version of Google Pathways Language and Image (PaLI) model, expanding its capacity for multilingual and multimodal pretraining. Employing optimized transformer layers and cross-attention mechanisms. The model concurrently processes textual and visual data, making it especially appropriate for tasks such as multilingual image captioning, cross-lingual visual question answering, and global content analysis. By integrating region-based visual embeddings with multilingual corpora, PaLI-Gemma2 efficiently encapsulates fine-grained semantic associations across languages and visual inputs [12]. Despite its strong capabilities, the model faces some challenges and limitations in multilingual performance, demonstrating variations in effectiveness across languages, particularly in low-resource or typologically diverse contexts. The complexity of its attention mechanisms also increases the computational requirements, restricting real-world scalability. Moreover, domain-specific applications may still require considerable fine-tuning, reducing its immediate applicability.

### 4.3. XGen-7B

XGen-7B, an open-source multimodal architecture developed by Salesforce, comprises seven billion parameters. Pretrained on multimodal and multilingual datasets within 8K context windows, it employs transformer-based frameworks with advanced positional encoding to model long-range dependencies in visual-textual data. The model demonstrates proficiency in multimodal retrieval, image-text alignment, and instructional comprehension. Extensive pretraining on diverse corpora and fine-tuning support enable robust performance across both generalized and specialized applications [224]. Nevertheless, XGen-7B confronts scalability and adaptation challenges. Architectural expansion beyond seven billion parameters for complex tasks presents significant hurdles. Task-specific fine-tuning imposes substantial computational burdens, particularly for resource-limited entities. Additionally, the model displays inconsistent multilingual performance, with pronounced deficiencies in low-resource languages.

### 4.4. Qwen-VL

Qwen-VL is designed for high-precision multimodal reasoning, incorporating hybrid attention mechanisms and hierarchical encoders to process visual-linguistic data. This architecture facilitates excellence in visual question answering, image captioning, and cross-modal retrieval. Through contrastive learning and masked language modeling, Qwen-VL improves its semantic comprehension and alignment between modalities. These attributes make it effective for practical applications, such as document analysis and product description generation [225]. However, Qwen-VL faces challenges that include substantial fine-tuning requirements for specialized tasks to achieve optimal outcomes. While hierarchically structured, the design creates computational inefficiencies during large-scale or complex operations. Furthermore, its dependence on high-quality multimodal data further limits effectiveness in data-scarce environments, resulting in performance degradation.

## 4.5. ChatGPT-4

Launched by OpenAI in 2024, ChatGPT-4 is an advanced multimodal model capable of interpreting both textual and visual information within a unified framework. It introduces improved self-attention mechanisms and adaptive scaling significantly enhancing its capability to handle high-dimensional multimodal data. ChatGPT-4 shows strong performance in domains such as diagram interpretation, multimodal dialogue, and VQA. Leveraging transfer learning after training on extensive image-text corpora, the model adapts to specialized applications with minimal labeled data [226]. Despite its strengths, it has some limitations. Its capabilities in detailed image analysis are less refined compared to domain-specific vision models. Furthermore, its deployment at scale involves considerable computational costs, causing challenges for widespread adoption. Another concern is the model lack of interpretability in complex decision-making processes, which raises concerns regarding its reliability in sensitive fields like healthcare or legal contexts.

## 4.6. Recent Innovations in VLMs (Post-2023)

Recent developments in vision-language models (VLMs) post-2023 have significantly advanced the capabilities and efficiency of multimodal understanding systems. This section provides a comparative overview of contemporary models such as IDEFICS [227], GPT-4V [228], Claude 3 Opus [229], Gemini 2.0 [230], PaliGemma 2 [12], Qwen 2.5 VL [231], Moondream [232], and others, highlighting their advancements over earlier foundational models like CLIP [5], ViLBERT [65], VisualBERT [233], Flamingo [9], SimVLM [234], and ALIGN [235]. Innovations are evident in architecture, pretraining strategies, and fine-tuning methodologies, as well as in the use of training data and scale for advancing multimodal tasks.

### *4.6.1. Architectural Innovations*

Recent VLMs have introduced sophisticated architectural enhancements to better integrate vision and language. For example, models such as BLIP-3, Gemini, and Claude 3 leverage advanced cross-attention and sparse transformer mechanisms, along with Mixture-of-Experts (MoE) designs, to enable efficient large-scale training. In contrast, earlier models like ViLBERT and VisualBERT [233] employed dual-stream or early-fusion transformers that lacked the scalability and precision of recent models. GPT-4o and Claude 3 further push the frontier with decoder-style multimodal transformers capable of real-time multimodal interaction, including audio and video.

### *4.6.2. Training Scale and Datasets*

Another significant shift is the use of more extensive, diverse datasets. While pre-2023 models such as CLIP and ALIGN relied on large but noisy image-text datasets (e.g., web-scale English-centric corpora), post-2023 models like Gemini, PaliGemma 2, and Qwen 2.5 VL are trained on highly curated multilingual and synthetic datasets. These include vision-language pairs across image, video, audio, and document modalities, enabling better generalization and real-world applicability.

### *4.6.3. Fine-Tuning Flexibility*

Modern VLMs emphasize parameter-efficient fine-tuning methods. While older models such as Flamingo and SimVLM required full-model fine-tuning or were few-shot prompt-based, newer models support modular architectures. Models like IDEFICS, Gemini, and DeepSeek Janus incorporate adapters, LoRA, and tool-calling interfaces to allow more efficient and flexible adaptation. Claude 3 and GPT-4o further integrate reinforcement learning with human feedback (RLHF) and instruction tuning to personalize model responses and control alignment.

### *4.6.4. Pretraining Objectives and Backbone Integration*

Pre-2023 models were typically pretrained on contrastive image-text alignment (as in CLIP and ALIGN), masked language modeling (VisualBERT), or region-level supervision (SimVLM), and relied on distinct vision and language backbones. While contrastive dual-encoder designs enabled impressive zero-shot performance, these approaches were largely limited to image-text pairs and required vast datasets for generalization. In contrast, post-2023 VLMs often employ joint multimodal pretraining objectives that combine instruction tuning, generative modeling, and unified representation learning. Recent models such as Gemini 2.0 and GPT-4V leverage large-scale vision-language-aligned corpora and seamless backbone integration, fusing vision transformers (ViT) with language models to enable richer cross-modal interactions and enhanced downstream versatility.

Table 7: Comparative Overview of Vision-Language Models (Pre-2023 vs Post-2023)

| **Model** | **Year** | **Fine-Tuning Strategy** | **Architecture** | **Pretraining Objective** | **Pretrained Backbone Model** | **Vision Encoder / Tokenizer** | **Parameters** | **Training Data** | **Key Innovations** |
|---|---|---|---|---|---|---|---|---|---|
| ViLBERT [65] | 2019 | Full fine-tuning | Two-stream Transformer | Image-text alignment with co-attentional modules | BERT | Object-based features (Faster R-CNN) | 110M | COCO, Conceptual Captions | Co-attentional streams for image-text fusion |
| VisualBERT [233] | 2019 | Full fine-tuning | Single-stream Transformer | Masked language modeling with visual embeddings | BERT | Object-based features (Faster R-CNN) | 110M | COCO, VQA | Shared encoder for text and image inputs |
| CLIP [5] | 2021 | Zero-shot inference | Encoder-decoder | Contrastive image-text learning | Pretrained from scratch | ViT/ ResNet | 63M-355M | 400M web image-text pairs | Contrastive learning with dual encoders; broad generalization via natural language supervision |
| ALIGN [235] | 2021 | Zero-shot inference | Dual encoder (EffNet-L2 + Transformer) | Contrastive image-text alignment | EfficientNet-L2 | EfficientNet | 1.8B | 1B+ noisy web image-text pairs | Large-scale noisy training with CLIP-style contrastive objectives |
| SimVLM [234] | 2021 | Full fine-tuning | Unified Transformer encoder-decoder | Prefix language modeling (unified vision-text sequence) | BERT + ResNet | Unified Transformer | 1B+ | Vision-language pairs | Simplified architecture with prefix modeling and no region-level supervision |
| Florence [236] | 2022 | Full fine-tuning | Unified Transformer | Unified encoder with multi-task supervision | Swin Transformer | Swin | 892M | Multilingual web-scale dataset | High-performance universal VL encoder |

| Model | Year | Fine-Tuning Strategy | Architecture | Pretraining Objective | Pretrained Backbone Model | Vision Encoder / Tokenizer | Parameters | Training Data | Key Innovations |
|---|---|---|---|---|---|---|---|---|---|
| Flamingo [9] | 2022 | Few-shot in-context learning | Perceiver Resampler + Decoder-only Transformer | Frozen vision-language backbones with trainable cross-attention | Chinchilla | ViT-L/14 + Perceiver | 80B | M3W, ALIGN | In-context few-shot learning with frozen backbones and cross-modal fusion |
| BLIP-2 [237] | 2023 | Modular fine-tuning via Q-Former | Vision encoder + frozen LLM + Q-Former | Two-stage: vision-text + vision-to-language generation | ViT-G + OPT/FlanT5 | ViT-G + Q-Former | 223M-400M | WebLI, COCO, CC3M, CC12M | Q-Former for modular downstream tasks |
| IDEFICS [227] | 2023 | Parameter-efficient tuning | Unified Transformer with vision encoder | Instruction-tuned vision-language | OPT + ViT | ViT | 80B | COCO, VQAv2, A-OKVQA | Open-source instruction-following VLM |
| PaliGemma 2 [12] | 2024 | LoRA, fine-grained adapters | Transformer encoder-decoder | Multilingual + synthetic datasets | Gemma + ViT | ViT | - | Synthetic + real data e.g (DOCCI, LAION, CC12M) | Multilingual generation + grounding |
| Gemini 2.0 [230] | 2024 | Modular fine-tuning | PaLM-based encoder-decoder + vision module(custom) | Multimodal pretraining with sparse transformers | PaLM 2 + ViT | ViT | - | Multilingual, synthetic corpus | Flexible and efficient multimodal reasoning |
| Kosmos-2.5 [238] | 2024 | Selective fine-tuning (frozen ViT, tuned resampler + decoder) | Decoder-only Transformer with ViT + resampler | Document text recognition + image-to-Markdown generation | - | ViT-G/14, ViT-L/14 | 1.3B | Document images, OCR, structured markup data | Layout-aware multimodal literacy via visual-text fusion with Markdown generation |

| Model | Year | Fine-Tuning Strategy | Architecture | Pretraining Objective | Pretrained Backbone Model | Vision Encoder / Tokenizer | Parameters | Training Data | Key Innovations |
|---|---|---|---|---|---|---|---|---|---|
| GPT-4V [228] | 2024 | No tuning (chat interface) | Unified Transformer with vision-text fusion | Text + image pretraining | GPT-4 | Custom ViT-like encoder | - | Vision-language aligned corpus | GPT-4 vision support with image-text joint encoding |
| Claude 3 Opus [229] | 2024 | Supervised fine-tuning via API | Encoder-decoder transformer | Proprietary encoder-decoder | Proprietary | - | - | Multimodal benchmarks | Safe and high-performance multimodal chat |
| LongVILA [239] | 2024 | Efficient parameter tuning | Video-based encoder-decoder transformer | Video-language transformer | Custom video model | Patch + frame tokenizer | - | Long video, image sequences | Long-context video QA and interleaved image-text reasoning |
| Molmo [240] | 2024 | Instruction tuning | Encoder-decoder transformer | Transformer-based VLM | - | ViT-L/14 (CLIP) | 72B | Open PixMo data | Open-source transparent training |
| Qwen 2.5 VL [231] | 2025 | Instruction tuning | Transformer decoder with visual patch input | Vision transformer + LLM fusion | Qwen 2.5 + ViT | ViT | 3B/7B/72B | Docs, images, audio | OCR + document QA specialization |
| DeepSeek Janus [241] | 2025 | Adapter-based fine-tuning | Dual-stream Transformer with MoE | Multimodal instruction-following | DeepSeek + ViT | ViT | 7B | Instruction + synthetic datasets | Efficient MoE-based dual-stream VLM |
| MiniCPM-o 2.6 [242] | 2025 | Plug-in modules + instruction tuning | Modular lightweight Transformer | Multimodal instruction-following + OCR | MiniCPM + LLaMA3 | Vision adapter | 8B | Instruction-tuned corpus | GPT-4V-level OCR + real-time video understanding on-device |
| Moondream [232] | 2025 | Minimal fine-tuning | Decoder-only Transformer | Multimodal pretraining | - | Compact encoder | 1.86B | Open efficient datasets | Small footprint with privacy focus |

| Model | Year | Fine-Tuning Strategy | Architecture | Pretraining Objective | Pretrained Backbone Model | Vision Encoder / Tokenizer | Parameters | Training Data | Key Innovations |
|---|---|---|---|---|---|---|---|---|---|
| Pixtral [243] | 2025 | Instruction tuning | Dual-stream compact transformer | Mistral-style ViT + LLM | Mistral + ViT | ViT | 12B | Multi-domain open-source corpus | ViT fusion in compact architecture |

These advancements collectively signify a paradigm shift in the field of vision-language understanding. **Table 7** summarizes the major VLMs developed before and after 2023, highlighting their architectural differences, fine-tuning mechanisms, training corpus, and performance. A comprehensive summary of Vision Language Model training configurations, including batch sizes, learning rates, hardware, and training strategies, is provided in **Appendix A 13**.

## 5. Datasets

Datasets play a crucial role to advance VLM, integrating visual and textual modalities. VLM are useful for tasks such as image captioning, VQA, image-text retrieval, and multimodal reasoning. Datasets used to train VLM consist of image-text pairs where a comprehensive text description or label is associated with each image supporting the complete understanding. In this section, we categorize datasets based on tasks they support. Additionally, we highlight challenges and ethical considerations associated with data collection. Before diving into task-specific categories, **Table 8** provides a high-level dataset audit summarizing key Vision-Language datasets in terms of size, modalities, linguistic diversity, and known limitations. **Table 9** outline a detailed overview of various datasets, organized by type, description, and applications. Furthermore, **Table 10** specifically gives the details of the VLM dataset used in the medical domain. A sample of images from various datasets are shown in **Figure 11, 12** and **13**.

General purpose datasets such as MS COCO [69] and VQAv2 [244] etc, focus on everyday scenes with object-centric or question-answering tasks in natural images, domain-specific datasets like RadGraph [245] and PMC-OA [246] significantly differ in scope, structure, and annotation requirements. RadGraph [245] consists of structured annotations from radiology reports, targeting clinical understanding and requiring domain expertise for accurate interpretation. Similarly, PMC-OA [246] comprises image-text pairs extracted from biomedical literature, covering a broad range of medical procedures and findings. Unlike conventional datasets, both RadGraph [245] and PMC-OA [246] demand high annotation quality, are limited to specialized domains, and introduce challenges related to data sensitivity, privacy, and expert-driven curation.

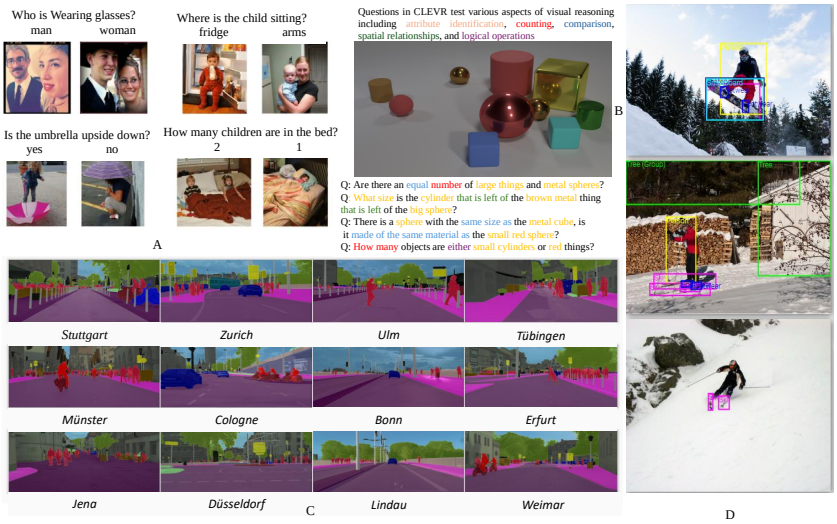



**Figure 11:** Image A illustrates a sample from the VQA dataset [244], while Image B presents an example from the CLEVR dataset. The associated questions in CLEVR are crafted to assess various aspects of visual reasoning, such as attribute identification, counting, comparative analysis, multi-focus attention, and logical operations [247]. Image C showcases a sample from the Cityscapes dataset [248], while Image D displays a sample from the Open Images dataset [249].

Table 8: Dataset Audit Table: Overview of key datasets used in Vision-Language research

| Dataset | Size | Modalities | Language(s) | Category Diversity | Known Biases / Limitations |
|---|---|---|---|---|---|
| MS COCO [69] | 328K images | Image-Text | English | 91 object categories | Western-centric content; limited cultural diversity; object-centric focus |
| VQAv2 [244] | 204K images; 1.1M Q&A pairs | Image-QA | English | Everyday scenes with varied Q&A | Language bias; answer priors; question redundancy |
| RadGraph [245] | 221K reports; 10.5M annotations | Text (Radiology reports) | English | Radiology findings | Domain-specific; requires medical expertise for annotation; limited to chest X-rays |
| GQA [250] | 113K images; 22M questions | Image-QA | English | Compositional reasoning | Synthetic question generation; potential over-reliance on scene graphs |
| GeoBench-VLM [16] | 10K+ tasks | Satellite-Text | English | Natural disasters, terrain, infrastructure | Sparse labels; coverage gaps |
| SBU Captions [251] | 1M images | Image-Text | English | Web-sourced everyday scenes | Noisy captions; duplicate entries |
| MIMIC-CXR [252] | 377K images; 227K studies | Image-Text | English | Chest X-rays | Hospital-centric; privacy restrictions |
| EXAMS-V [253] | 20,932 questions | Mixed Multimodal | 11 languages | Exam-style reasoning across disciplines | Regional bias; multilingual challenge |
| RS5M [254] | 5M images | Satellite-Text | English | Remote sensing imagery | Sparse labels; class imbalance; varying image quality |
| VLM4Bio [255] | 30K instances | Image-Text-QA | English | Biodiversity, taxonomy | Domain-specific; taxonomic bias; limited generalizability |
| PMC-OA [246] | 1.65M image-text pairs | Image-Text-QA | English | High diversity within the biomedical domain; Covers a wide range of diagnostic procedures; disease types, and medical findings; | Caption noise, Requires medical expertise; |
| WebLI-100B [256] | 100 Billion image-text pairs | image-text | 100+ languages | Global content | Cultural/geographic bias, noisy data |

## 5.1. VLM Dataset

VLM depend on diverse datasets adapted to different tasks, ranging from classification and detection to reasoning and retrieval etc,. The following is an overview of the most commonly used dataset categories for training VLM.

### *5.1.1. Detection Datasets*

Detection datasets are vital for tasks like object detection, segmentation, and scene understanding, providing detailed annotations for visual objects.

**COCO (Common Objects in Context):** COCO is a widely used dataset that contains over 330,000 images, annotated with 2.5 million object instances comprising 80 categories. It supports object detection, segmentation, and image captioning tasks [69].

**Applications:** Object detection, segmentation, image captioning, visual question answering.

**Open Images:** A massive dataset with more than 9 million images, annotated with bounding boxes and object labels in 600 categories, ideal for training models in detection and multi-label classification [249].

**Applications:** Object detection, multi-label classification, instance segmentation.

### *5.1.2. Classification Datasets*

Classification datasets are applied to train models on image classification, object recognition and related tasks, where images are classified into different classes.

**ImageNet:** One of the largest datasets for image classification which consists of more than 14 million images across 21,000 categories. It is commonly employed for object recognition and image-text alignment tasks [257].

**Applications:** Image classification, object recognition, image-text alignment.

**Visual Genome:**A dataset with over 100,000 images annotated with object level information and relationships between objects. It is valuable for tasks such as scene graph generation and captioning.

**Applications:** Image-text alignment, scene graph generation, object relationships.

### *5.1.3. Segmentation Datasets*

Segmentation datasets focus on pixel-level tasks, such as semantic segmentation and instance segmentation, where the goal is to assign each pixel of an image to a specific category.

**ADE20K:** A semantic segmentation dataset with 20,000 images, covering various scene categories. It includes dense annotations for both objects and parts, making it useful for detailed scene understanding [258].

**Applications:** Semantic segmentation, scene parsing, image segmentation.

**Cityscapes:** This dataset focuses on metropolitan street scenes, providing high-quality pixel-level annotations for semantic segmentation tasks. It contains more than 5,000 images from 50 different cities [248].

**Applications:** Urban scene analysis, autonomous driving, semantic segmentation.

### *5.1.4. Text-to-Image Generation Datasets*

These datasets are critical for tasks involving the generation of images from textual descriptions, as well as for multimodal retrieval and understanding.

**Flickr30k:** This dataset contains 31,000 images and each has five captions which is commonly used for image captioning and text-to-image generation tasks [259].

**Applications:** Image captioning, text-to-image generation, multimodal retrieval.

**COCO Captions:** It is the subset of the COCO dataset, which includes more then 300,000 images and each paired with five textual descriptions. It is one of the largest datasets utilized for image captioning and text-to-image synthesis [69].

**Applications:** Image captioning, text-to-image synthesis, multimodal understanding.

### *5.1.5. Multimodal Alignment Datasets*

These datasets are utilized for tasks that require matching images with corresponding textual descriptions, labels or questions.

**VQA Dataset** A dataset for visual question answering that contains more than 200,000 questions related to 100,000 images. It is used to assess multimodal reasoning and understanding [260].

**Applications:** multimodal reasoning, visual question answering.

**MS-COCO Text-Only:** A subset of the MS-COCO dataset, specifically curated for text-based retrieval tasks, comprises images systematically paired with descriptive captions, facilitating research in multimodal information retrieval and analysis [69].
**Applications:** Text-based retrieval, text-to-image matching, textual understanding of visual content.

#### *5.1.6. Vision-Language Pre-training Datasets*

These datasets are predominantly used for the pre-training of VLM, allowing them to effectively learn and comprehend the relationship between image-caption pairs on a large scale.
**Conceptual Captions:** This extensive dataset comprises 3.3 million image-caption pairs which is collected from online sources, rendering it highly suitable for the pre-training of VLM in tasks related to image-caption generation [261].
**Applications:** image-caption generation, Pre-training Vision-Language models, multimodal pre-training.
**SBU Captions:** A dataset featuring 1 million image-caption pairs, also gathered from the web, which spans diverse domains and is employed for the pre-training of Vision-Language Models [262].
**Applications:** Pre-training, caption generation, multimodal learning.

Table 9: Datasets for Vision-Language Models

| Dataset Type | Dataset Name | Description | Applications |
|---|---|---|---|
| **Detection Datasets** | COCO [69] | Contains 330k images with annotations for object detection and segmentation. | Object detection, instance segmentation, image captioning. |
| | Open Images [249] | A large-scale dataset with over 9 million annotated images for object detection. | Object detection, image captioning, visual relationship detection. |
| **Classification Datasets** | ImageNet [257] | A large dataset with over 14 million labeled images across 1000 classes. | Image classification, transfer learning, visual recognition. |
| | Visual Genome [70] | A dataset containing 108k images annotated with object detection and scene graphs. | Object detection, visual question answering, scene understanding. |
| **Segmentation Datasets** | ADE20K [258] | A semantic segmentation dataset with 20k images across 150 categories. | Semantic segmentation, scene parsing, object localization. |
| | Cityscapes [248] | A dataset for urban scene understanding with high-quality annotations for segmentation. | Semantic segmentation, autonomous driving, road scene understanding. |
| **Text-to-Image Generation Datasets** | Flickr30k [259] | Contains 31k images with five captions per image, often used for image captioning. | Image captioning, text-to-image generation, multimodal retrieval. |
| | COCO Captions [69] | A subset of COCO with over 300,000 images and 5 textual descriptions per image. | Image captioning, text-to-image synthesis, multimodal understanding. |
| **Multimodal Alignment Datasets** | VQA [260] | A dataset for visual question answering with 200k+ questions across 100k images. | Visual question answering, multimodal reasoning, image-text understanding. |

| Dataset Type | Dataset Name | Description | Applications |
|---|---|---|---|
| | EndoVis-18-VLQA [263] | A dataset for vision-language question answering (VLQA) in surgical and endoscopic procedures. It contains video frames and question-answer pairs focused on tasks like identifying tools and recognizing actions during surgeries | Multimodal Question Answering, Medical Applications, Surgical Assistance. |
| | VLM4Bio [255] | The VLM4Bio dataset evaluates the performance of 12 SOTA Vision-Language Models in answering biologically relevant questions. It contains 469K question-answer pairs and 30K images of fishes, birds, and butterflies, covering five biological tasks. | Multimodal Question Answering, Biodiversity research, Image-Based Scientific Discovery. |
| | MS-COCO Text-Only [69] | A subset of MS-COCO focusing on text-only tasks with descriptive captions. | Text-based retrieval, text-to-image matching, textual understanding of visual content. |
| **Vision-Language Pre-training Datasets** | Conceptual Captions [261] | A dataset with 3.3 million image-caption pairs sourced from the web. | Pre-training Vision-Language models, multimodal pre-training, image-caption generation. |
| | PathQA Bench Public [264] | The dataset consists of 456,916 instruction-response pairs, including multi-turn conversations, multiple-choice questions, and short answers, specifically curated for pathology. The data was used to train PathChat, a vision-language AI assistant for human pathology | Pathology Education, models, multimodal pre-training, Clinical Decision Support. |
| | SBU Captions [262] | Contains 1 million image-caption pairs collected from the web. | Pre-training, caption generation, multimodal learning. |
| **Multimodal Retrieval Datasets** | Flickr30k Entities [265] | An extended version of Flickr30k, with object-level entity annotations. | Object detection, image-text retrieval, image-caption alignment. |
| | RS5M [254] | The RS5M dataset contains 5 million remote sensing images with English descriptions, created by filtering and captioning existing datasets. It bridges the gap between general pre-trained Vision-Language Models (VLMs) and domain-specific remote sensing tasks | Remote Sensing Classification, Cross-Modal Retrieval, Semantic Localization,Domain-Specific Fine-Tuning. |

| Dataset Type | Dataset Name | Description | Applications |
|---|---|---|---|
| | Visual Semantic Role Labeling (v-SRL) [266] | Focuses on assigning semantic roles to visual elements of an image. | Semantic role labeling, image-text alignment, multimodal grounding. |
| **Multimodal Reasoning Datasets** | CLEVR [247] | A synthetic dataset for visual reasoning and answering compositional questions. | Visual reasoning, question answering, synthetic data generation. |
| | GMAI-MMBench [267] | GMAI-MMBench is a comprehensive benchmark for evaluating Large Vision-Language Models (LVLMs) in medical applications. It spans 284 datasets across 38 medical image modalities, 18 clinical tasks, 18 departments, and 4 perceptual granularities, organized in a Visual Question Answering (VQA) format. The benchmark enables customizable evaluation through a lexical tree structure, aiming to improve medical AI research | Medical Diagnosis and Treatmen, Clinical AI Evaluation, Benchmarking Medical AI. |
| | NavGPT-Instruct-10k [268] | NavGPT-Instruct-10k is a dataset for training Vision-Language Models (VLMs) in navigational reasoning. It includes 10,000 intermediate navigation steps generated with GPT-4V, where each step provides environment descriptions and directions for the next action, based on equirectangular panoramic images of the agent's surroundings | Navigational Reasoning in AI, Autonomous Systems Training, Pathfinding AI. |
| | GQA [269] | A dataset for evaluating compositional question answering, consisting of 22 million questions. | Visual reasoning, compositional question answering, multimodal reasoning. |

| Dataset Type | Dataset Name | Description | Applications |
|---|---|---|---|
| | EXAMS-V [253] | EXAMS-V is a multi-discipline, multimodal, multilingual benchmark for evaluating vision-language models. It contains 20,932 multiple-choice questions across 20 school subjects in 11 languages, with features like text, images, tables, and diagrams. The dataset, sourced from diverse education systems, requires advanced reasoning over both textual and visual content, making it a challenging test for models like GPT-4V and Gemini | Multimodal AI Testing, Education Technology, Cross-Language and Cross-Cultural Learning |
| | MMVP-VLM [15] | MMVP-VLM evaluates CLIP-based models on matching image-text pairs representing visual patterns (e.g., color, shape, spatial relationships). It includes 15 pairs per pattern, drawn from the MMVP dataset but simplified for easier language understanding | Visual reasoning, compositional question answering, multimodal reasoning. |
| **Semantic Segmentation and Instance Segmentation Datasets** | ADE20K [258] | 20k images annotated for semantic segmentation across multiple scene categories. | Semantic segmentation, scene understanding, instance segmentation. |
| | Cityscapes [248] | Urban scene dataset with pixel-level annotations for semantic segmentation. | Urban scene analysis, autonomous driving, semantic segmentation. |
| **Cross-Modal Transfer Datasets** | MIMIC-CXR [252] | Large dataset of chest radiographs paired with radiology reports. | Medical image analysis, multimodal health informatics, cross-modal learning. |
| | MedNLI [270] | A dataset for natural language inference in the medical domain. | Medical text understanding, NLP for healthcare, cross-modal reasoning. |

Table 10: Overview of the VLM datasets for Medical domain

| Dataset Name | Image-Text pairs | QA pairs | Description | Application |
|---|---|---|---|---|

| Dataset Name | Image-Text pairs | QA pairs | Description | Application |
|---|---|---|---|---|
| **VQA-Med 2020** [271] | ✗ | ✓ | VQA-Med-2020 is a dataset for Visual Question Answering (VQA) and Visual Question Generation (VQG) tasks in the medical domain, specifically for radiology images. It includes training, validation, and test sets for answering questions about medical abnormalities and generating questions from images captioning. | Medical Diagnosis,Clinical Decision Support, Multimodal Question Answering |
| **ROCO** [272] | ✓ | ✗ | The Radiology Objects in Context (ROCO) dataset is a large-scale, multimodal medical imaging dataset sourced from PubMed Central Open Access. It contains radiology and non-radiology images, each accompanied by captions, keywords, and UMLS (Unified Medical Language System) Semantic Types and Concept Unique Identifiers (CUIs) | Image Captioning, Image Classification & Tagging, Content-Based Image Retrieval, Medical Visual Question Answering, Multimodal Retrieval |
| **VQA-Med 2019** [273] | ✗ | ✓ | VQA-Med-2019 is a dataset for Visual Question Answering (VQA) focused on radiology images, featuring 3,200 images with 12,792 question-answer pairs. It includes four categories: Modality, Plane, Organ system, and Abnormality. The dataset is designed for questions that can be answered directly from the images without external medical knowledge | Medical Image Analysis, Radiology AI Development, Medical Educational Tools, Medical VQA' |
| **MIMIC-NLE** [274] | ✓ | ✗ | The MIMIC-CXR-JPG dataset is a public collection of 377,110 chest X-ray images in JPG format, with structured labels from 227,827 free-text radiology reports | Medical Image Understanding, Natural Language Processing for Radiology, Decision Support Systems. |
| **SLAKE** [275] | ✗ | ✓ | SLAKE is a bilingual dataset for medical Visual Question Answering (Med-VQA), featuring semantic labels annotated by physicians and a new medical knowledge base. It covers more human body parts and richer modalities than existing datasets. | Visual Annotations, Diverse Questions,Knowledge-Based Medical AI |

| Dataset Name | Image-Text pairs | QA pairs | Description | Application |
|---|---|---|---|---|
| **GEMeX** [276] | ✓ | ✓ | GEMeX is the largest chest X-ray Med-VQA dataset, designed to support diverse question types and enhance explainability in medical VQA systems. It is consider to be the first to incorporate multimodal explainability, aiming to improve the visual reasoning ability of Large Vision-Language Models through fine-tuning | Medical Visual Question Answering Visual Reasoning in Healthcare AI, Explainable AI in Medical Imaging |
| **MS-CXR** [277] | ✓ | ✗ | The MS-CXR dataset supports semantic modeling in biomedical vision-language processing, offering 1,162 image-sentence pairs with bounding boxes and phrases across eight cardiopulmonary radiological findings. It icludes both reviewed (1,026 pairs) and manually labeled (136 pairs) annotations | Biomedical Vision-Language Processing, Image-Text Reasoning in Medical AI, Radiology Image Annotation and Analysis, Contrastive Learning in Vision-Language Models, Semantic Modeling in Medical Imaging, |
| **MedICaT** [278] | ✓ | ✗ | MedICaT is a comprehensive medical image dataset containing 217,060 figures from 131,410 open-access papers, along with captions, subfigure annotations, and inline textual references. Sourced from PubMed Central and supplemented with text from S2ORC, it provides rich multimodal data for training and evaluating models in medical image understanding, captioning, and information retrieval. | Medical Image Captioning,Multimodal Learning,Information Retrieval. |
| **3D-RAD** [279] | ✗ | ✓ | 3D-RAD is a large-scale dataset for 3D Medical VQA using **over 4,000 radiology CT scans** and **more than 12,000 QA pairs**. It encompasses diverse VQA tasks, including anomaly detection, medical computation, existence detection, and multi-temporal analysis, addressing limitations of 2D Med-VQA. | 3D Medical VQA, Multi-temporal Diagnosis, Clinical Decision Support, 3D Medical Image Understanding. |

| Dataset Name | Image-Text pairs | QA pairs | Description | Application |
|---|---|---|---|---|
| **ImageCLEFmed-MEDVQA-GI** [280] | ✓ | ✓ | This challenge dataset focuses on integrating VQA with synthetic gastrointestinal (GI) data (e.g., endoscopy images) to enhance diagnostic accuracy. The 2025 iteration includes **over 10,000 endoscopic images with more than 30,000 QA pairs** (including synthetic). | Gastrointestinal Image Analysis, Synthetic Data Generation, Endoscopy VQA, Diagnostic AI Enhancement. |
| **BIOMEDICA** [281] | ✓ | ✗ | BIOMEDICA is a scalable, open-source framework and archive derived from the PubMed Central Open Access subset, containing over 24 million unique image-text pairs from over 6 million biomedical articles. It covers a wide range of disciplines (pathology, radiology, ophthalmology, etc.) for generalist biomedical VLM pre-training. | Biomedical VLM Pre-training, Image-Text Retrieval in Scientific Literature, General Medical AI Development across diverse modalities. |
| **PMC-OA** [282] | ✓ | ✗ | PMC-OA (PubMed Central Open Access) is a large-scale biomedical dataset with over 1.6 million image-caption pairs extracted from PubMed Central Open Access articles. It supports multimodal learning and has been used to train models like PMC-CLIP for tasks such as image-text retrieval and classification. | Biomedical Image-Text Retrieval, Medical Image Classification, Multimodal Learning. |
| **ReasonMed** [283] | ✗ | ✓ | ReasonMed offers 370K samples for complex medical reasoning and Visual Question Answering, generated from multi-agent Chain-of-Thought (CoT) paths to ensure high-quality, explainable answers. | Medical Reasoning, Complex VQA, Clinical Decision Support, Explainable Medical AI. |
| **Lingshu** [284] | ✓ | ✓ | Lingshu is a large-scale medical VLM dataset, unifying 9.3M samples from over 60 existing datasets, designed for diverse tasks including VQA, report generation, and medical consultation. | Generalist Multimodal QA, Medical Report Generation, AI-powered Medical Consultation, Broad Medical VLM Development. |

| Dataset Name | Image-Text pairs | QA pairs | Description | Application |
|---|---|---|---|---|
| GMAI-VL-5.5M [285] | ✓ | ✓ | GMAI-VL-5.5M is a large-scale medical image-text dataset with 5.5 million pairs, created by merging and aligning various existing medical datasets, suitable for training general medical AI models. | General Medical AI, Medical Diagnosis, Multimodal QA, Clinical Decision Support Systems. |

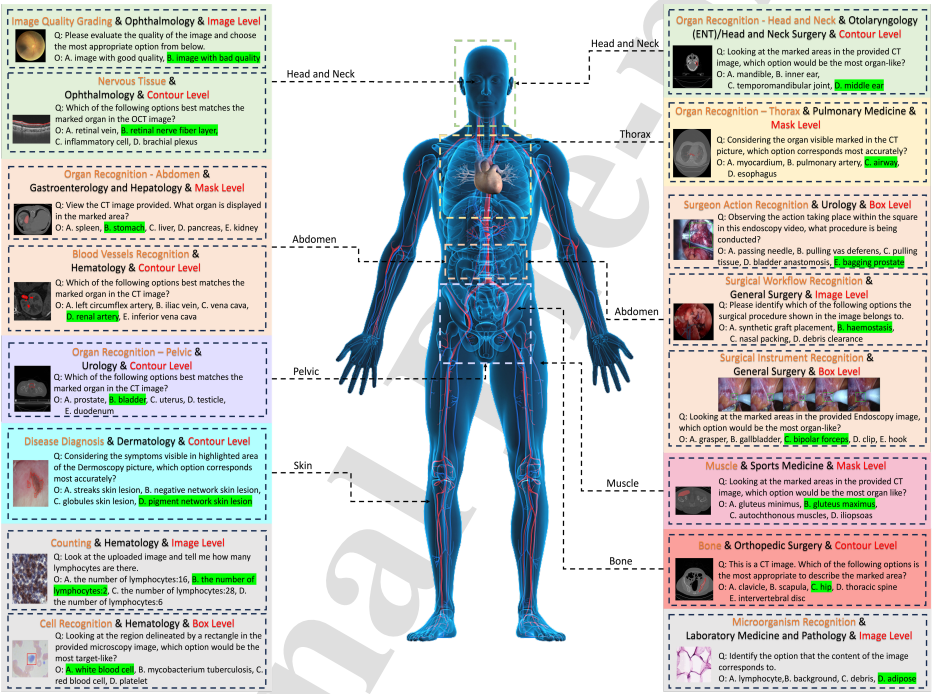



**Figure 12:** Sample of the GMAI Dataset (Medical related) [267].

### *5.1.7. Multimodal Retrieval Datasets:*

These datasets support tasks involving the retrieval of images or text based on queries, promoting enhanced alignment between visual and textual modalities.

**Flickr30k Entities:** An extension version of the Flickr30k dataset, incorporating object-level entity annotations alongside captions, enabling precise retrieval tasks [265].

**Applications:** Object detection, image-caption alignment, image-text retrieval.

**Visual Semantic Role Labeling (v-SRL):** This dataset emphasizes the assignment of semantic roles to visual components within images, aiding in multimodal language grounding [266].

**Applications:** Semantic role labeling, image-text alignment, multimodal reasoning.

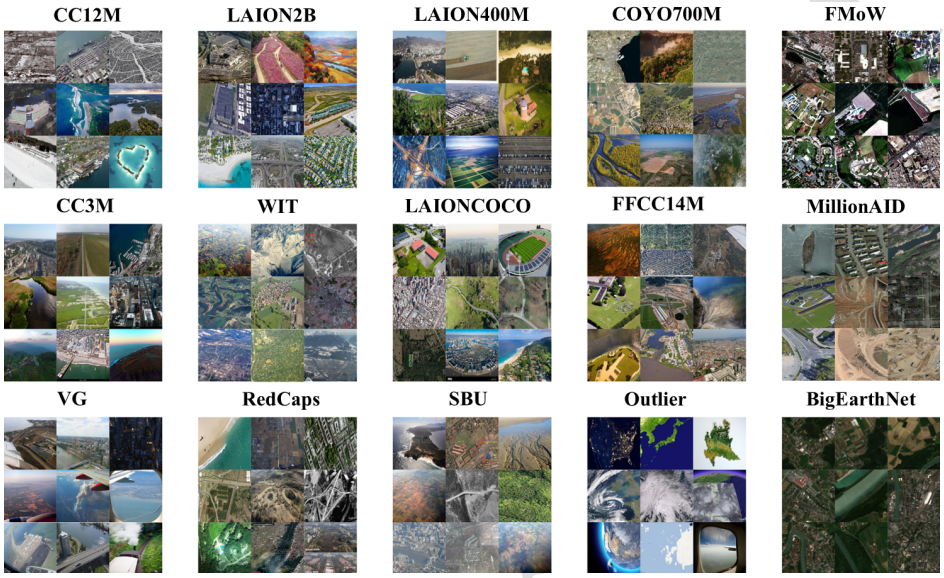



**Figure 13:** Data Sample of the RS5M (Remote sensing Dataset) [254].

#### *5.1.8. Multimodal Reasoning Datasets:*

These datasets are tailored for assessing multimodal reasoning capabilities, where models are required to draw inferences from both visual and textual inputs.

**CLEVR:** A synthetically generated dataset crafted for visual reasoning, particularly for answering questions about scenes that demand compositional reasoning [247].

**Applications:** Visual reasoning, question answering, synthetic data generation.

**AltChart** The AltChart dataset which consists of 10,000 real chart images with semantically rich summaries, enhances Human-Computer Interaction by improving accessibility and user experience for visually impaired individuals through detailed chart summarization [286].

**Applications:** Visual reasoning, question answering, chart summarization.

**GQA:** A dataset consisting of 22 million questions about images, designed to evaluate compositional question answering over natural images [269].

**Applications:** Visual reasoning, compositional question answering, multimodal reasoning.

### 5.2. Empirical Benchmarking of VLMs

**Table 11** provides a comparative evaluation of SOTA Vision-Language Pre-training models (VLP) across three main multimodal tasks: image captioning, VQA, and image retrieval. Performance is evaluated by applying widely recognized benchmarks datasets including VQA 2.0, MS-COCO, and Flickr30K and standard evaluation metrics such as BLEU-4, CIDEr, METEOR, and SPICE for image captioning, accuracy for VQA, and Recall@1 (R@1) for retrieval. These metrics collectively capture syntactic accuracy, semantic richness, and retrieval precision, allowing a detailed analysis of cross-model capabilities.

VLM models such as Unified VLP and VinVL exhibit strong results in both caption generation and VQA, using integrated vision-language architectures optimized for joint generation and comprehension. In contrast, contrastive learning-based models like CLIP and DreamLIP achieve high performance in retrieval tasks, benefiting from large-scale pretraining on diverse image-text pairs. Architectures like BLIP and BLIP-2 advance the state-of-the-art by utilizing modular design strategies and bootstrapped pretraining techniques. Notably, BLIP-2 attains leading results on both captioning (BLEU-4 score of 43.7) and VQA (accuracy of 79.3%), while maintaining parameter efficiency relative to larger-scale models such as Flamingo.

Furthermore, models including FIBER and SimVLM prioritize architectural efficiency and task generalization, adopting strategies such as cross-modal fusion within the backbone and prefix-based language modeling. Emerging frameworks such as VILA and NLIP underscore the growing focus on robustness, noise resistance, and multilingual adaptability. Collectively, the performance trends across these models demonstrate a trajectory toward scalable, modular, and computationally efficient VLP systems tailored for implementation in diverse and resource-constrained environments.

Table 11: Comparison of models across Image Captioning, VQA, and Retrieval tasks

| **Model** | **Task** | **Dataset** | **Metric** | **Score** |
|---|---|---|---|---|
| Unified VLP [287] | Image Captioning | COCO, Flickr30K | BLEU-4 / CIDEr | 36.5 / 116.9 (COCO), 30.1 / 67.4 (Flickr) |
| VinVL [288] | Image Captioning | COCO | BLEU-4 / CIDEr | 40.9 / 140.9 |
| SimVLM [289] | Image Captioning | COCO | BLEU-4 / CIDEr | 40.3 / 143.3 |
| BLIP [34] | Image Captioning | COCO | BLEU-4 / CIDEr | 41.7 / 143.5 |
| RegionCLIP [290] | Image Captioning | COCO | BLEU-4 / CIDEr | 40.5 / 139.2 |
| BLIP-2 [237] | Image Captioning | COCO, NoCaps | BLEU-4 / CIDEr | 43.7 / 123.7 (COCO), – (NoCaps) |
| FIBER [291] | Image Captioning | COCO | CIDEr | 42.8 |
| NLIP [292] | Image Captioning | Flickr30k | CIDEr | 135.2 |
| LCL [293] | Image Captioning | COCO | CIDEr | 87.5 |
| Unified VLP [287] | VQA | VQA 2.0 | VQA Score | 70.3% |
| VinVL [288] | VQA | VQA 2.0 | VQA Score | 76.6% |
| FewVLM [124] | VQA | VQA 2.0 | VQA Score | 51.1% |
| SimVLM [289] | VQA | VQA 2.0 | VQA Score | 24.1% |
| BLIP [34] | VQA | VQA 2.0 | VQA Score | 77.5% |
| BLIP-2 [237] | VQA | VQA 2.0 | VQA Score | 79.3% |
| VILA [165] | VQA | VQA 2.0, GQA | VQA Score | 80.8% (VQA 2.0), 63.3% (GQA) |
| LCL [293] | VQA | VQA 2.0 | VQA Score | 73.4% |
| TCL [294] | Image Retrieval | COCO, Flickr30K | R@1 | 62.3% / 88.7% |
| CLIP [5] | Image Retrieval | COCO, Flickr30K | R@1 | 58.4% / 88.0% |
| NLIP [292] | Image Retrieval | COCO | R@1 | 82.6% |
| Cross-Attn [295] | Image Retrieval | COCO, Flickr30K | R@1 | 67.8% / 88.9% |
| DreamLIP [296] | Image Retrieval | COCO, Flickr30K | R@1 | 58.3% / 87.2% |

## 5.3. Challenges and Open Problems

Datasets are foundational for the development of vision language models, as they directly influence model performance and generalization in various multimodal tasks. However, several key challenges persist in dataset creation, annotation, and utilization [72, 21]. The data imbalance in many datasets leads to models that perform well on common categories but struggle with rare or underrepresented objects. This issue, compounded by the

annotation complexity involved in large-scale multimodal datasets, creates scalability challenges. As the demand for high-quality, large datasets increases, finding efficient methods for annotation without sacrificing quality remains a key open problem. Additionally, ensuring multimodal alignment between images and their textual descriptions is essential, as poor alignment can hinder model training and affect performance in tasks like image captioning or visual question answering. Models need well-aligned data to accurately understand and generate meaningful outputs [297].

Furthermore, domain specificity presents a challenge in terms of model generalization. While datasets like MS-COCO [69] are effective for general tasks, specialized domains such as medical imaging or autonomous driving require distinct datasets that may not be widely available [298]. This highlights the need for domain-specific datasets and methods that enable models to generalize across domains without extensive retraining [299]. Finally, the dynamic nature of real-world data necessitates the creation of continually updated datasets that can evolve over time. In dynamic fields like robotics or autonomous driving, datasets need to adapt to new scenarios and objects, making lifelong learning and dataset evolution a pressing concern [300].

To address these challenges, future research should focus on creating more diverse and unbiased datasets, improving automated annotation techniques, and ensuring the cross-domain generalization of models. Additionally, fine-grained multimodal annotations and methods for handling dynamic datasets will be crucial to push the boundaries of Vision-Language Models. Tackling these open problems will facilitate the development of more robust, scalable, and ethically sound Vision-Language Models.

## 6. Evaluation Metrics and Benchmarks

This section provides a comprehensive discussion of the most widely used evaluation metrics, benchmarks, and associated challenges.

### 6.1. Evaluation Metrics

The performance of VLM is evaluated using a set of metrics that assess various aspects of multimodal understanding. These metrics are critical to quantify how well VLM processes and generates meaningful results when working with visual and textual data.

- **Accuracy:** Accuracy is one of the most prominent and commonly used metrics, evaluating the percentage of correctly predicted results or matches. In tasks like VQA or Image-Text Matching, accuracy helps evaluate the models ability to generate correct textual responses or match images with appropriate captions [301].
- **BLEU** (Bilingual Evaluation Understudy Score): Originally developed for machine translation task, BLEU is extensively employed to assess image captioning. It calculates the overlap between n-grams in the generated description and reference descriptions. A higher BLEU score indicates a closer match with human-written text. BLEU is especially relevant for tasks where the precise match of words is crucial [302].
- **CIDEr** (Consensus-based Image Description Evaluation): CIDEr is designed to solve some drawbacks of BLEU by considering both the recall and precision of n-grams in generated captions. It is especially useful in image captioning tasks, where several valid descriptions can exist for the same image. CIDEr is based on the consensus of multiple human-generated descriptions, providing a more nuanced evaluation [303].
- **SPICE** (Semantic Propositional Image Caption Evaluation): Unlike BLEU or CIDEr, which focus on n-gram matching, SPICE evaluates the semantic content of a generated caption. It breaks down captions into semantic propositions (e.g., subject-action-object relationships) and compares them with human-provided annotations. This metric is aligned with understanding the meaning and intent of the caption, making it valuable for tasks that require deeper understanding, like image captioning and multimodal retrieval [304].

Table 12: Comprehensive Overview of Benchmark Datasets for Vision Language Models

| Dataset Name | Description | Data Information |
|---|---|---|
| SEEDBench Series [305] | Image captioning and multi-modal reasoning tasks. | 73K images/questions |
| VLM4Bio [255] | Evaluates the effectiveness of vision-language models in answering biologically relevant questions using images of fishes, birds, and butterflies across five tasks. | 469K QA pairs, 30K images |
| MM-Vet [306] | Visual reasoning tasks for VLMs | - |
| MMBench Series [307] | MMbenchmark is the collection of datasets. It is designed to evaluate the fine-grained capabilities of Vision-Language Models (VLMs) across multi-modal tasks. | 6.4 k samples |
| MME [308] | Robustness testing for multi-modal tasks. | - |
| MMVet V2 [309] | MM-Vet v2 is a benchmark dataset for integrated multi-modal reasoning, challenging VLMs with diverse tasks in real-world and abstract scenarios to advance vision-language understanding. | 3K tasks with diverse multimodal questions |
| HallusionBench [310] | A diagnostic benchmark for evaluating large vision-language models on entangled language hallucination and visual illusion, focusing on image-context reasoning with challenging yes/no questions. | 254 questions across 69 figure |
| OCRBench [311] | A comprehensive benchmark designed to evaluate VLMs on OCR-related tasks, including Text Recognition, Scene Text-Centric VQA, Document-Oriented VQA, Key Information Extraction, and Handwritten Mathematical Expression Recognition, using 1,000 manually verified question-answer pairs. | 1,000 manually verified QA pairs across 29 datasets |
| VCR Series [312] | A benchmark to evaluate VLMs on restoring partially obscured text within images, leveraging pixel-level hints and contextual cues. Includes 2.11M English and 346K Chinese entities sourced from Wikipedia, offered in easy and hard variants. | 2.46M image-caption pairs across English and Chinese datasets |
| COCO_VAL [69] | Object detection, segmentation, and image captioning. | 330K images, 1.5M object instances, 80 object categories, and 5 captions per image |
| ScienceQA_VAL [313] | Science-based question answering using multi-modal inputs. | 21k multimodal science questions with annotations |
| MTVQA [314] | A multilingual benchmark for evaluating Vision-Language Models on Text-Centric Visual Question Answering (TEC-VQA) tasks. | 6,778 QA pairs, 2,116 images across 9 languages |
| MMStar [315] | An advanced multi-modal benchmark designed to evaluate Vision-Language Models on vision-indispensable tasks, addressing issues of unnecessary visual content and data leakage in current benchmarks. | 1,500 high-quality challenge samples |
| POPE [316] | Post-OCR performance evaluation with reasoning tasks. | - |

Continued on the next page

| Dataset Name | Description | Data Information |
|---|---|---|
| TextVQA_VAL [317] | Text recognition-based VQA tasks. | 28,408 images, 45,336 questions, 453,360 answers |
| ChartQA_TEST [318] | A benchmark designed for visual and logical reasoning over charts, featuring 9.6K human-written questions and 23.1K machine-generated questions, with annotations for chart images, bounding boxes, and data tables. | 33,000+ questions and chart annotations |
| GEOBench-VLM [16] | GEOBench-VLM is a comprehensive benchmark designed to evaluate Vision-Language Models on geospatial tasks like scene classification, temporal analysis, and disaster detection, addressing unique challenges in Earth Observation applications. | 10,000+ tasks across 8 categories |
| NL-EYE [319] | NL-EYE is a benchmark designed to evaluate Vision-Language Models on visual abductive reasoning tasks. | 350 triplets, 6 categories & Plausibility prediction and explanation |
| MMT-Bench_VAL [18] | A comprehensive benchmark to evaluate Vision-Language Models on multimodal tasks requiring visual recognition, reasoning, localization, and planning | 31,325 visual questions across 162 subtasks |
| MMInA [320] | Benchmark for evaluating VLMs and LLM agents on multihop reasoning tasks involving multimodal data (images + text) across multiple domains (e.g., webpages, Wikipedia, shopping). | 1,050 tasks, 6 subfolders |
| MileBench [321] | Benchmarks VLMs on multimodal long-context tasks requiring comprehension and generation, using 6,440 samples with an average of 15.2 images and 422.3 words per sample. | 6,440 samples |

- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation): ROUGE evaluates the overlap of textual units, such as n-grams, word sequences, or sentence pairs, between the generated and reference text. Although it is used primarily for text summarization, ROUGE is increasingly applied to vision language tasks that require textual alignment, such as image captioning and text-to-image retrieval [322]. It measures recall, ensuring the model does not overlook important details in reference captions.

While these metrics have been instrumental in evaluating Vision-Language Models, they also present limitations. For example, BLEU and CIDEr can overemphasize exact word matches and fail to capture the broader semantic meaning of captions. SPICE, on the other hand, provides richer semantic evaluation, but still has room for improvement in accounting for more complex multimodal reasoning.

### *6.1.1. Benchmarks*

The development of robust benchmarks has been essential for evaluating the performance of Vision Language Models across various tasks. These benchmarks typically consist of both standardized datasets and domain-specific tasks, enabling researchers to compare models under consistent conditions.

- **VLM4Bio:** VLM4Bio is a benchmark dataset comprising scientific question-answer pairs designed to assess pretrained Vision Language Models for trait discovery in biological images. It includes images from three taxonomic groups: fish, birds, and butterflies, with approximately 10,000 images in total for each group and 469k QA pairs [255].
- **Visual Question Answering**: VQA is one of the most widely used benchmarks for evaluating Vision-Language Models. It involves answering natural language questions about images, which may require visual reasoning or commonsense knowledge. The VQA dataset, introduced by Antol et al., [260]

contains over 200,000 questions spanning 100,000 images, providing a diverse set of challenges for model evaluation.

- **MS-COCO** (Microsoft Common Objects in Context): The MS-COCO dataset is another central benchmark for evaluating image captioning, object detection, and other vision-language tasks. It consists of over 330,000 images with five human-generated captions per image. It has been used extensively to evaluate models on tasks like image captioning, image retrieval, and object detection. Its diversity and large-scale nature make it an ideal benchmark for testing the generalization ability of Vision-Language Models [69].
- **Flickr30k:** Similar to MS-COCO, the Flickr30k dataset includes 31,000 images with five captions per image, making it a useful benchmark for image captioning and cross-modal retrieval tasks. Its relatively smaller size compared to MS-COCO allows for faster experimentation, while still providing a rich source of evaluation data [259].
- **OK-VQA:** OK-VQA was introduced to test Vision-Language Models in scenarios where external knowledge beyond the image is required to answer questions. This dataset addresses limitations in standard VQA by introducing questions that involve commonsense reasoning and external knowledge. It consists of questions paired with images from MS-COCO [323].

While these benchmarks serve as standard evaluation tools, they have limitations, such as their inability to measure models' general multimodal capabilities. As a result, researchers have started to develop more holistic benchmarks that assess models on a wider range of tasks and reasoning abilities. **Table 12** provides a comprehensive overview of benchmark datasets designed to evaluate Vision Language Models across diverse tasks, including image captioning, multimodal reasoning, geospatial analysis, text recognition, and visual abductive reasoning, highlighting their purposes, data characteristics.

### 6.2. Challenges

Despite the variety of metrics and benchmarks, significant challenges remain in evaluating Vision-Language Models effectively:

- **Limitations of Existing Metrics:** Traditional metrics like BLEU, CIDEr, and SPICE are effective in their respective areas but do not fully capture the richness of multimodal reasoning or generalization across tasks. These metrics often fail to account for the complex interplay between text and visual modalities. Moreover, there is an inherent challenge in measuring subjective aspects of Vision Language Models, such as creativity, coherence, and fluency of generated content using purely automated metrics [324].
- **Need for More Comprehensive Metrics:** To address these limitations, there is a growing need for holistic evaluation frameworks that assess Vision-Language Models from multiple perspectives. These frameworks should incorporate both automated and human evaluations. Human evaluations are critical in tasks like image captioning, where creativity, relevance, and semantic quality matter more than exact word matches. Multimodal reasoning tasks like MMInA [320] also require evaluating models on their ability to combine visual and textual information effectively [325].
- **Ethical Considerations and Biases:** Another key challenge in VLM evaluation is the bias present in many existing benchmarks. Datasets such as MS-COCO [69] and VQA [260] have been criticized for embedding gender, racial, and cultural biases, which can skew the evaluation of model performance. As Vision-Language Models are increasingly deployed in real-world applications, it is crucial to ensure that evaluation metrics and benchmarks address these biases and promote fairness [326].
- **Data Leakage and Unfair Comparisons:** The risk of data leakage is also a significant concern, particularly when training and evaluation datasets overlap. This issue can lead to inflated performance scores, making it difficult to assess the true capabilities of Vision-Language Models. Ensuring proper separation of training and evaluation data is critical for fair evaluation [327].

## 7. Challenges and Future Directions

As VLM develops, it faces many challenges related to their development, scalability, and real-world applicability. Although significant progress has been made, but still remain some key hurdles that need to be addressed for the further development of Vision-Language Models, specifically in the domains of fine-tuning, pre-trained models, prompt engineering, adapters, and datasets. Below, some of the key challenges emerging

trends, and proposals for research opportunities for enhancing VLM performance and deployment have been discussed.

## 7.1. Key Challenge

**Robustness, Bias, and Ethical Concerns in VLM Development** One of the key challenges in developing VLMs is ensuring their robustness and fairness across diverse environments and real time applications. A critical issue arises from biases embedded in both training datasets and model architectures, which can produce in discriminatory or unreliable results. For example, widely used datasets such as MS-COCO [69] and VQA [260] have been found to contain biases related to gender, race, and cultural representations. When these biases are incorporated into VLMs, they can reinforce harmful stereotypes, particularly when deployed in critical real time applications such as autonomous healthcare, driving or law enforcement [328]. Furthermore, the ethical concerns surrounding the use of these models in sensitive fields where biased or inaccurate results could lead to serious consequences underscore the importance of prioritizing fairness, transparency, and accountability in model design [329].

**Scalability** Scalability of VLM especially with large size pre-trained models is another significant challenge. Training and fine-tuning SOTA VLM such as Llava-grounding [330] and PaliGemma [331] requires extensive computational resources which are not available to many researchers and developers. This also poses a considerable environmental challenge because of high energy consumption of large model training. To address these challenges, novel approaches are crucial focusing on energy-efficient training methodologies and the optimization of both hardware designing and algorithmic frameworks. Such developments are vital to ensuring the sustainability and broader accessibility of VLM [332].

**Practical Deployment Challenges** Practical Deployment Challenges and Limitations of Vision-Language Models Despite the remarkable capabilities demonstrated by Vision-Language Models (VLMs), their widespread real-world deployment is frequently hindered by significant practical limitations. A primary concern is inference latency, as VLMs, particularly larger models, demand substantial computational resources for real-time applications. The multi-modal processing pipeline, often involving separate image encoding, text encoding, cross-attention, and autoregressive generation, can lead to sequential bottlenecks and suboptimal GPU utilization, thereby increasing response times. This computational intensity directly translates to high energy consumption, making VLM operation costly, especially for continuous inference. Consequently, deploying VLMs on edge devices or in low-resource settings like smartphones or IoT devices remains a formidable challenge. While progress is being made with techniques such as model quantization, pruning, and knowledge distillation to reduce model size and accelerate inference, significant performance-accuracy trade-offs often exist. Adapting large, cloud-optimized VLMs to the constrained memory, processing power, and battery life of edge hardware necessitates considerable engineering effort and platform-specific optimizations, often resulting in reduced performance or limited functionality.

Furthermore, the deployment of VLMs, especially in sensitive domains, introduces critical privacy concerns and compliance challenges. VLMs process and interpret potentially sensitive visual and textual data (e.g., personal identifiable information in images, medical records, surveillance footage). Ensuring compliance with stringent regulations like the General Data Protection Regulation (GDPR) requires robust mechanisms for data anonymization, consent management, secure data handling, and transparent data processing practices. Models must be designed with privacy-by-design principles, incorporating techniques such as federated learning or differential privacy to minimize the exposure of raw sensitive data. Additionally, the 'black-box' nature of complex VLMs can pose interpretability challenges, making it difficult to ascertain how decisions are made, which can be a barrier to trust and accountability in critical applications. Addressing these operational constraints is paramount for unlocking the full potential of VLMs across diverse real-world scenarios.

## 7.2. Emerging Trends

**Generalization:** A significant trend is the growing focus on model generalization and deployment of VLM in real world applications. While VLM has shown impressive success in controlled environments, it can fail when applied to unseen or domain-specific cases. To address this challenge, it is important to develop more effective generalization strategies that allow models to adapt to dynamic and diverse real-world scenarios [21]. Furthermore, there is an increasing interest in integrating VLM into real world applications including autonomous vehicles, robotics and healthcare where multimodal reasoning is crucial for decision making [333].

**Unified Foundational Models:** Another noticeable trend in VLM is the shift towards unified foundational models that utilize the same architecture to handle a wide range of vision and language tasks. Models like PaliGemma2 [12] and Cogvlm2 [334] show the potential for cross-domain learning and can handle multiple tasks with minimal specific tuning. This might lead to general-purpose models that can effortlessly transition across different domains, such as image captioning, visual question answering, object detection, and even multimodal reasoning [335].

**Multimodal Knowledge Adaptation:** The future of VLM is likely to focus more on multimodal knowledge transfer. Models that can transfer knowledge between multiple modalities like vision, language and audio can open up new research directions for cross-modal reasoning [336]. This will not only enhance VLM more flexibility but also enable them to adapt to new unseen tasks by utilizing existing multimodal knowledge [202]. Transferring and integrating knowledge across modalities will be one of the key factors of progress in vision-language models.

### 7.3. Research Opportunities

To unlock the full potential of Vision-Language Models and address the current challenges of robustness, scalability, practical deployment, and ethical concerns, several promising research avenues demand focused exploration. These opportunities involve both fundamental methodological innovations and advancements in training and deployment paradigms.

**Energy-Efficient Training and Deployment:** Due to the enormous computational cost of training large-scale Vision-Language Models, there is a dire need for an energy-efficient training strategies. Techniques such as quantization-aware training, mixed-precision computation, and pruned transformers (e.g., Tiny-ViT or MobileViT) offer promising paths to compress VLMs without significant loss in performance. Additionally, parameter-efficient tuning via LoRA or BitFit can reduce the memory demand for downstream tasks. Federated learning, combined with differential privacy, can enable distributed VLM training on edge devices while preserving data confidentiality.

**Multilingual and Low-Resource Cross-Modal Training:** Another promising research direction involves improving VLM generalizability across diverse languages and underrepresented modalities. Cross-lingual vision-language alignment can be achieved using multilingual contrastive learning with datasets like mVLT, or synthetic image-text pairs generated by multilingual LLMs. Additionally, cross-modal translation supervision and low-resource multitask adapters offer viable mechanisms for scaling VLMs to linguistically and culturally diverse settings while task consistency.

**Advancing Multimodal Continual Learning and Domain Adaptation:** Multimodal continual learning presents a compelling yet underexplored avenue for enabling VLMs to operate reliably in dynamic, real-world environments. Research should concentrate on catastrophic forgetting mitigation and domain adaptation, using mechanisms like Elastic Weight Consolidation (EWC) and AdapterFusion to retain previously learned knowledge. Integration with domain-specific memory replay and inter-task knowledge distillation can further enhance transferability and adaptability of VLMs in high-stakes applications such as healthcare, finance, or legal analysis.

**Integrating Symbolic Reasoning and External Knowledge:** Another interesting research direction for future VLM research is the integration of symbolic reasoning and structured knowledge. Combining VLMs with neuro-symbolic modules such as Neural Module Networks or knowledge graph-augmented models enables explicit logical reasoning beyond surface-level correlations. Applications like scene graph-based VQA or ontology-guided image captioning benefit from this synergy. These systems can be effectively implemented through graph attention networks or knowledge-injection layers within transformer-based architectures, bridging perceptual understanding with interpretable, knowledge-grounded reasoning.

## 8. Discussion

The evolution of Vision-Language Models has been marked by significant progress in architectural innovation, task adaptability, and multimodal representation learning. Our analysis of 115 publications reveals that recent VLMs have transitioned from monolithic, fully fine-tuned systems toward more modular, parameter-efficient frameworks driven by prompt engineering and adapter-based methods.

A key insight from this survey is the growing preference for parameter-efficient tuning techniques such as LoRA, BitFit, and adapter modules, which significantly reduce computational costs while maintaining competitive performance. These methods are particularly valuable in scenarios where resource constraints limit the feasibility of full model retraining. Similarly, prompt engineering has emerged as a lightweight yet powerful

method for steering pretrained models toward task-specific behavior. Soft and hybrid prompts, in particular, have shown notable improvements in low-resource and few-shot learning tasks, underscoring their utility in real-world applications.

The surveyed VLMs demonstrate increasingly strong performance across a range of multimodal tasks including image captioning, VQA, and retrieval. However, the lack of standardization in evaluation benchmarks and metrics remains a significant limitation. While widely used datasets like MS COCO and VQAv2 support comparative evaluation, inconsistencies in task setups and reporting conventions complicate cross-model comparisons. Furthermore, many current benchmarks fail to sufficiently test generalization to low-resource domains, multilingual contexts, and unseen modalities.

Another observation is the trade-off between model performance, interpretability, and efficiency. While large-scale models such as Gemini and GPT-4o offer strong generalization, they are often opaque and difficult to deploy in constrained environments. In contrast, adapter and prompt-based methods allow modular adaptation and reusability, but may underperform on complex reasoning tasks unless carefully optimized.

The integration of symbolic reasoning and structured knowledge into VLMs remains an underexplored but promising direction. Approaches that incorporate scene graphs or knowledge graphs could improve logical reasoning, interpretability, and consistency in model outputs. Similarly, the demand for cross-lingual and domain-adaptive capabilities highlights the importance of multilingual training strategies, synthetic data generation, and the development of language-agnostic tuning methods.

Overall, this survey illustrates that while VLMs have achieved significant milestones, several open challenges persist. Addressing these will require coordinated efforts across dataset development, model interpretability, and efficient deployment strategies. Future work must not only optimize performance but also ensure scalability, robustness, and fairness in real-world multimodal AI systems.

## 9. Conclusion

This survey offers a comprehensive synthesis of recent advancements in Vision Language Models, examining 115 peer-reviewed studies across five core components of VLM: fine-tuning methods, prompt engineering, adapter-based tuning, pre-trained model architectures, and benchmark datasets. We analyze 21 VLMs developed between 2018 and 2025 and evaluated eight adapter variants, including LoRA, BitFit, Houlsby, and Compacter. Our findings indicate that adapter-based methods can reduce trainable parameters by up to 98% while preserving 80 to 95% of the full model performance. Prompt engineering approaches, particularly soft and hybrid strategies, consistently yield 4% to 7% improvements over hard prompts in few-shot and zero-shot tasks.

Despite these advancements, significant challenges remain in areas such as generalization, multilingual adaptability, and evaluation consistency. To overcome these challenges, future work should explore symbolic reasoning by integrating structured knowledge sources, such as scene graphs, commonsense knowledge graphs (e.g., ConceptNet) or rule-based engines into VLM pipelines. This can be achieved through neuro-symbolic architectures that link distributed representations with symbolic modules to support logical inference, entity grounding, commonsense inference, and relational understanding. For instance, scene graphs can enhance object relationship modeling in image understanding, while commonsense graphs can improve contextual reasoning in question answering. Enhancing few-shot learning in multilingual settings requires dedicated language-specific adapters, cross-lingual prompts, and synthetic data generation to support low-resource domains. Additionally, improving model efficiency through sparse architectures, quantization, and adapter compression will be essential for deployment on edge devices. These future directions underscore the need for VLMs that are not only computationally efficient but also scalable, interpretable, and capable of robust performance across diverse real-world tasks and domains.

## Appendix A: Vision-Language Model Training Configurations

| Model Name | Batch Size | Learning Rate | Epochs / Tokens | Hardware | Training Strategy |
|---|---|---|---|---|---|
| CLIP (e.g., ViT-L/14) | 32k–65k | 1e-3 (cos. decay) | ~32 Epochs (LAION-400M) | 256–600+ V100/A100 GPUs | Contrastive pretraining on image-text pairs for zero-shot transfer |
| Flamingo (9B/80B) | ~1k–4k | 2e-5–1e-4 | Multi-stage finetune | Dozens–100s A100/H100 | Frozen LLM, gated cross-attn, few-shot tuning |
| LLaVA (1.5/1.6) | 16–128 | 2e-5 (finetune) | 1–3 Epochs (instr. data) | 8× A100 (40/80GB) | Two-stage instruction tuning; uses LoRA |
| GIT | 512–2048 | ~1e-4 (warm + decay) | Pretraining on 100B + tokens | 100s A100 GPUs | Unified VL modeling for captioning, VQA |
| Med-PaLM M | N/A | N/A | Billions of tokens | Google TPU Pods | Fine-tuned on multimodal clinical data |
| PaLM-E | N/A | N/A | Billions of tokens | Google TPU Pods | General-purpose multimodal LLM |
| GMAI-VL (5.5M) | ~16–64 | 2e-5 | ~5–10 Epochs | 8× A100 | Finetuned on merged medical datasets |
| Lingshu | ~64–256 | 1e-5–2e-5 | ~10–20 Epochs | 8–32 A100 GPUs | Multi-stage training on unified med data |
| HealthGPT | ~32–128 | 2e-5 | ~5–10 Epochs | 8× A100 | Instruction-tuned for medical comprehension and gen. |

## Acknowledgments

This research was supported by the Culture, Sports and Tourism R& D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Training Global Talent for Copyright Protection and Management of On-Device AI Models, Project Number: RS-2025-02221620, Contribution Rate: 100%).

## Conflicts of Interest

The authors declare no conflicts of interest.

## References

[1] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024.

[2] Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5689–5700, 2024.

[3] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence*, 7:1430984, 2024.

[4] Xiang Li, Like Li, Yuchen Jiang, Hao Wang, Xinyu Qiao, Ting Feng, Hao Luo, and Yong Zhao. Vision-language models in medical image analysis: From simple fusion to general large models. *Information Fusion*, page 102995, 2025.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

[7] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[8] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, 2021.

[9] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[10] Hrishikesh Singh, Aarti Sharma, and Millie Pant. Pixels to prose: Understanding the art of image captioning. *arXiv preprint arXiv:2408.15714*, 2024.

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[12] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

[13] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[15] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024.

[16] Muhammad Sohail Danish, Muhammad Akhtar Munir, Syed Roshaan Ali Shah, Kartik Kuckreja, Fahad Shahbaz Khan, Paolo Fraccaro, Alexandre Lacoste, and Salman Khan. Geobench-vlm: Benchmarking vision-language models for geospatial tasks. *arXiv preprint arXiv:2411.19325*, 2024.

[17] Zhenjie Cao, Zhuo Deng, Jie Ma, Jintao Hu, and Lan Ma. Mammovlm: A generative large vision-language model for mammography-related diagnostic assistance. *Information Fusion*, page 102998, 2025.

[18] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

[19] Tomas Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. The promises and pitfalls of llm annotations in dataset labeling: a case study on media bias detection. *arXiv preprint arXiv:2411.11081*, 2024.

[20] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[21] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024.

[22] Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Fluent and accurate image captioning with a self-trained reward model. *arXiv preprint arXiv:2408.16827*, 2024.

[23] Raihan Kabir, Naznin Haque, Md Saiful Islam, et al. A comprehensive survey on visual question answering datasets and algorithms. *arXiv preprint arXiv:2411.11150*, 2024.

[24] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning*, pages 39755–39769. PMLR, 2023.

[25] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing advanced visual reasoning ability of large language models. *arXiv preprint arXiv:2409.13980*, 2024.

[26] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, page 102963, 2025.

[27] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27831–27840, 2024.

[28] J Mao. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[29] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017.

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[31] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2025.

[32] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. Brave: Broadening the visual encoding of vision-language models. In *European Conference on Computer Vision*, pages 113–132. Springer, 2025.

[33] Kanyuan Dai, Ji Shao, Bo Gong, Ling Jing, and Yingyi Chen. Clip-fssc: A transferable visual model for fish and shrimp species classification based on natural language supervision. *Aquacultural Engineering*, 107:102460, 2024.

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[35] Hanning Chen, Wenjun Huang, Yang Ni, Sanggeon Yun, Yezi Liu, Fei Wen, Alvaro Velasquez, Hugo Latapie, and Mohsen Imani. Taskclip: Extend large vision-language model for task oriented object detection. *arXiv preprint arXiv:2403.08108*, 2024.

[36] Zhengfeng Lai, Haotian Zhang, Bowen Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, et al. Veclip: Improving clip training via visual-enriched captions. In *European Conference on Computer Vision*, pages 111–127. Springer, 2025.

[37] Shentong Mo and Pedro Morgado. Audio-visual generalized zero-shot learning the easy way. In *European Conference on Computer Vision*, pages 377–395. Springer, 2025.

[38] Chunpeng Zhou, Zhi Yu, Xilu Yuan, Sheng Zhou, Jiajun Bu, and Haishuai Wang. Less is more: A closer look at semantic-based few-shot learning. *Information Fusion*, 114:102672, 2025.

[39] Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, et al. Beyond efficiency: A systematic survey of resource-efficient large language models. *arXiv preprint arXiv:2401.00625*, 2024.

[40] Lukas Hoyer, David Joseph Tan, Muhammad Ferjad Naeem, Luc Van Gool, and Federico Tombari. Semivl: semi-supervised semantic segmentation with vision-language guidance. In *European Conference on Computer Vision*, pages 257–275. Springer, 2025.

[41] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[42] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.

[43] Ruifeng Guo, Jingxuan Wei, Linzhuang Sun, Bihui Yu, Guiyong Chang, Dawei Liu, Sibo Zhang, Zhengbing Yao, Mingjun Xu, and Liping Bu. A survey on image-text multimodal models. *arXiv preprint arXiv:2309.15857*, 2023.

[44] Shunsuke Koga and Wei Du. From text to image: challenges in integrating vision into chatgpt for medical image interpretation. *Neural Regeneration Research*, 20(2):487–488, 2025.

[45] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Vitamin: Designing scalable vision models in the vision-language era. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12954–12966, 2024.

[46] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

[47] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[48] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[49] Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. Vision-and-language pretrained models: A survey. *arXiv preprint arXiv:2204.07356*, 2022.

[50] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024.

[51] Yue Zhou, Litong Feng, Yiping Ke, Xue Jiang, Junchi Yan, Xue Yang, and Wayne Zhang. Towards vision-language geo-foundation model: A survey. *arXiv preprint arXiv:2406.09385*, 2024.

[52] Feng Li, Hao Zhang, Yi-Fan Zhang, Shilong Liu, Jian Guo, Lionel M Ni, PengChuan Zhang, and Lei Zhang. Vision-language intelligence: Tasks, representation learning, and large models. *arXiv preprint arXiv:2203.01922*, 2022.

[53] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*, 2022.

[54] Jindong Gu, Zhen Han, Shuo Chen, Ahmad Beirami, Bailan He, Gengyuan Zhang, Ruotong Liao, Yao Qin, Volker Tresp, and Philip Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

[55] Mayank Vatsa, Anubhooti Jain, and Richa Singh. Adventures of trustworthy vision-language models: A survey. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22650–22658, 2024.

[56] Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq Joty, Yixuan Li, Hai Li, et al. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *arXiv preprint arXiv:2407.21794*, 2024.

[57] Kun Ding, Ying Wang, Gaofeng Meng, and Shiming Xiang. A survey of low-shot vision-language model adaptation via representer theorem. *arXiv preprint arXiv:2410.11686*, 2024.

[58] Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171, 2022.

[59] Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. A survey on efficient vision-language models. *arXiv preprint arXiv:2504.09724*, 2025.

[60] Mohammad Ghiasvand Mohammadkhani, Saeedeh Momtazi, and Hamid Beigy. A survey on bridging vlms and synthetic data. *Authorea Preprints*, 2025.

[61] Ngoc Dung Huynh, Mohamed Reda Bouadjenek, Sunil Aryal, Imran Razzak, and Hakim Hacid. Visual question answering: from early developments to recent advances–a survey. *arXiv preprint arXiv:2501.03939*, 2025.

[62] Jialu Xing, Jianping Liu, Jian Wang, Lulu Sun, Xi Chen, Xunxun Gu, and Yingfei Wang. A survey of efficient fine-tuning methods for vision-language models—prompt and adapter. *Computers & Graphics*, 119:103885, 2024.

[63] Sungyeon Kim, Boseung Jeong, Donghyun Kim, and Suha Kwak. Efficient and versatile robust fine-tuning of zero-shot models. In *European Conference on Computer Vision*, pages 440–458. Springer, 2025.

[64] Xiaoqing Zhao, Miaomiao Xu, Wushour Silamu, and Yanbing Li. Clip-llama: A new approach for scene text recognition with a pre-trained vision-language model and a pre-trained language model. *Sensors*, 24(22):7371, 2024.

[65] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[66] Yi Xin, Siqi Luo, Haodi Zhou, Junlong Du, Xiaohong Liu, Yue Fan, Qing Li, and Yuntao Du. Parameter-efficient fine-tuning for pre-trained vision models: A survey. *arXiv preprint arXiv:2402.02242*, 2024.

[67] Yang Bai, Yang Zhou, Jun Zhou, Rick Siow Mong Goh, Daniel Shu Wei Ting, and Yong Liu. From generalist to specialist: Adapting vision language models via task-specific visual instruction tuning. *arXiv preprint arXiv:2410.06456*, 2024.

[68] Fan Liu, Tianshu Zhang, Wenwen Dai, Chuanyi Zhang, Wenwen Cai, Xiaocong Zhou, and Delong Chen. Few-shot adaptation of multi-modal foundation models: A survey. *Artificial Intelligence Review*, 57(10):268, 2024.

[69] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[70] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

[71] Zhichao Han, Azreen Azman, Mas Rina Mustaffa, and Fatimah Binti Khalid. Cross-modal retrieval: A review of methodologies, datasets, and future perspectives. *IEEE Access*, 2024.

[72] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

[73] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

[74] Shihong Liu, Samuel Yu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12687–12697, 2024.

[75] Congcong Wen, Yisiyuan Huang, Hao Huang, Yanjia Huang, Shuaihang Yuan, Yu Hao, Hui Lin, Yu-Shen Liu, and Yi Fang. Zero-shot object navigation with vision-language models reasoning. In *International Conference on Pattern Recognition*, pages 389–404. Springer, 2025.

[76] Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Benchmarking spurious bias in few-shot image classifiers. In *European Conference on Computer Vision*, pages 346–364. Springer, 2025.

[77] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *Meta-Radiology*, page 100047, 2023.

[78] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15670–15680, 2023.

[79] Jihwan Bang, Sumyeong Ahn, and Jae-Gil Lee. Active prompt learning in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27004–27014, 2024.

[80] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*, 2024.

[81] Asim Khan, Warda Asim, Anwaar Ulhaq, and Randall W. Robinson. A multiview semantic vegetation index for robust estimation of urban vegetation cover. *Remote Sensing*, 14(1):228, 2022.

[82] Yingjun Du, Wenfang Sun, and Cees GM Snoek. Ipo: Interpretable prompt optimization for vision-language models. *arXiv preprint arXiv:2410.15397*, 2024.

[83] Gahyeon Kim, Sohee Kim, and Seokju Lee. Aapl: Adding attributes to prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1572–1582, 2024.

[84] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Advances in Neural Information Processing Systems*, 36, 2024.

[85] Yunseon Choi, Sangmin Bae, Seonghyun Ban, Minchan Jeong, Chuheng Zhang, Lei Song, Li Zhao, Jiang Bian, and Kee-Eung Kim. Hard prompts made interpretable: Sparse entropy regularization for prompt tuning with rl. *arXiv preprint arXiv:2407.14733*, 2024.

[86] Sayash Raaj Hiraou. Optimising hard prompts with few-shot meta-prompting. *arXiv preprint arXiv:2407.18920*, 2024.

[87] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28578–28587, 2024.

[88] Dana Ramati, Daniela Gottesman, and Mor Geva. Eliciting textual descriptions from representations of continuous prompts. *arXiv preprint arXiv:2410.11660*, 2024.

[89] Sifan Long, Zhen Zhao, Junkun Yuan, Zichang Tan, Jiangjiang Liu, Jingyuan Feng, Shengsheng Wang, and Jingdong Wang. Mutual prompt leaning for vision language models. *International Journal of Computer Vision*, pages 1–19, 2024.

[90] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models. *arXiv preprint arXiv:2401.02418*, 2024.

[91] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[92] A Tuan Nguyen, Kai Sheng Tai, Bor-Chun Chen, Satya Narayan Shukla, Hanchao Yu, Philip Torr, Tai-Peng Tian, and Ser-Nam Lim. ucap: An unsupervised prompting method for vision-language models. In *European Conference on Computer Vision*, pages 425–439. Springer, 2025.

[93] Shuanghao Bai, Yuedi Zhang, Wanqi Zhou, Zhirong Luan, and Badong Chen. Soft prompt generation for domain generalization. In *European Conference on Computer Vision*, pages 434–450. Springer, 2025.

[94] Li-Wu Tsao, Hao-Tang Tsui, Yu-Rou Tuan, Pei-Chi Chen, Kuan-Lin Wang, Jhih-Ciang Wu, Hong-Han Shuai, and Wen-Huang Cheng. Trajprompt: Aligning color trajectory with vision-language representations. In *European Conference on Computer Vision*, pages 275–292. Springer, 2025.

[95] Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, and Abdenour Hadid. Fidavl: Fake image detection and attribution using vision-language model. In *International Conference on Pattern Recognition*, pages 160–176. Springer, 2025.

[96] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024.

[97] Shubin Huang, Qiong Wu, and Yiyi Zhou. Adapting pre-trained language models to vision-language tasksvia dynamic visual prompting. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[98] Yu Du, Tong Niu, and Rong Zhao. Mixture of prompt learning for vision language models. *arXiv preprint arXiv:2409.12011*, 2024.

[99] Qian Zhang. Generalizable prompt tuning for vision-language models. *arXiv preprint arXiv:2410.03189*, 2024.

[100] Ankit Jha. In the era of prompt learning with vision-language models. *arXiv preprint arXiv:2411.04892*, 2024.

[101] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5749–5757, 2024.

[102] Tingwei Zhang, Collin Zhang, John X Morris, Eugene Bagdasarian, and Vitaly Shmatikov. Soft prompts go hard: Steering visual language models with hidden meta-instructions. *arXiv preprint arXiv:2407.08970*, 2024.

[103] Anna Scius-Bertrand, Michael Jungo, Lars Vögtlin, Jean-Marc Spat, and Andreas Fischer. Zero-shot prompting and few-shot fine-tuning: Revisiting document image classification using large language models. In *International Conference on Pattern Recognition*, pages 152–166. Springer, 2025.

[104] M Jehanzeb Mirza, Leonid Karlinsky, Wei Lin, Sivan Doveh, Jakub Micorek, Mateusz Kozinski, Hilde Kuehne, and Horst Possegger. Meta-prompting for automating zero-shot visual recognition with llms. In *European Conference on Computer Vision*, pages 370–387. Springer, 2025.

[105] Yabin Zhang, Wenjie Zhu, Chenhang He, and Lei Zhang. Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In *European Conference on Computer Vision*, pages 271–288. Springer, 2025.

[106] Marc Lafon, Elias Ramzi, Clément Rambour, Nicolas Audebert, and Nicolas Thome. Gallop: Learning global and local prompts for vision-language models. In *European Conference on Computer Vision*, pages 264–282. Springer, 2025.

[107] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review*, 55(5):3503–3568, 2022.

[108] Atheer Algherairy and Moataz Ahmed. Prompting large language models for user simulation in task-oriented dialogue systems. *Computer Speech & Language*, 89:101697, 2025.

[109] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023.

[110] Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2305.18010*, 2023.

[111] Jameel Abdul Samadh, Mohammad Hanan Gani, Noor Hussein, Muhammad Uzair Khattak, Muhammad Muzammal Naseer, Fahad Shahbaz Khan, and Salman H Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[112] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1593–1603, 2024.

[113] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1401–1411, 2023.

[114] Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, and Biplab Banerjee. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5542–5552, 2024.

[115] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. *arXiv preprint arXiv:2411.15232*, 2024.

[116] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*, 2023.

[117] Linhao Qu, Dingkang Yang, Dan Huang, Qinhao Guo, Rongkui Luo, Shaoting Zhang, and Xiaosong Wang. Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In *European Conference on Computer Vision*, pages 196–212. Springer, 2025.

[118] Shounak Sural, Ragunathan Rajkumar, et al. Contextvlm: Zero-shot and few-shot context understanding for autonomous driving using vision language models. *arXiv preprint arXiv:2409.00301*, 2024.

[119] Xiang Gu, Shuchao Pang, Anan Du, Yifei Wang, Jixiang Miao, Jorge Díez, et al. Dynamic multimodal prompt tuning: Boost few-shot learning with vlm-guided point cloud models. In *European Conference on Artificial Intelligence*. Ulle Endriss, Francisco S. Melo, Kerstin Bach, Alberto Bugarín-Diz, José M …, 2024.

[120] Junda Wu, Rui Wang, Handong Zhao, Ruiyi Zhang, Chaochao Lu, Shuai Li, and Ricardo Henao. Few-shot composition learning for image retrieval with prompt tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4729–4737, 2023.

[121] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[122] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *AI Open*, 5:30–38, 2024.

[123] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.

[124] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021.

[125] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022.

[126] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023.

[127] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. In *Proceedings of EMNLP*, 2022.

[128] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*, 2022.

[129] Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022.

[130] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9516–9526, 2023.

[131] Muhammad Uzair khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[132] Changsheng Xu Hantao Yao, Rui Zhang. Visual-language prompt tuning with knowledge-guided context optimization. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[133] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23232–23241, 2023.

[134] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10878–10887, 2023.

[135] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023.

[136] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J. Kim. Read-only prompt optimization for vision-language few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[137] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023.

[138] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024.

[139] Changsheng Xu Hantao Yao, Rui Zhang. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[140] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23826–23837, June 2024.

[141] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models, 2023.

[142] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. *arXiv preprint arXiv:2306.01195*, 2024.

[143] Ge Wu, Xin Zhang, Zheng Li, Zhaowei Chen, Jiajun Liang, Jian Yang, and Xiang Li. Cascade prompt learning for vision-language model adaptation. In *European Conference on Computer Vision*, pages 304–321. Springer, 2025.

[144] Haoyang Li, Liang Wang, Chao Wang, Jing Jiang, Yan Peng, and Guodong Long. Dpc: Dual-prompt collaboration for tuning vision-language models, 2025.

[145] Matteo Farina, Massimiliano Mancini, Giovanni Iacca, and Elisa Ricci. Rethinking few-shot adaptation of vision-language models in two stages, 2025.

[146] Xiaocheng Lu, Ziming Liu, Song Guo, Jingcai Guo, Fushuo Huo, Sikai Bai, and Tao Han. Drpt: Disentangled and recurrent prompt tuning for compositional zero-shot learning. *arXiv preprint arXiv:2305.01239*, 2023.

[147] Nihal V Nayak, Peilin Yu, and Stephen H Bach. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574*, 2022.

[148] Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, Peng Wang, and Yanning Zhang. Dual modality prompt tuning for vision-language pre-trained model. *IEEE Transactions on Multimedia*, 2023.

[149] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023.

[150] Eulrang Cho, Jooyeon Kim, and Hyunwoo J Kim. Distribution-aware prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22004–22013, 2023.

[151] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[152] Minh Le, Chau Nguyen, Huy Nguyen, Quyen Tran, Trung Le, and Nhat Ho. Revisiting prefix-tuning: Statistical benefits of reparameterization among prompts. *arXiv preprint arXiv:2410.02200*, 2024.

[153] Shouchang Guo, Sonam Damani, and Keng-hao Chang. Lopt: Low-rank prompt tuning for parameter efficient language models. *arXiv preprint arXiv:2406.19486*, 2024.

[154] Bruce XB Yu, Jianlong Chang, Lingbo Liu, Qi Tian, and Chang Wen Chen. Towards a unified view on visual parameter-efficient transfer learning. *arXiv preprint arXiv:2210.00788*, 2022.

[155] Chengming Xu, Siqian Yang, Yabiao Wang, Zhanxiong Wang, Yanwei Fu, and Xiangyang Xue. Exploring efficient few-shot adaptation for vision transformers. *arXiv preprint arXiv:2301.02419*, 2023.

[156] Cheng-Hao Tu, Zheda Mai, and Wei-Lun Chao. Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7725–7735, 2023.

[157] Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, and Jian Zhang. A unified continual learning framework with general parameter-efficient tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11483–11493, 2023.

[158] Zining Chen, Weiqiu Wang, Zhicheng Zhao, Fei Su, Aidong Men, and Hongying Meng. Practicaldg: Perturbation distillation on vision-language models for hybrid domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23501–23511, 2024.

[159] Fan Yang, Xiabing Zhou, Min Zhang, and Guodong Zhou. Mixture of hybrid prompts for cross-domain aspect sentiment triplet extraction. *IEEE Transactions on Affective Computing*, 2024.

[160] Fouad Trad and Ali Chehab. Evaluating the efficacy of prompt-engineered large multimodal models versus fine-tuned vision transformers in image-based security applications. *arXiv preprint arXiv:2403.17787*, 2024.

[161] Chong Bian, Xue Han, Zhiyu Duan, Chao Deng, Shunkun Yang, and Junlan Feng. Hybrid prompt-driven large language model for robust state-of-charge estimation of multi-type li-ion batteries. *IEEE Transactions on Transportation Electrification*, 2024.

[162] Yuzhen Zhong, Tong Xu, and Pengfei Luo. Contextualized hybrid prompt-tuning for generation-based event extraction. In *International Conference on Knowledge Science, Engineering and Management*, pages 374–386. Springer, 2023.

[163] Qinglong Cao, Yuntian Chen, Lu Lu, Hao Sun, Zhenzhong Zeng, Xiaokang Yang, and Dongxiao Zhang. Promoting ai equity in science: Generalized domain prompt learning for accessible vlm research. *arXiv preprint arXiv:2405.08668*, 2024.

[164] Xiaohong Liu, Guoxing Yang, Yulin Luo, Jiaji Mao, Xiang Zhang, Ming Gao, Shanghang Zhang, Jun Shen, and Guangyu Wang. Expert-level vision-language foundation model for real-world radiology and comprehensive evaluation. *arXiv preprint arXiv:2409.16183*, 2024.

[165] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024.

[166] Rongyu Zhang, Zefan Cai, Huanrui Yang, Zidong Liu, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, Baobao Chang, Yuan Du, et al. Vecaf: Vision-language collaborative active finetuning with training objective awareness. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5451–5459, 2024.

[167] Thong Nguyen, Yi Bin, Xiaobao Wu, Xinshuai Dong, Zhiyuan Hu, Khoi Le, Cong-Duy Nguyen, See-Kiong Ng, and Luu Anh Tuan. Meta-optimized angular margin contrastive framework for video-language representation learning. In *European Conference on Computer Vision*, pages 77–98. Springer, 2025.

[168] Ali Abdollah, Amirmohammad Izadi, Armin Saghafian, Reza Vahidimajd, Mohammad Mozafari, Amirreza Mirzaei, Mohammadmahdi Samiei, and Mahdieh Soleymani Baghshah. Comalign: Compositional alignment in vision-language models. *arXiv preprint arXiv:2409.08206*, 2024.

[169] Ziyang Song, Qincheng Lu, He Zhu, and Yue Li. Bidirectional generative pre-training for improving time series representation learning. *arXiv preprint arXiv:2402.09558*, 2024.

[170] Yi Zhang, Ce Zhang, Ke Yu, Yushun Tang, and Zhihai He. Concept-guided prompt learning for generalization in vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 7377–7386, 2024.

[171] Shaozhi Wu, Han Li, Xingang Liu, Dan Tian, and Han Su. Class-aware patch based contrastive learning for medical image segmentation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2024.

[172] Salma Haidar and José Oramas. Enhancing hyperspectral image prediction with contrastive learning in low-label regime. *arXiv preprint arXiv:2410.07790*, 2024.

[173] Mélanie Roschewitz, Fabio De Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Robust image representations with counterfactual contrastive learning. *arXiv preprint arXiv:2409.10365*, 2024.

[174] Pengxiang Ouyang, Jianan Chen, Qing Ma, Zheng Wang, and Cong Bai. Distinguishing visually similar images: Triplet contrastive learning framework for image-text retrieval. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[175] Sameer Khanna, Daniel Michael, Marinka Zitnik, and Pranav Rajpurkar. Learning generalized medical image representations through image-graph contrastive pretraining. In *Machine Learning for Health (ML4H)*, pages 232–243. PMLR, 2023.

[176] Tianxiang Wu and Shuqun Yang. Contrastive enhanced learning for multi-label text classification. *Applied Sciences*, 14(19):8650, 2024.

[177] Ji Ma, Wei Suo, Peng Wang, and Yanning Zhang. C3l: Content correlated vision-language instruction tuning data generation via contrastive learning. *arXiv preprint arXiv:2405.12752*, 2024.

[178] Fei He, Kai Liu, Zhiyuan Yang, Yibo Chen, Richard D Hammer, Dong Xu, and Mihail Popescu. pathclip: Detection of genes and gene relations from biological pathway figures through image-text contrastive learning. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[179] Jingqi Hu, Chen Mao, Chong Tan, Hui Li, Hong Liu, and Min Zheng. Progeo: Generating prompts through image-text contrastive learning for visual geo-localization. In *International Conference on Artificial Neural Networks*, pages 448–462. Springer, 2024.

[180] Lin Zhu, Weihan Yin, Yiyao Yang, Fan Wu, Zhaoyu Zeng, Qinying Gu, Xinbing Wang, Chenghu Zhou, and Nanyang Ye. Vision-language alignment learning under affinity and divergence principles for few-shot out-of-distribution generalization. *International Journal of Computer Vision*, pages 1–33, 2024.

[181] Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. *arXiv preprint arXiv:2406.11832*, 2024.

[182] Chris Kelly, Luhui Hu, Bang Yang, Yu Tian, Deshun Yang, Cindy Yang, Zaoshan Huang, Zihao Li, Jiayin Hu, and Yuexian Zou. Visiongpt: Vision-language understanding agent using generalized multimodal framework. *arXiv preprint arXiv:2403.09027*, 2024.

[183] Yihuan Zhu, Honghua Xu, Ailin Du, and Bin Wang. Image–text matching model based on clip bimodal encoding. *Applied Sciences*, 14(22):10384, 2024.

[184] Ran Jin, Tengda Hou, Tao Jin, Jie Yuan, and Chenjie Du. A method for image–text matching based on semantic filtering and adaptive adjustment. *EURASIP Journal on Image and Video Processing*, 2024(1):23, 2024.

[185] Zhe Li, Lei Zhang, Kun Zhang, Yongdong Zhang, and Zhendong Mao. Improving image-text matching with bidirectional consistency of cross-modal alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[186] Soojin Jang, Jungmin Yun, Junehyoung Kwon, Eunju Lee, and Youngbin Kim. Dial: Dense image-text alignment for weakly supervised semantic segmentation. *arXiv preprint arXiv:2409.15801*, 2024.

[187] Bo Li, You Wu, and Zhixin Li. Team huge: Image-text matching via hierarchical and unified graph enhancing. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 704–712, 2024.

[188] Yan Huang, Yuming Wang, Yunan Zeng, Junshi Huang, Zhenhua Chai, and Liang Wang. Unpaired image-text matching via multimodal aligned conceptual knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[189] Jinguo Zhu, Xiaohan Ding, Yixiao Ge, Yuying Ge, Sijie Zhao, Hengshuang Zhao, Xiaohua Wang, and Ying Shan. Vl-gpt: A generative pre-trained transformer for vision and language understanding and generation. *arXiv preprint arXiv:2312.09251*, 2023.

[190] T Sanjay et al. Enhancing image generation by fusing auto encoder & transformative generation approach. In *2024 International Conference on Electrical Electronics and Computing Technologies (ICEECT)*, volume 1, pages 1–6. IEEE, 2024.

[191] Zhiming Mao, Haoli Bai, Lu Hou, Jiansheng Wei, Xin Jiang, Qun Liu, and Kam-Fai Wong. Visually guided generative text-layout pre-training for document intelligence. *arXiv preprint arXiv:2403.16516*, 2024.

[192] Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Shutao Xia, and Ke Xu. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. *arXiv preprint arXiv:2406.05491*, 2024.

[193] Jiawei Chen, Dingkang Yang, Yue Jiang, Yuxuan Lei, and Lihua Zhang. Miss: A generative pre-training and fine-tuning approach for med-vqa. In *International Conference on Artificial Neural Networks*, pages 299–313. Springer, 2024.

[194] Dominykas Seputis, Serghei Mihailov, Soham Chatterjee, and Zehao Xiao. Multi-modal adapter for vision-language models. *arXiv preprint arXiv:2409.02958*, 2024.

[195] Daniel P Jeong, Saurabh Garg, Zachary C Lipton, and Michael Oberst. Medical adaptation of large language and vision-language models: Are we making progress? *arXiv preprint arXiv:2411.04118*, 2024.

[196] Rui Cai, Zhiyu Dong, Jianfeng Dong, and Xun Wang. Dynamic adapter with semantics disentangling for cross-lingual cross-modal retrieval. *arXiv preprint arXiv:2412.13510*, 2024.

[197] Md Shohel Sayeed, Varsha Mohan, and Kalaiarasi Sonai Muthu. Bert: A review of applications in sentiment analysis. *HighTech and Innovation Journal*, 4(2):453–462, 2023.

[198] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[199] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237, 2022.

[200] Mateusz Klimaszewski, Piotr Andruszkiewicz, and Alexandra Birch. No train but gain: Language arithmetic for training-free language adapters enhancement. *arXiv preprint arXiv:2404.15737*, 2024.

[201] Yi-Chen Li, Fuxiang Zhang, Wenjie Qiu, Lei Yuan, Chengxing Jia, Zongzhang Zhang, and Yang Yu. Q-adapter: Training your llm adapter as a residual q-function. *arXiv e-prints*, pages arXiv–2407, 2024.

[202] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23826–23837, 2024.

[203] Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36, 2024.
[204] Haoyu Lu, Yuqi Huo, Guoxing Yang, Zhiwu Lu, Wei Zhan, Masayoshi Tomizuka, and Mingyu Ding. Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. *arXiv preprint arXiv:2302.06605*, 2023.
[205] Junfei Xiao, Zheng Xu, Alan Yuille, Shen Yan, and Boyu Wang. Palm2-vadapter: Progressively aligned language model makes a strong vision-language adapter. *arXiv preprint arXiv:2402.10896*, 2024.
[206] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199*, 2021.
[207] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019.
[208] Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. Compacter: Efficient low-rank hypercomplex adapter layers, 2021.
[209] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
[210] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366*, 2021.
[211] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model, 2023.
[212] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning, 2021.
[213] Ting Liu, Xuyang Liu, Siteng Huang, Honggang Chen, Quanjun Yin, Long Qin, Donglin Wang, and Yue Hu. Dara: Domain-and relation-aware adapters make parameter-efficient tuning for visual grounding. *arXiv preprint arXiv:2405.06217*, 2024.
[214] Juncheng Yang, Zuchao Li, Shuai Xie, Weiping Zhu, Wei Yu, and Shijun Li. Cross-modal adapter: Parameter-efficient transfer learning approach for vision-language models. *arXiv preprint arXiv:2404.12588*, 2024.
[215] Julio Silva-Rodriguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23681–23690, 2024.
[216] Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14151, 2024.
[217] Ming Li, Jike Zhong, Chenxin Li, Liuzhuozheng Li, Nie Lin, and Masashi Sugiyama. Vision-language model fine-tuning via simple parameter-efficient modification. *arXiv preprint arXiv:2409.16718*, 2024.
[218] Joey Hong, Anca Dragan, and Sergey Levine. Q-sft: Q-learning for language models via supervised fine-tuning. *arXiv preprint arXiv:2411.05193*, 2024.
[219] Zhizhao Duan, Hao Cheng, Duo Xu, Xi Wu, Xiangxie Zhang, Xi Ye, and Zhen Xie. Cityllava: Efficient fine-tuning for vlms in city scenario. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7180–7189, 2024.
[220] Chengwei Sun, Jiwei Wei, Yujia Wu, Yiming Shi, Shiyuan He, Zeyu Ma, Ning Xie, and Yang Yang. Svfit: Parameter-efficient fine-tuning of large pre-trained models using singular values. *arXiv preprint arXiv:2409.05926*, 2024.
[221] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. Lora vs full fine-tuning: An illusion of equivalence. *arXiv preprint arXiv:2410.21228*, 2024.
[222] Yuhang Zang, Hanlin Goh, Josh Susskind, and Chen Huang. Overcoming the pitfalls of vision-language model finetuning for ood generalization. *arXiv preprint arXiv:2401.15914*, 2024.
[223] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
[224] Erik Nijkamp, Tian Xie, Hiroaki Hayashi, Bo Pang, Congying Xia, Chen Xing, Jesse Vig, Semih Yavuz, Philippe Laban, Ben Krause, et al. Xgen-7b technical report. *arXiv preprint arXiv:2309.03450*, 2023.
[225] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.
[226] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
[227] Hugging Face. Introducing idefics: An open reproduction of state-of-the-art multimodal models. https://huggingface.co/blog/idefics, 2023.
[228] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie

Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[229] AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1:1, 2024.

[230] Google DeepMind. Introducing gemini 2.0: Our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/, 2024.

[231] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[232] Vikhyat Korrapati. moondream2 (revision 92d3d73), 2024.

[233] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[234] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision, 2022.

[235] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[236] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[237] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[238] Tengchao Lv, Yupan Huang, Jingye Chen, Yuzhong Zhao, Yilin Jia, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, Shaoxiang Wu, Guoxin Wang, Cha Zhang, and Furu Wei. Kosmos-2.5: A multimodal literate model, 2024.

[239] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han.

Longvila: Scaling long-context visual language models for long videos, 2024.

[240] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024.

[241] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[242] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[243] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral-12b, 2024.

[244] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017.

[245] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021.

[246] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents, 2023.

[247] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[248] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[249] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

[250] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[251] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011.

[252] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[253] Rocktim Jyoti Das, Simeon Emilov Hristov, Haonan Li, Dimitar Iliyanov Dimitrov, Ivan Koychev, and Preslav Nakov. Exams-v: A multi-discipline multilingual multimodal exam benchmark for evaluating vision language models. *arXiv preprint arXiv:2403.10378*, 2024.

[254] Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[255] M Maruf, Arka Daw, Kazi Sajeed Mehrab, Harish Babu Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James Balhoff, Yasin Bakis, Bahadir Altintas, et al. Vlm4bio: A benchmark dataset to evaluate pretrained vision-language models for trait discovery from biological images. *Advances in Neural Information Processing Systems*, 37:131035–131071, 2024.

[256] Xiao Wang, Ibrahim Alabdulmohsin, Daniel Salz, Zhe Li, Keran Rong, and Xiaohua Zhai. Scaling pre-training to one hundred billion data for vision language models. *arXiv preprint arXiv:2502.07617*, 2025.

[257] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[258] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[259] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[260] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[261] Edwin G. Ng, Bo Pang, Piyush Sharma, and Radu Soricut. Understanding guided image captioning performance across domains. *arXiv preprint arXiv:2012.02339*, 2020.

[262] Haowen Pan, Yixin Cao, Xiaozhi Wang, and Xun Yang. Finding and editing multi-modal neurons in pre-trained transformer. *arXiv preprint arXiv:2311.07470*, 2023.

[263] Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6859–6865. IEEE, 2023.

[264] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024.

[265] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017.

[266] Ahmed Sabir, Francesc Moreno-Noguer, and Lluís Padró. Textual visual semantic dataset for text spotting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 542–543, 2020.

[267] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024.

[268] Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pages 260–278. Springer, 2025.

[269] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[270] Chaitanya Shivade. Mednli — a natural language inference dataset for the clinical domain, 2017.

[271] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021.

[272] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.

[273] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

[274] Maxime Kayser, Cornelius Emde, Oana Camburu, Guy Parsons, Bartlomiej Papiez, and Thomas Lukasiewicz. Explaining chest x-ray pathologies in natural language. In *Proceedings of the 25th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2022, Singapore, 18–22 September 2022*, Lecture Notes in Computer Science (LNCS). Springer, September 2022.

[275] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[276] Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. *arXiv preprint arXiv:2411.16778*, 2024.

[277] Benedikt Boecking et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[278] Lucy Lu Wang Sanjay Subramanian and other. MedICaT: A Dataset of Medical Images, Captions, and Textual References. In *Findings of EMNLP*, 2020.

[279] Xiaotang Gai, Jiaxiang Liu, Yichen Li, Zijie Meng, Jian Wu, and Zuozhu Liu. 3d-rad: A comprehensive 3d radiology med-vqa dataset with multi-temporal analysis and diverse diagnostic tasks, 2025.

[280] ImageCLEFmed Organization. Imageclefmed medvqa-gi 2025 challenge overview. https://www.imageclef.org/2025/medical/vqa, 2025. Accessed: [18 june 2025 ].

[281] Alejandro Lozano, Min Woo Sun, James Burgess, Liangyu Chen, Jeffrey J Nirschl, Jeffrey Gu, Ivan Lopez, Josiah Aklilu, Austin Wolfgang Katzer, Collin Chiu, Anita Rau, Xiaohan Wang, Yuhui Zhang, Alfred Seunghoon Song, Robert Tibshirani, and Serena Yeung-Levy. Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature, 2025.

[282] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023.

[283] Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. *arXiv preprint arXiv:2506.09513*, 2025.

[284] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[285] Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. *arXiv preprint arXiv:2411.14522*, 2024.

[286] Omar Moured, Jiaming Zhang, M Saquib Sarfraz, and Rainer Stiefelhagen. Altchart: Enhancing vlm-based chart summarization through multi-pretext tasks. In *International Conference on Document Analysis and Recognition*, pages 349–366. Springer, 2024.

[287] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049, 2020.

[288] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

[289] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021.

[290] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022.

[291] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *Advances in neural information processing systems*, 35:32942–32956, 2022.

[292] Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. Nlip: Noise-robust language-image pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 926–934, 2023.

[293] Chenyu Yang, Xizhou Zhu, Jinguo Zhu, Weijie Su, Junjie Wang, Xuan Dong, Wenhai Wang, Bin Li, Jie Zhou, Yu Qiao, et al. Vision model pre-training on interleaved image-text data via latent compression learning. *Advances in Neural Information Processing Systems*, 37:23912–23938, 2024.

[294] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022.

[295] Yikang Li, Jen-hao Hsiao, and Chiuman Ho. Object prior embedded network for query-agnostic image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4969–4974, 2022.

[296] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, pages 73–90. Springer, 2024.

[297] Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, et al. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*, 2024.

[298] Zhenlin Xu, Yi Zhu, Siqi Deng, Abhay Mittal, Yanbei Chen, Manchen Wang, Paolo Favaro, Joseph Tighe, and Davide Modolo. Benchmarking zero-shot recognition with vision-language models: Challenges on granularity and specificity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1827–1836, 2024.

[299] Kai Chen, Yunxin Li, Xiwen Zhu, Wentai Zhang, and Baotian Hu. A vision-language model with multi-granular knowledge fusion in medical imaging. *World Wide Web*, 28(1):1–21, 2025.

[300] Yuan Liu, Le Tian, Xiao Zhou, Xinyu Gao, Kavio Yu, Yang Yu, and Jie Zhou. Points1. 5: Building a vision-language model towards real world applications. *arXiv preprint arXiv:2412.08443*, 2024.

[301] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179, 2024.

[302] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[303] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[304] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.

[305] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*, 2024.

[306] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

[307] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025.

[308] Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*, 2023.

[309] Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *arXiv preprint arXiv:2408.00765*, 2024.

[310] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[311] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xucheng Yin, Cheng lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: On the hidden mystery of ocr in large multimodal models, 2024.

[312] Tianyu Zhang, Suyuchen Wang, Lu Li, Ge Zhang, Perouz Taslakian, Sai Rajeswar, Jie Fu, Bang Liu, and Yoshua Bengio. Vcr: Visual caption restoration. *arXiv preprint arXiv:2406.06462*, 2024.

[313] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[314] Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, et al. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*, 2024.

[315] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024.

[316] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[317] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

[318] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

[319] Mor Ventura, Michael Toker, Nitay Calderon, Zorik Gekhman, Yonatan Bitton, and Roi Reichart. Nl-eye: Abductive nli for images. *arXiv preprint arXiv:2410.02613*, 2024.

[320] Ziniu Zhang, Shulin Tian, Liangyu Chen, and Ziwei Liu. Mmina: Benchmarking multihop multimodal internet agents. *arXiv preprint arXiv:2404.09992*, 2024.

[321] Song Dingjie, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. In *First Conference on Language Modeling*, 2024.

[322] Ahmad Sami Al-Shamayleh, Omar Adwan, Mohammad A Alsharaiah, Abdelrahman H Hussein, Qasem M Kharma, and Christopher Ifeanyi Eke. A comprehensive literature review on image captioning methods and metrics based on deep learning technique. *Multimedia Tools and Applications*, 83(12):34219–34268, 2024.

[323] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019.

[324] Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M Frankland, Thomas L Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *arXiv preprint arXiv:2411.00238*, 2024.

[325] Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024.

[326] Gabriele Ruggeri, Debora Nozza, et al. A multi-dimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, 2023.

[327] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024.

[328] Xiyang Wu, Ruiqi Xian, Tianrui Guan, Jing Liang, Souradip Chakraborty, Fuxiao Liu, Brian M Sadler, Dinesh Manocha, and Amrit Bedi. On the safety concerns of deploying llms/vlms in robotics: Highlighting the risks and vulnerabilities. In *First Vision and Language for Autonomous Driving and Robotics Workshop*, 2024.

[329] Anetta Jedličková. Ethical approaches in designing autonomous and intelligent systems: a comprehensive survey towards responsible development. *AI & SOCIETY*, pages 1–14, 2024.

[330] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2025.

[331] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[332] Aditi Singh, Nirmal Prakashbhai Patel, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. A survey of sustainability in large language models: Applications, economics, and challenges. *arXiv preprint arXiv:2412.04782*, 2024.

[333] Saman Hashemi. Integrating qt and llms on the nvidia jetson board for controlling a patient-assisting robot arm. Master's thesis, S. Hashemi, 2024.

[334] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

[335] Tanmay Gupta, Amita Kamath, Aniruddha Kembhavi, and Derek Hoiem. Towards general purpose vision systems: An end-to-end task-agnostic vision-language architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16399–16409, 2022.

[336] Dongha Choi, Jung Jae Kim, and Hyunju Lee. Transfercvlm: Transferring cross-modal knowledge for vision-language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16733–16746, 2024.

# A comprehensive VLM Survey: Fine-Tuning, Pre-trained Models, Prompt Engineering, Adapter, and Benchmarking Data

- **Comprehensive Analysis of VLM Components**: The survey offers a comprehensive exploration of fundamental Vision Language Model (VLM) components such as fine-tuning, pre-trained models, prompt engineering, adapters, and benchmarking datasets. These components play an important role in multi-sensor, multi-source information fusion by enhancing the model's capability to process diverse input types.

- **Optimization Techniques for Improved Efficiency**: The study delves into contemporary optimization approaches such as adapter-based fine-tuning and low-resource learning strategies to enhance computational efficiency and adaptability across diverse tasks.

- **Advancements in Pre-training and Prompt Engineering**: The study highlights innovations in pre-training, including contrastive and generative approaches, and examines the role of prompt engineering in refining VLM performance for various downstream applications.

- **Benchmarking and Dataset Challenges**: The paper addresses the significance of benchmarking datasets, focusing on data diversity, annotation quality, and bias mitigation to ensure fair and robust model evaluation.

- **Future Directions and Ethical Considerations**: It outlines key challenges in VLM research, including scalability, domain-specific adaptation, and ethical deployment, providing a forward-looking agenda for future advancements in the field.

**Declaration of interests**

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: