Efficient Adaptation of Large Vision-Language Models: Transfer Learning Methods and Applications

Anonymous authors
Paper under double-blind review

Abstract

Pre-trained large vision-language models (VLMs) have become the dominant choice for handling vision-language tasks, covering from multimodal reasoning to text-image generation. However, these models heavily depend on large-scale training datasets, primarily composed of image-text pairs sourced from web data, which are typically confined to general domains rather than specific downstream tasks. Given the scarcity of data in such specialized domains, transfer learning emerges as a remedy, enabling the adaptation of a model's preexisting knowledge to new tasks with limited data, thereby mitigating the reliance on extensive datasets. Following the current trend of the transfer learning application with vision-language tasks, we provide a systematic study of existing transfer learning techniques adopted for vision-language models, including: (1) a summary of the existing state-of-the-art VLMs, (2) a comprehensive taxonomy of transfer learning approaches for VLMs, (3) the discussion of real-world applications of transfer learning methods for VLMs, (4) a summary of commonly used vision-language dataset and benchmarks in variant vision-language tasks.

1 Introduction

Vision-language models (VLMs) present a dynamic frontier in machine learning where the modalities of vision and language intersect to address complex tasks that require the inputs of both modalities. These models process and interpret visual data alongside textual descriptions, enabling machines to understand and generate content that reflects both visual perception and text context. The significance of VLMs lies in their ability to perform a broad spectrum of tasks much better compared to architectures that process modalities independently Shangguan et al. (2025a), such as image captioning (Zhou et al., 2020; Li et al., 2023; Alayrac et al., 2022; Li et al., 2022b; Zhang et al., 2023a), visual question answering (Antol et al., 2015; Kim et al., 2021; Li et al., 2021; Floridi & Chiriatti, 2020), image classification (Radford et al., 2021; Alayrac et al., 2022; Dosovitskiy et al., 2021), etc. These capabilities make VLMs advance numerous practical applications, including robotics system (Chen et al., 2024a; Kuzmenko & Shvai, 2024; Liu et al., 2024b), medical assistance (Chambon et al., 2022; Chen et al., 2022c; Chang et al., 2024), and human sentimental understanding (Wang et al., 2023a; Shi et al., 2024). Despite the potential of VLMs for the broad applications, the development of VLMs that effectively integrate visual and text information to fit in variant domains remains challenging. With the growing trend of pre-training VLMs, model size has grown from 200 million (Li et al., 2021) to 562 billion (Driess et al., 2023) in a few years. At the same time, the size of dataset needed to train the VLMs also has grown from million-scale to billion-scale, making it impractical to develop independent pre-trained VLMs for every sub-domain and every downstream task. Finally, the computational cost to develop a pre-trained VLM has grown to the extent that most of the state-of-the-art VLMs can be only trained by a few large companies.

To address the complexity of the VLM training, transfer learning has emerged as a pivotal strategy in the "Pre-training, Finetuning, Prediction" approach for using VLMs (Xing et al., 2024a). Transfer learning for VLMs involves taking a model pre-trained on one task with massive amount of data, and finetuning it for another specific task using the relatively small amount of data. This approach not only accelerates the training process by utilizing pre-learned features but also enhances the model's ability to generalize from one task to another within or across domains. By transferring knowledge from large, general datasets to

more specialized tasks, VLMs can achieve state-of-the-art performance across a wide range of applications. Currently, Transfer learning and VLMs have been studied with several surveys in each domain. For example, Du et al. 2022 and Zhang et al. Zhang et al. (2024b) survey pretrained VLMs by categorizing them according to the training objective function, architecture, and the used pre-training tasks. Several extensive papers summarize transfer learning strategies in a general setting Zhuang et al. (2020); Weiss et al. (2016); Tan et al. (2018). However, these works do not study transfer learning in the context of VLMs. Recently, Xing et al. Xing et al. (2024a) provides insights to this specific field, but they mainly focus on the adapter-based and the prompt-based transfer learning method which, as we will see in this paper, is not the comprehensive picture of the field. Compared to their work, we provide a comprehensive survey of efficient transfer learning methods used for VLMs. Expanded from prior works, we carefully summarize and classify the transfer learning methods for VLMs into four general categories: **Adapter-based** methods, **Prompt-based** methods, **Model-based** methods, and **Knowledge-based** methods. We also study the corresponding VLMs, datasets, and realistic applications for the VLM transfer learning methods.

The paper is organized as follows: Section 2 introduces a background on vision-language tasks and transfer learning. Section 3 provides insight about pre-trained vision language models, with their architectures and training objectives. Section 4 discusses the taxonomy of efficient transfer learning method for vision-language tasks. Section 5 introduces the real-world application of vision-language transfer learning in various domains. Section 6 discusses the common benchmarks and datasets for vision-language tasks. The paper is concluded in Section 7.

2 Background & Foundations

Recent advances in the vision-language models and the transfer learning techniques have significantly expanded the models' understanding of the vision language tasks . In this section, we review prior work on vision–language tasks and discuss how transfer learning has facilitated efficient adaptation across domains and modalities.

2.1 Vision Language Tasks

Vision-language tasks involve jointly understanding and reasoning over visual and textual information by enabling cross-modal alignment and comprehension. The early machine learning approaches for handling these tasks relied on manually engineered features and separate models for vision and language processing, such as pre-trained convolutional neural networks (CNNs) for vision and recurrent neural networks (RNNs) for language (Wang et al., 2016; Khamparia et al., 2020; Yu et al., 2017). The outputs of these paths then are integrated to perform the downstream task. Despite obtaining some success, these approaches are limited because of their reliance on task-specific feature engineering and separate processing pipelines, which often struggled to achieve seamless multimodal alignment. Recent advancements in deep learning have enabled the development of unified architectures that use large-scale pre-training to achieve remarkably better performance than the simple combination of CNNs and RNNs (Radford et al., 2021; Chen et al., 2020; Kim et al., 2021; Wang et al., 2021; Li et al., 2021; 2020; Jia et al., 2021; Li et al., 2022b). Building on the foundation laid by unified architectures, more recent large-scale models, which have billion and trillion of parameters, mark a significant step forward in performing vision-language tasks. These recent models leverage massive datasets and transformer-based architectures to process and generate both visual and textual information seamlessly (Achiam et al., 2023; Alayrac et al., 2022; Ramesh et al., 2021; Driess et al., 2023; Huang et al., 2023; Zhang et al., 2023a). For example, GPT-4V (Achiam et al., 2023) exemplifies the trend of incorporating vision into language-dominant models, enabling complex reasoning over image-text inputs via 1.76 trillion parameters. On the side of generative task, DALL · E (Ramesh et al., 2021) boosts test-to-image generation task by synthesizing high-quality images from textual descriptions, showcasing the promising potential of the multimodal generation. Overall, these models not only scale up in terms of parameters and training data but also expand the scope of vision-language field into more complicated and challenging tasks, which brings vision language tasks from experiments to realistic applications.

2.2 Transfer Learning

Transfer learning has become a cornerstone technique in machine learning, which enables large models trained on massive datasets to generalize effectively to domain-specific tasks with limited data Socher et al. (2013); Zhuang et al. (2020); Rostami et al. (2022); Zhang et al. (2024c). This paradigm is particularly effective in domains where collecting data is costly or impractical. Early research on transfer learning focused on feature-based approaches, where pretrained models serve as feature extractors, often leveraging architectures such convolutional neural networks (CNNs) pretrained on large image datasets like ImageNet (Deng et al., 2009) for handling vision tasks. Finetuning has since emerged as a flexible and powerful approach, allowing the adaptation of pretrained models to target tasks by updating their parameters in a partial or full manner.

Transfer learning has diversified into several sub-domains, each addressing specific challenges, including: (i) **Domain Adaptation**, which focuses on transferring knowledge from a source domain with labeled data to a target domain with different distribution in which we only have unlabeled data. Techniques in domain adaptation often involve aligning feature spaces across domains. (ii) **Zero/Few-Shot Learning**, which aims to generalize the knowledge from seen classes to unseen classes or tasks without access or with very limited access to labeled training data for those classes. (iii) **Continual Learning**, which addresses the challenge of learning sequentially from a stream of tasks without forgetting previously learned tasks. (iv) **Multi-Task Learning**, which focuses on training models on multiple tasks simultaneously to encourage shared representations that help improve the overall tasks' performance. Transfer learning has evolved from simple feature reuse to sophisticated frameworks that span a diverse set of applications and challenges. Its sub-domains continue to address critical gaps for the case vision-language tasks, enabling models to perform robustly in dynamic and resource-constrained environments. Our goal is to survey transfer learning methods that are designed to be used on VLMs for handling vision-language tasks more efficiently.

As vision-and-language tasks continue to evolve, transfer learning has played a crucial role in improving model generalization across diverse datasets and applications. Early approaches relied on task-specific models trained from scratch or finetuned on relatively small datasets. However, with the rise of large-scale vision-language datasets and the success of self-supervised learning, a shift toward pre-trained large Vision-Language Models (VLMs) has emerged. These models, trained on large-scale multimodal corpora, serve as powerful foundations that can be adapted to downstream tasks with minimal finetuning. In the next section, we explore the architectures and training objectives of these pre-trained Large VLMs, highlighting how they enable transfer learning at scale and achieve state-of-the-art performance across vision-language benchmarks.

3 Pre-trained Large Vision Language Models

In recent years, pre-trained large Vision-Language Models have emerged as a powerful class of AI models that bridge the gap between computer vision and natural language processing. These models are trained on massive datasets containing both images and text, enabling them to understand and generate complex visual and textual content. Pre-trained large vision-language models are the fundamental modern tools to solve vision-language tasks and are the backbones that transfer learning methods in the area built on (Kim et al., 2021; Radford et al., 2021). In recent works, Du et al. (2022); Zhang et al. (2024b); Wu et al. (2023a); Liang et al. (2024) provide detailed summarization about pre-trained large vision language models. In our paper, to offer a more comprehensive insight about VLMs and help understand the transfer learning methods in this area, we survey the most popular VLM architectures in section 3.1 and the objective functions that are used for training them in section 3.2. Additionally, we present the most popular large-scale pre-trained VLMs in temporal order in Figure 1.

3.1 VLM Architectures

Various architectures have been designed to bridge the gap between visual and textual modalities (See Figure 2). A general approach to categorize VLM architecture is based on how the vision and text features are integrated, namely, **Contrastive Dual Encoder** architecture, **Encoder-Decoder** architecture and **Multi-modal LLM** architecture. We briefly survey these categories.

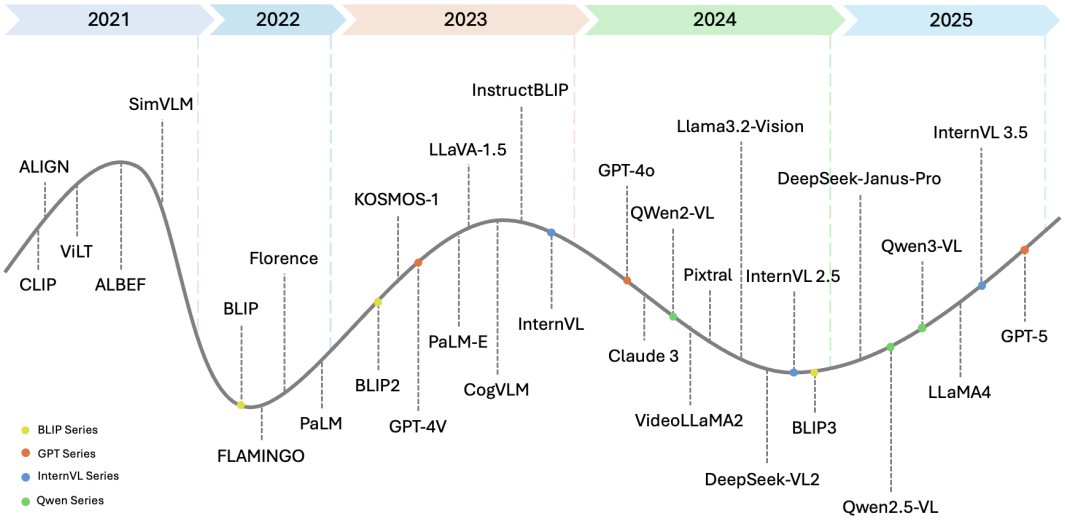

Figure 1: Development of large-scale pretrained vision language model

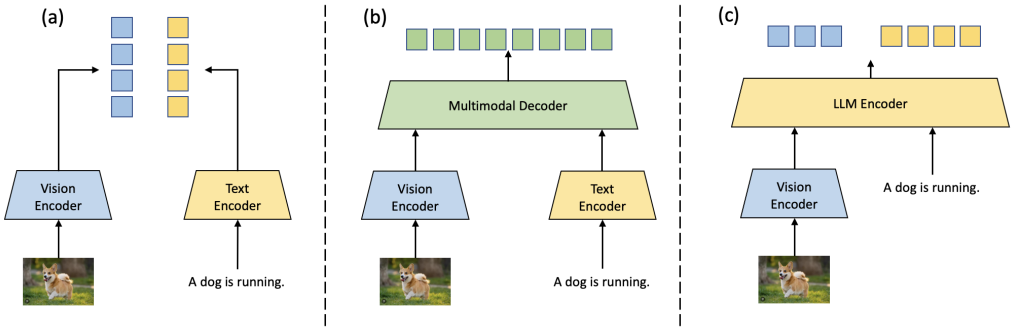

Figure 2: Diagram shows the three different architecture of pretrained large vision language model. **(a)** Contrastive Dual Encoder Architecture, **(b)** Encoder-Decoder Architecture, **(c)** Multi-modal LLM Architecture.

3.1.1 Contrastive Dual Encoder Architecture

Some vision-language models utilize independent vision and text encoders to generate features from different modalities and align them into the shared embedding space (see Figure 2 (a)). During pretraining, the model is optimized using a contrastive loss, such as InfoNCE loss, that encourages matched image–text pairs to have similar embeddings, while pushing apart mismatched pairs.

For example, Radford et al. (2021) use vision transformer (Dosovitskiy et al., 2021) as vision encoder and the BERT model (Kenton & Toutanova, 2019) as the text encoder. The encoders are further trained via a large-scale image-pair dataset with contrastive loss to guarantee the alignment of feature from both encoders for matching image and text features. Jia et al. (2021) adopt a similar architecture and train the model on a more general and noisy dataset to achieve more advanced performance. Yuan et al. (2021), on the other hand, expand the feature representation from scene-level to object level and enable the video and depth image retrieval. However, although the contrastive dual encoder architecture models are computationally efficient and well-suited for retrieval-style tasks (e.g., image–text retrieval or zero-shot classification), they generally lack fine-grained cross-modal reasoning since the encoders operate independently.

3.1.2 Encoder-Decoder Architecture

Different from contrastive dual encoder architecture models, the encoder–decoder architecture models integrate visual and textual information within a single multimodal transformer framework(see Figure 2 (c)). Both image and text inputs are first tokenized into unified embeddings, with a vision encoder producing patch tokens and a text tokenizer producing word tokens, which are then jointly processed by a shared transformer. This allows bidirectional cross-modal attention at every layer, enabling deep fusion and reasoning between modalities.

For example, Kim et al. (2021) use transformer blocks as multimodal decoder. The input image and text are first converted to feature tokens via the modality-specific encoder. The image feature and the text feature are concatenated together, after applying the modality-specific type embedding and the positional embedding, to form the input of the multimodal decoder. Li et al. (2021) take a further step such that they apply the image-text contrastive loss to the unimodal features before the fusion. During the pre-training, the model learns to align the image feature and the text feature from the same image-text pair, via the contrastive loss, before fusion. The image-text matching loss and the masked-language-modeling loss are then applied to learn the multimodal interaction between the image and the text.

However, although encoder-decoder architecture models are more powerful solving the reasoning task than the contrastive dual encoder architecture models, the multimodal decoders often smaller or specially-designed multimodal transformers, rather than very large language models (LLMs) with hundreds of billions of parameters. This limits their pure language reasoning capacity. On the other hand, as the current large-scale LLMs are naturally capable of complicated reasoning tasks, building a relatively small multi-modal decoder module on the top of the LLM may harm its performance on reasoning-centric tasks.

3.1.3 Multi-Modal LLM Architecture

To fully utilize the reasoning capacity of the SOTA large language models, multi-modal LLM architecture models aims to integrate the image features into the LLMs to enable multi-modal reasoning. In this design, image features extracted by a powerful vision backbone are projected into the LLM's token space and fused via cross-attention or lightweight connector networks. This enables the model to perceive and reason jointly across modalities while leveraging the extensive world knowledge, linguistic fluency, and reasoning capability of the underlying LLM. Compared with traditional encoder–decoder or unified multimodal architectures, multimodal LLMs offer superior scalability, modularity, and instruction-following ability, allowing the reuse of existing text-only LLMs and rapid adaptation to multimodal tasks through instruction tuning.

For example, LLaVA (Liu et al., 2023) adopts CLIP's ViT as the visual encoder and the pre-trained Vicuna model as the LLM backbone. The image features from the visual encoder are projected, by a trainable projection matrix, to the LLM backbone's embedding space and become part of the input to the LLM. BLIP-2 (Li et al., 2023) model, on the other hand, finetune a light-weight querying transformer (Q-Former) as the vision-language project to adopt the image feature into the LLM. BLIP-2 freezes the visual encoder and the LLM during the finetuning and only update the Q-Former, enabling a generic and efficient modality interaction. Flamingo (Alayrac et al., 2022) applies novel perceiver resampler as the vision-language projector and inserts *gated xattn-dense layers* in the LLM layers to align image feature into the text encoder.

After discussing the three major architectural paradigms of the VLMs, we now turn to the training objectives that guide how these models learn cross-modal alignment and reasoning. While architectures define the interaction between vision and language, training objectives determine the nature of this interaction—ranging from the contrastive alignment in dual encoders to the matching and generative learning in multimodal LLMs.

3.2 VLM Training Objective Functions

The training objective functions for VLMs play a pivotal role in enabling effective cross-modal alignment and understanding. We summarize the commonly used objective functions into three categories, including contrastive objectives, generative objectives and matching objectives.

Contrastive objectives aim to align the representations of different modalities within the same pair closer to each other and to be distant from the representations of the other pairs. The most common contrastive loss used for training VLMs is based on the InfoNCE loss (Radford et al., 2021; Jia et al., 2021; Chen et al., 2024b) which aims to maximize the alignment of features for positive pairs while minimizing the similarity to the features of negative pairs. The image to text loss is expressed as:

$$\mathcal{L}_{\text{I2T}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(f_I(x_i), f_T(y_i))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f_I(x_i), f_T(y_j))/\tau)}, \tag{1}$$

while the text to image loss is:

$$\mathcal{L}_{\text{T2I}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\text{sim}(f_T(y_i), f_I(x_i))/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(f_T(y_i), f_I(x_j))/\tau)} \tag{2}$$

By combining the two, the final contrastive loss is formed as below:

$$\mathcal{L}_{\text{contrastive}} = \mathcal{L}_{\text{I2T}} + \mathcal{L}_{\text{T2I}}, \tag{3}$$

where $f_I(x_i)$ represents the embedding of image x_i, $f_T(y_i)$ represents the embedding of text y_i, and τ is the temperature hyper-parameter to control the sharpness of the softmax distribution.

Generative objectives generative objectives for image modality are easier to define because visual data has a clear, structured pixel-based representation that allows straightforward metrics like pixel-wise differences Goodfellow et al. (2014); Shin et al. (2017); Rostami et al. (2019). In addition to bi-modality, the challenge for defining generative objectives for VLMs is that language data is not as structured as visual data. let the model learn the semantic property of features to generate the corresponding image or text for either prediction or filling a missing part (Chowdhery et al., 2023; Driess et al., 2023; Achiam et al., 2023). Masked language modeling (MLM) is a common approach to learn the probability distribution of the sequences of words or tokens in a given context by measuring how well a model can predict the word in a sequence given the preceding or surrounding context. The objective function is denoted as:

$$\mathcal{L}_{\text{LM}} = -\frac{1}{T}\sum_{t=1}^{T}\log P(w_t \mid w_1, w_2, \ldots, w_{t-1}), \tag{4}$$

where w_t is the ground truth next token at position t.

On the other hand, masked image modeling (MIM) (Li et al., 2022b; Wang et al., 2021) is used to learn the image representation by predicting the masked part of the image patch as:

$$\mathcal{L}_{\text{MIM}} = -\sum_{i\in M}\log P(v_M \mid v_U), \tag{5}$$

where v_M is the masked token and v_U represents the unmasked tokens.

Matching objectives help understand the relationship between images and their corresponding text descriptions. The goal of image-text matching is to determine whether a given image and a piece of text (e.g., a caption or description) are semantically aligned or not. Different from contrastive objective, matching objectives are generally a type of binary classification loss instead of measuring similarity as:

$$\mathcal{L}_{\text{ITM}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right], \tag{6}$$

where y_i is the ground truth that 1 means match and 0 for otherwise, and $\hat{y}_i$ is the predicted probability that whether the paired image and text matches with each other.

As pre-trained Large VLMs continue to advance, their widespread adoption has highlighted both their strengths and challenges. While these models achieve state-of-the-art performance across various vision-language tasks, the massive scale introduces significant computational and deployment costs, making finetuning or adapting them for specific applications impractical in many scenarios. To address these limitations, efficient transfer learning methods have been developed to leverage the rich knowledge encoded in large VLMs while reducing resource consumption. In the next section, we explore different methods which enable effective model adaptation with minimal compute- and temporal- cost, making large VLMs more accessible and scalable for real-world applications.

4 Transfer Learning Approaches for Vision-Language Tasks

Transfer learning aims to enable large models, which are trained on massive datasets, to generalize effectively to domain-specific tasks with limited data. In this section, we categorize the efficient transfer learning approaches for VLMs into four categories, namely **Adapter Based Method** in section 4.1 , **Prompt Based Method** in section 4.2, **Knowledge Based Method** in section 4.4 and **Pipeline Based Method** in section 4.3. We summarize methods in each category.

4.1 Adapter-Based Methods

End-to-end finetuning is inefficient for large models and often would lead the model to become overfit on the finetuning datasets. Adapter-based transfer learning methods introduce trainable small modules during the finetuning stage to enable efficient adoption of pre-trained models to new tasks without updating all the parameters Houlsby et al. (2019); Jin et al. (2021b); Sung et al. (2022). Houlsby et al. (2019) insert small trainable feedforward networks modules in every transformer blocks, and consider those moducles to be **adapters**. Since only a small number of parameters are updated during finetuning, adapter-based methods significantly reduce the memory requirement for model storage and computational cost when finetuning a pretrained VLM on several downstream tasks. When dealing with VLMs, adapters are not constrained to layer-wise FFNs only. Among all the vision-language adapter-based transfer learning methods, we further divide them into **layer-wise** adapters, which are repeatedly inserted in layers of the original model, and **module-wise** adapters, which is the independent modules that only plugged into the VLM once.

4.1.1 Layer-Wise Adapter

The Layer-wise adapters are often inserted into transformer layers, shown in Figure3(a). The adapter layers receive the output from the preceding transformer blocks as their input and generate a light-weight transformation. The transformed features are then passed into the next transformer layer.

Li & Sun (2023) propose LiFT-Adapter which consists of two layers: bottleneck one fully-connected layer and one non-linear activation function. The adapter modules are inserted after the MLP layers in the transformer blocks of the vision encoder while the vision encoder remains frozen. Yang et al. (2022a), similarly, apply adapters on both of the frozen pre-trained vision model and the language model. As the backbones are frozen, only the adapters and normalization layers for each transformer block are updated during the finetuning stage, enabling both mitigating catastrophic forgetting and modality alignment. While the above two methods apply regular static adapter to the base-model's architecture, other methods propose more different designs of adapter. Chen et al. (2025a) introduce a shared-weight multi-modal spatio-temporal adapter (MSTA) for video-text matching. MSTA introduces a spatio-temporal description-guided consistency constraint such that in the vision encoder branch, the output of adapter goes to both the spatial up-sampling layer and the temporal up-sampling layer to mitigate over-fitting and enhance generalizability. Zhang et al. (2023b) and Gao et al. (2023) introduce llama-adapter, for LLaMA (Touvron et al., 2023) based models, with adaptation prompt and gated zero-init attention to efficiently finetune the LLaMA model for a variety of downstream tasks. Yu et al. (2024) introduce MoE-Adapters for continual learning. MoE-Adapters contain multiple adapters, inserted after the multi-head attention block of each transformer layer in both the visual encoder and the text encoder. The adapters are selected by the task-specific routers for different tasks. Wu et al. (2024), on the other hand, propose Dynamic Architecture Skipping by utilizing adapters to replace parts of the vision-language layers. The method observes the significance of each module in a reinforcement

Category	Sub-Category	Method	Transfer Learning Domain	Downstream Task
Layer-Wise	-	LiFT-Adapter (Li & Sun, 2023)	Few Shot Learning, Novel Class Discovery	Image Classification
		LLaMA-Adapter (Zhang et al., 2023b)	Efficient Finetuning	Multi Modal Reasoning
		LLaMA-Adapter v2 (Gao et al., 2023)	Efficient Finetuning	Multi Modal Reasoning
		MoE-Adapter (Yu et al., 2024)	Continual Learning	Image Classification
		CLIP-LoRA (Zanella & Ben Ayed, 2024)	Few Shot Learning	Image Classification
		DAS (Wu et al., 2024)	Efficient Finetuning	VQA, NLVR, Retrieval
		BiLM (Yang et al., 2022a)	Zero Shot Learning	Video Question Answering
		MSTA (Chen et al., 2025a)	Efficient Finetuning	Video Text Retreival
		MMA (Yang et al., 2024)	Domain Generalization	Image Classification
Module-Wise	Cross-Modal	Tip-Adapter (Zhang et al., 2021)	Few Shot Learning, Efficient Finetuning	Image Classification
		CLIP-Adapter Gao et al. (2024)	Few Shot Learning, Efficient Finetuning	Image Classification
		TDA (Karmanov et al., 2024)	Domain Adaptation	Image Classification
		Meta-Adapter (Song et al., 2023)	Few Shot Learning	Image Classification
		Prompt-Aware Adapter (Zhang et al., 2025)	Efficient Finetuning	Visual Question Answering
		Cross-Modal Adapter (Jiang et al., 2025)	Efficient Finetuning	Image Text Retrieval
	Uni-Modal	SgVA-CLIP (Peng et al., 2023)	Few Shot Learning	Image Classification
		SVL-Adapter (Pantazis et al., 2022)	Few Shot Learning	Image Classification
		RAIL (Xu et al., 2024)	Continual Learning	Image Classification

Table 1: Summary of adapter-based methods

learning manner, and the less significant module is skipped, or replaced by adapters to avoid drastic change in hidden features from previous layers. Yang et al. (2024) propose a multi-modal adapter (MMA) for Clip-based few shot learning. MMA connects the frozen higher-layers of text encoder and visual encoder through a shared projection layer and makes features more discriminative and generalizable.

Moreover, other than providing novel designs such as those mentioned above, Zanella & Ben Ayed (2024) offer analysis of applying LoRA Hu et al. (2021) to enable few-shot learning with VLMs to explore the application of LoRA-like adapters in vision-language transfer learning. Similarly, Sung et al. (2022) analyze the performance of different parameter-efficient training techniques, such as adapter, hyperformer and compactor, for image-text and video-text tasks. The methods are summarized in Table 1 layer-wise section.

4.1.2 Module-Wise Adapter

Different from layer-wise adapters which are repeatedly added to the original model, module-wise adapters take a different strategy through designing independent modules outside of transformer layers Song et al. (2023); Cai & Rostami (2024a); Gao et al. (2024); Zhang et al. (2021). The designed module is usually used only once. The adapter is then plugged in between the original backbone's modules, e.g., between the visual encoder and the multimodal encoder. Compared to the layer-wise adapters, module-wise adapters aim for high-level feature alignment and are more modular and flexible. As adapters are independent and used only once, they increase the scalability for large models. Within module-wise adapters, they can be further categorized into cross-modal adapters and uni-modal adapters.

Cross-Modal Adapter Cross-modal adapters take both image and text modalities as inputs, either by a single adapter layer or separate adapter layers. The general architecture is shown in Figure3(b). Jiang et al. (2025) introduce a cross-modal adapter, it enables encoder-level cross-modal interactions by sharing adapters' weights between two modalities rather than introducing explicit feature interactions. Such a scheme allows an implicit cross-modal interaction, which will facilitate the re-alignment of vision and language feature spaces for the cross-modal retrieval task. Song et al. (2023) propose a single cross-modal adapter layer for aligning inputs from vision and text encoders using the attention-based adapter to refine the CLIP features, guided by a few samples in an online manner. The adapter layer takes queries from the support images and Key/Query from text labels to refine the category embedding for the target image. On the other hand, Clip-Adapter (Gao et al., 2024) adopts two separate adapters, one for each modality, which consist of small bottleneck linear layers, $Av(\cdot)$ and $At(\cdot)$, to handle image features and text features. Inspired by Clip-Adapter, Zhang et al. (2021) adopt the idea of multi-layer perceptron and residual connection, but also propose a novel cache-based model to obtain the weights from few-shot visual features and ground truth labels. Following the idea of cache-based method, Training-free Dynamic Adapter (TDA) (Karmanov et al., 2024) expands the design of cache models. It constructs and updates two key-value caches to store the knowledge of a stream of test samples, and uses the two caches to generate positive and negative predictions which are combined with the CLIP predictions to produce the final prediction. Zhang et al. (2025) propose

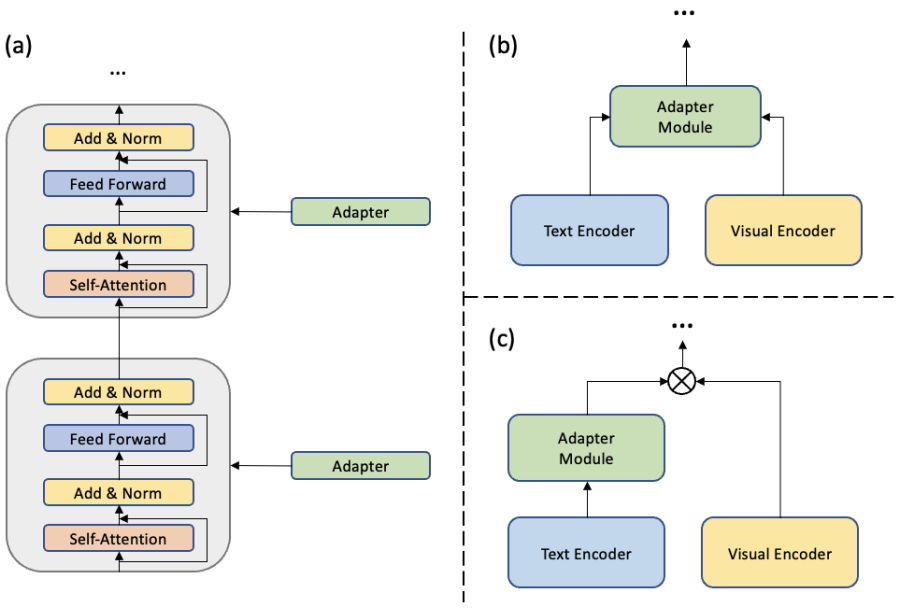

Figure 3: Taxonomy of Adapter-based Method. **(a):** layer-wise adapter, **(b):** cross-modal adapter, **(c):** uni-modal adapter.

a novel prompt-aware adapter to align the image and the text input. It is designed to dynamically embed the visual inputs based on the specific focus of the text prompt, by using both global and local textual features to capture the most relevant visual clues from the prompt at both coarse and fine granularity levels. The methods are summarized in Table1 cross-modal section. TAM-CL Cai et al. enables continual learning through a task-attentive layer which receives its input form the cross-modal transfer layers and adapt the model to generalize on a sequence of continually observed tasks.

Uni-Modal Adapter Different from cross-modal adapters, uni-modal adapters take one modality as input to boost the generalizablity of backbone model in transfer learning scenarios. The diagram is shown in Figure3(c). Existing uni-modal adapters focus on the enhancement of visual features. Xu et al. (2024) design regression-based analytic incremental learning (RAIL) which utilizes a recursive ridge regression-based visual adapter to learn from a sequence of domains in a non-forgetting manner and decouple the cross-domain correlations by projecting features to a higher-dimensional space. Similarly, Pantazis et al. (2022) propose a vision-only adapter that takes input from an additional CLIP visual encoder. The adapted visual features are used to fuse with the output at classification level to guide zero-shot and low-shot image classification. Moreover, Peng et al. (2023) adopt a two-layer perceptron as vision adapters to support visual-specific contrastive loss between query images and support images in few-shot learning setting. By adopting visual adapters, visual features can be better aligned with the text features extracted by LLMs, which in turn improve the overall performance of VLMs. The methods are summarized in Table 1 uni-modal section.

4.2 Prompt-Based Methods

Prompt-based methods enable transfer learning of VLMs using additional image or text features and trainable variables to guide the VLM to perform the downstream task without extensive finetuning Ge et al. (2023); Jia et al. (2022); Qian et al. (2023); Cai & Rostami (2024b) . By freezing the backbone model, prompt-based methods adopt the original knowledge from pretrained VLMs to the downstream task with the guidance of well-designed additional inputs and thus make these methods parameter-efficient and flexible. In this section, we survey textual prompts in Sec 4.2.1, visual prompts in Sec 4.2.2 and multimodal prompts in Sec 4.2.3.

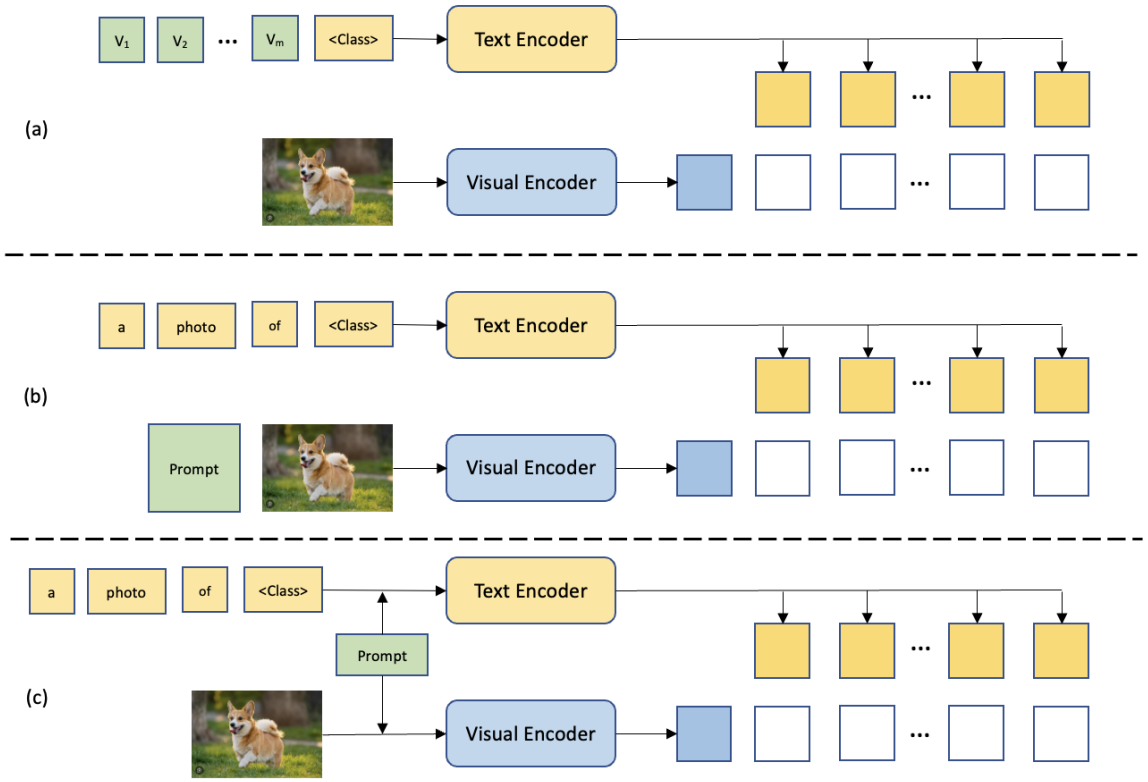

Figure 4: Taxonomy of Prompt-based Methods. **(a):** Textual Prompt, **(b):** Visual Prompt, **(c)**: Multimodal Prompt.

Category	Method	Transfer Learning Domain	Downstream Task
Textual Prompt	PLOT (Chen et al., 2022a)	Few Shot Learning	Image Classification
	FEWVLM (Jin et al., 2021a)	Zero/Few Shot Learning	Image Classification, Image Captioning, VQA
	CoCoOp (Zhou et al., 2022a)	Domain Generalization	Image Classification
	UDA (Ge et al., 2023)	Domain Adaptation	Image Classification
	CoOp (Zhou et al., 2022a)	Few Shot Learning	Image Classification
	MPA (Chen et al., 2023)	Unsupervised Domain Adaptation	Image Classification
	PODA (Fahes et al., 2023)	Zero Shot Learning, Domain Adaptation	Semantic Segmentation
	ProDA (Lu et al., 2022)	Few Shot Learning	Image Classification, Object Detection
	ProGrad (Zhu et al., 2023)	Few Shot Learning, Domain Generalization	Image Classification
	StyleGAN-NADA (Gal et al., 2022)	Domain Generalization	Image Generation
	TPT (Shu et al., 2022)	Zero Shot Learning, Domain Adaptation	Image Classification
	TriMPL (Liu et al., 2024c)	Domain generalization, Few Shot Learning	Image Classification
	PMP (Liu et al., 2025)	Domain Generalization, Few Shot Learning	Image Classification
Visual Prompt	Visual Prompt (Bahng et al., 2022)	Efficient Fine Tuning	Image Classification
	SHIP (Wang et al., 2023d)	Zero Shot Learning	Image Classification
	RePrompt (Rong et al., 2023)	Few Shot Learning, Domain Generalization	Image Classification
	VPT (Jia et al., 2022)	Efficient Fine Tuning	Image Classification
	VL-PTM (Wen et al., 2022)	Few Shot Learning	Text Classification
	CPT (Yao et al., 2024)	Few/Zero Shot Learning	Visual Relation Detection, Visual Reasoning, VQA
Multimodal Prompt	APT (Wu et al., 2023b)	Efficient Fine Tuning	Text-Image Generation, Image Classification
	TRIPLET (Qian et al., 2023)	Continual Learning	Visual Question Answering
	MaPLe (Khattak et al., 2023)	Domain Generalization	Image Classification
	MVLPT (Shen et al., 2024)	Few Shot Learning, Domain Generalization	Image Classification
	PADCLIP (Lai et al., 2023)	Unsupervised Domain Adaptation	Image Classification
	PMHANet (Liu et al., 2022)	Efficient Fine Tuning	Image Classification
	UPL (Huang et al., 2022)	Efficient Fine Tuning	Image Classification
	ViTiS (Engin & Avrithis, 2023)	Zero/Few Shot Learning	Visual Question Answering
	HiCroPL (Zheng et al., 2025)	Efficient Fine Tuning	Image Classification

Table 2: Prompt-based methods

4.2.1 Textual Prompt

Textual prompts, presented in Figure4(a), are the natural language input or the add-on parameters to the textual inputs or features which aim to guide the model's understanding of the input task by enriching the semantic knowledge of the given input text. It has been observed that enriching input text can help understanding the vision modality better and can improve generalization Shangguan et al. (2025b).

The most intuitive textual prompt is based on the human-readable natural language. Ge et al. (2023) introduce text-based DAPrompt for image domain adaption tasks, which aims to embed domain information into domain-specific prompts, which is a form of representation generated from natural language, and then align the image features of the same domain and push them distant from the ones of other domains. Fahes et al. (2023) propose PODA for zero-shot image domain adaptation using prompts. PODA uses the natural language description of target image to match the source image via CLIP-based model, and bring the source image features in CLIP feature space closer to its imaginary counterpart in the target domain. Besides using the natural language as the textual prompt, other methods rely on learnable context parameter to enrich the textual features. Zhou et al. (2022b) propose Context Optimization (CoOp) such that instead of adopting a context such as "a photo of" or any other phrase for CLIP text input, the method turns the context into learnable vectors. By freezing the text encoder and the image encoder, the learnable contexts are updated through the back-propagation and are made to be efficient for few-shot learning. Following CoOp, Conditional Context Optimization (CoCoOp) (Zhou et al., 2022a) extends the idea by further learning a light weight neural network that generates an input-conditional token for each image, combined with the learnable context vector. Zhu et al. (2023) develope another variant of CoOp by training learnable context with gradient regularization strategy. The method first measures the gradient *general direction*, which is the KL loss between the zero-shot original CLIP and the few-shot finetuned model, and then decomposes the gradient into components which are orthogonal to the *general direction* and the gradient parallel to the *general direction*. The prompts are then updated accordingly. Similarly, Shu et al. (2022) provide Test-time Prompt Tuning (TPT). TPT tunes the prompts, which are learnable context features, on the fly using only the test sample. The tuned prompt is adapted to each task, making it suitable for zero-shot generalization without requiring any task-specific training data or annotations.

While the above methods adopt the single prompt, other methods utilize the combination of multiple prompts. Chen et al. (2022a) introduce multiple learnable context vectors as the prompt and use optimal transport as the metric to align the prompt features with the local features of an image, achieving the alignment of different modalities in a more fine-grained and comprehensive way. Similarly, ProDA (Lu et al., 2022) generates multiple learnable prompts. By integrating the text with all the prompts, ProDA forms a weight distribution of the output embedding features for each object class in the feature space. The image features are then aligned with the best-matched distribution. Chen et al. (2023) provide a multi-prompt method for unsupervised domain adaptation. For source domain and target domain, they propose both domain-invariant features and domain-specific features and align them with the input text to learn domain-shared and domain-specific knowledge. FEWVLM (Jin et al., 2021a) provide a prompt-based method for low resource learning via analyzing the performance of different vision-lamguage tasks on zero/few shot learning with different prompts. On the other hand, Liu et al. (2025) propose a progressive implementation of multi-prompt (PMP). PMP introduces multiple prompts in a step-by-step manner to focus on various information and utilizes a late attaching mechanism to defer the interactions of prompts and features to a deeper encoding layer to reduce the overfitting. All the methods are summarized in Table 2 textual prompt section.

4.2.2 Visual Prompt

Visual prompts refer to the prompts attached to the visual inputs to augment the visual semantics. While textual prompts can be crafted manually, visual prompts often require complex transformations that need to be learned or optimized. The diagram of the visual prompt is presented in Figure 4(b).

Similar to the textual prompt based method, some visual prompt based methods adopt trainable variables on the vision side as visual prompts. Jia et al. (2022) propose Visual Prompt Tuning (VPT) for downstream recognition tasks. VPT prepends learnable prompt parameters to each visual encoder transformer layer's input while keeping the whole model frozen. Bahng et al. (2022) propose a trainable visual prompt for input image. The image feature is summed with the visual prompt to form a prompted image, which is then goes through the frozen visual encoder.

Other than directly applying trainable parameters to the visual input, some methods explore generating pseudo images as visual prompts. Wen et al. (2022) freeze CLIP and use the class name to generate pseudo images. The pseudo image features are combined with the class name to match with the text content

through CLIP-style contrastive loss. Wang et al. (2023d) introduce Synthesized Prompt (SHIP) that trains a generator model to synthesize image features. The image features from the visual encoder are sent to a variantional autoencoder (VAE) and are confined to a prior distribution. The sample from the distribution is then sent to the text encoder, prepending to the class name, for feature reconstruction.

Different from methods above, some methods apply existing images as the visual prompts. Rong et al. (2023) propose retrieval-enhanced visual prompt learning (RePrompt) for few-shot classification with additional realistic images. RePrompt constructs a retrieval database from either training examples or external data, if available, and uses a retrieval mechanism to enhance multiple stages of a simple visual prompt learning baseline, thus narrowing the domain gap. Yao et al. (2024) propose the colorful prompt tuning (CPT) which reformulates visual grounding into a fill-in-the-blank problem with the color-based co-referential markers in image and text. CPT masks the proposed region with the natural visual marker of a different color as visual sub-prompt, and rewrite the text to select the color of bounding box of the targeted object as text sub-prompt. Hence, the visual grounding can be reformulated into a simple fill-in-the-blank problem. All the visual prompt methods are summarized in Table2 visual prompt section.

4.2.3 Multi-modal Prompt

Unlike uni-modal prompts that rely solely on either vision or language, multi-modal prompts provide a cohesive mechanism to influence how models process and align information across both modalities. The architecture of multi-modal prompt is shown in Figure 4(c).

An intuitive approach to generate the multi-modal prompt is to treat the prompt from both modalities separately, such that the prompt of each modalities remains relatively independent and balanced. Qian et al. (2023) introduce a multimodal prompt for VLMs with multi-modal encoders such as ALBEF (Li et al., 2021) for the continual learning of VQA tasks. By freezing the backbone, they train the visual prompt for image input, the text prompt for text input, and the fusion prompt for multi-modal encoder input. The fusion prompt is made by the weighted interaction of the two single-modality prompts, where each prompt is divided into the task-agnostic general prompt G and the task-specific expert prompt E. Similarly, Cai & Rostami (2024b) propose CluMo, which is also adopted for continual learning of VQA tasks using VLMs with a multi-modal encoder. Instead of adopting uni-modal prompt for fusion prompt, CluMo designs uni-modal prompt keys, which are trained via K-means clustering on the training data. The visual prompt key and the text prompt key are then combined to determine the best matched fusion prompt which is used for the multi-modal encoder. Zang et al. (2022) propose the Unified Prompt Tuning (UPT), which learns a tiny neural network to jointly optimize the prompts across different modalities. UPT designs a layer-wise unified modal-agnostic prompt that can be split into the corresponded text prompt and image prompt, which are separately input to the text encoder and the image encoder.

Other than treating the different modalities' prompt relatively independently, some methods rely on generating the prompt of one modality conditioned on the other modality's prompts and features. For example, Huang et al. (2022) provide a novel design of unsupervised learning in multi-modal prompt setting for unlabeled datasets. While the text prompt is shared learnable prompt, the method applies CLIP-style encoding for both the image and the text inputs to generate pseudo-labels for unlabeled images, which serve as the visual prompt for the visual input. Khattak et al. (2023) propose Multi-modal Prompt Learning (MaPLe) which attaches the text prompt to the text input, attaches the visual prompt to the image input. The image prompt is conditioned on the text prompt via a vision-language coupling function, which is trained along with the prompt. Similarly, Engin & Avrithis (2023) propose ViTiS for few-shot video question answering. It trains a visual mapping network via a LLM so that the visual prompts are transformed, conditioned on the text, into the textual feature space and input to the frozen language model. ViTiS also proposes a novel attention-level text prompt such that the prompts are used as inputs to every attention layer. Zheng et al. (2025), on the other hand, introduce a Hierarchical Cross-modal Prompt Learning (HiCroPL) framework that establishes bidirectional knowledge flow between text and vision modalities. It adopts a layer-wise modality fusion strategy such that in early layers, text prompts inject relatively clear semantics into visual prompts through a hierarchical knowledge mapper, enhancing the representation of low-level visual semantics. In later layers, visual prompts encoding specific task-relevant objects flow back to refine text prompts, enabling deeper alignment. All the methods are summarized in Table 2 multimodal prompt section.

Category	Method	Transfer Learning Domain	Downstream Task
Cross-Model Interaction	CLAP (Jha et al., 2024)	Continual Learning	Image Classification
	VT-CLIP (Qiu et al., 2021)	Few Shot Learning	Image Classification
	VL-Few (Ma et al., 2024)	Few Shot Learning	Visual Question Answering
	-(Chen et al., 2022b)	Domain Adaptation	Visual Understanding
	MAPL (Mañas et al., 2022)	Few Shot Learning	VQA, Image Captioning
	CDCIN (Zhang et al., 2024a)	Few Shot Learning	Visual Question Answering
	CIRPLANT (Liu et al., 2021)	Domain Adaptation	Image Retrieval
	LingoCL (Ni et al., 2024)	Continual Learning	Image Classification
	CaFo (Zhang et al., 2023c)	Few Shot Learning	Image Classification
	RanPAC (McDonnell et al., 2024)	Continual Learning	Image Classification
	TaskRes (Yu et al., 2023)	Domain Generalization	Image Classification

Table 3: Summary of pipeline-based methods

4.3 Model-Based Methods

In the subsections above, we discussed the adapter-based methods and the prompt-based methods, and both of them introduce extra learnable parameters to boost the model's performance on downstream tasks. Besides inserting additional parameters, the model-based methods focus on the modification of the model's architecture itself and come up with the novel design of the model's training pipeline, loss objective and architecture. In this section, we survey and categorize these methods into cross-modal pipeline and novel-architecture pipeline methods.

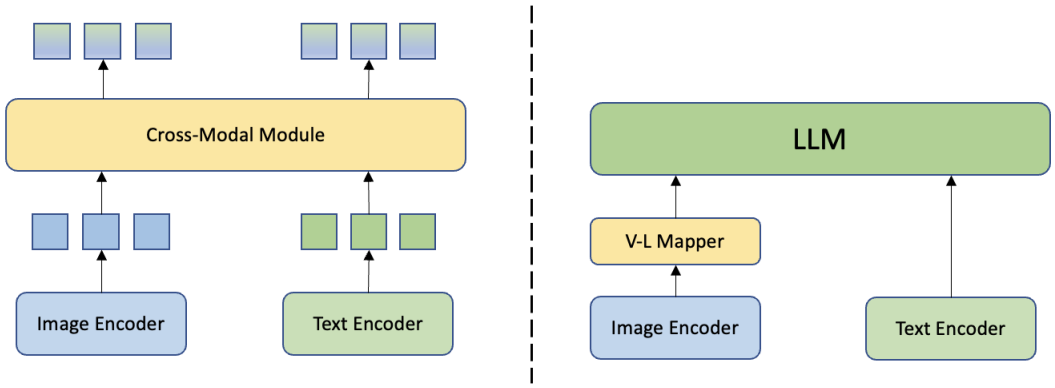

Figure 5: Diagram for different model-based cross-modality pipelines. **(left)**: Methods that use features from one modality to guide the features of another modality. **(right)**: Methods that map the visual features into the textual space via LLM.

A group of methods apply cross-modality pipelines by using features from one modality to guide the features of another modality. Jha et al. (2024) propose probabilistic modeling for downstream tasks. Using the designed visual-guided attention inference module, the proposed method learns the functional space of task-specific posterior distributions based on text features that are aligned with their visual counterpart. Similarly, Qiu et al. (2021) propose visual-guided attention module that takes text features as query and image features as key and value. The module contains co-attention layers that helps explore the informative regions of image that is related to the text of different category, and associates them with more attention weights. While the above two methods use visual-guided textual features, Zhang et al. (2024a) propose a textual-guided visual feature method, named *visual information entropy*, which represents the spatial distribution of visual features guided by questions. To minimize visual information entropy, a multi-modal feature adaptive module is designed to improve cross-modal interaction which calculates the visual information entropy before and after interaction. The novel information consistency loss is then used to minimize the difference between the visual information entropy of the two stages.

Another approach to ensure cross-modal interaction is to map the visual features into the textual space via LLMs. Ma et al. (2024) explicitly extract the cross-modal token through a vision mapper and a vision-language mapper. After the encoder, the vision feature is sent to the vision mapper to extract vision representation alignment token, while vision features and text features are sent to the vision language

Category	Method	Transfer Learning Domain	Downstream Task
Knowledge-Based	Selective Decomposition (Khan et al., 2024)	Zero Shot Learning	Visual Question Answering
	KGENVQA (Cao & Jiang, 2024)	Zero Shot Learning	Visual Question Answering
	ZPVQA (Hu et al., 2025a)	Zero Shot Learning	Visual Question Answering
	Prophet (Yu et al., 2025b)	Efficient Fine Tuning	Visual Question Answering
	PromptCAP (Hu et al., 2023)	Efficient Fine Tuning, Zero Shot Learning	Visual Question Answering
	Img2LLM (Guo et al., 2023)	Zero Shot Learning	Visual Question Answering
	RQP (Lan et al., 2023)	Zero Shot Learning	Visual Question Answering
	ViDIL (Wang et al., 2022)	Few Shot Learning	Video Lanuguage Tasks
	cross-modal adaptation (Lin et al., 2023b)	Few Shot Learning	Image Classification
	PNP-VQA (Tiong et al., 2022)	Zero Shot Learning	Visual Question Answering
	F-VQA (Wu et al., 2023c)	Zero Shot Learning	Visual Question Answering
	R2A (Pan et al., 2023)	Zero Shot Learning	Video Question Asnwering
	LADS (Dunlap et al., 2023)	domain adaptation	Image Classification
	Cross-Modal Adaptation (Lin et al., 2023b)	Few Shot Learning	Image Classification
	Z-LaVI (Yang et al., 2022c)	Zero Shot Learning	language understanding
	protext (Khattak et al., 2024)	Efficient Fine Tuning	Image Classification

Table 4: Transfer learning methods based Knowledge

mapper to extract vision-language representation token. The two tokens, along with task instruction and text features, are then sent to a multimodal fusion mapper for alignment across different modalities. Similarly, Chen et al. (2022b) use a V-L mapper that maps the visual features into the language feature space, which allosw the vision and language module to be progressively and independently pre-trained. Mañas et al. (2022) propose a method to adopt pre-trained unimodal models for VL tasks without intensive training or finetuning. It directly adopts a frozen pre-trained vision encoder and frozen LLM, connecting them with a trainable mapping network, which projects the visual feature into the dimension of text features. The mapping network is then updated with an image captioning objective. Liu et al. (2021) propose a cross-modal interaction method for composed image retrieval. While the images are encoded through ResNet, the reference image is sent to the VLM, OSCAR (Li et al., 2020), along with the tokenized text. After alignment within the VLM, the image feature token is compared with the candidate target image feature tokens to select the top match. The cross-modal methods are summarized in Table3 cross-modal interaction section.

Some methods, similar to adapter-based and prompt-based methods, propose additional novel modules to the VLM, but these novel modules are different from adapter, prompts or any other seen methods. For example, McDonnell et al. (2024) adopt random projection for continual learning tasks. The random projection layer, along with non-linear activation function, is inserted between the pre-trained model's feature extraction and output head in order to capture interactions between features with expanded dimensionality, providing enhanced linear separability for class-prototype-based CL. Yu et al. (2023) introduce a TaskRes for expanding the classifier for efficient transfer learning. Different from using prompts or adapters that update the classifier through training, TaskRes freezes the whole model, including the classifiers. It adds tunable parameters which are not dependent on based classifier, directly to the classifier by weighted sum to achieve better old knowledge inheritance while flexible task-specific knowledge exploration. Ni et al. (2024) propose LingoCL which relies on the significance of the semantic information within the label names while others encode label as one-hot label. LingoCL uses a pre-trained language model to generate semantic target based on label name, and uses the output features as the weights for the frozen classifier, guiding the learning of encoders.

Other than creating novel modules, Zhang et al. (2023c) introduce a few-shot learning pipeline, CaFo, by cascade combination of the prior-knowledge from four existing pre-trained models. Given the input image-text pair, CaFo firstly utilizes GPT-3 to generate diverse textual descriptions of the input image, and uses DALL-E to generate synthesized few-shot training images which are used to build cache of prior-knowledge with the visual output from both CLIP and DINO. During inference, the target image is sent to both CLIP and DINO to generate visual features, which are used to retrieve knowledge from cache for prediction.

4.4 Knowledge-Based Methods

Different from all the methods above, which rely on improving the model to better understand the input semantics for better performance, the knowledge-based methods leverages pre-existing knowledge from large-scale pre-trained models or external sources of knowledge to improve performance on new, unseen tasks or domains that combine visual and textual data. The diagram shown the general architecture can be found in

Figure 6, in which the inputs are sent to a pre-trained LLM to generate supplement knowledge to enhance the semantic information.

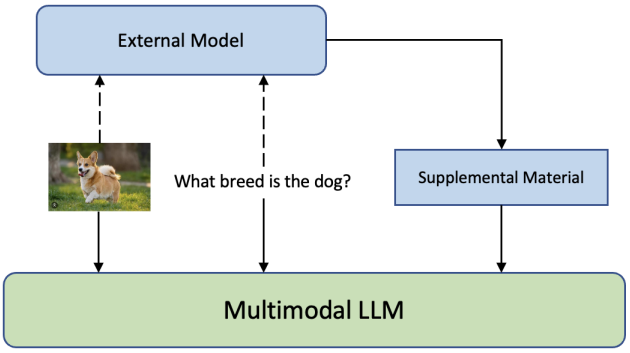

Figure 6: Taxonomy of knowledge-based Methods. For knowledge-based methods, the input image, text, or both, is sent to an external LLM to retrieve the supplemental knowledge. The supplemental knowledge are then set to the model to help the comprehensive understanding of task.

Knowledge-based methods are widely used in visual question answering (VQA) tasks. Cao & Jiang (2024) propose KGenVQA to enable the zero-shot learning for VQA. KGenVQA converts the image into a text description via a LLM, and then uses the LLM-generated description and question to form the initial knowledge. The selected knowledge, along with the question and description, are input to the LLM for answer generation. Hu et al. (2023) propose a method that generates captions based on the image and provide it as additional knowledge to align the different modalities to improve the model's performance on VQA tasks. Hu et al. (2025a), similarly, augment the text input with the LLM's generation. It generates caption prompts from images, along with samples of synthesized question-answer pairs to serve as example prompts in context, thus enhancing the model's understanding of the associations between the images and the questions. Yu et al. (2025b) introduce a framework, Prophet, designed to prompt LLM with answer heuristics for knowledge-based VQA. Prophet first trains a vanilla VQA model on a specific knowledge-based VQA dataset without external knowledge. Then, it extracts two types of complementary answer heuristics from the VQA model: answer candidates and answer-aware examples. The two types of answer heuristics are jointly encoded into a formatted prompt to facilitate the LLM's understanding of both the image and question, thus generating a more accurate answer. Khan et al. (2024) propose a question decomposition strategy to provide external knowledge and offer second-guessing for a LLM to answer the question. Given the question and the image, they first decompose the question into several sub-questions and provide a sub-answer for each of the questions. The generated sub-questions and sub-answers are served as the context, along with original question, of the input to the LLM. Guo et al. (2023) offer an additional text input to the text encoder by generating captions using their proposed Img2LLM module. The Img2LLM generates an image caption based on a text-guided image feature and exemplar question-answer pair. Similarly, Lan et al. (2023) rely on image-guided caption generation for solving zero-shot VQA tasks. The method contains two stages. On the first stage, the image caption is generated, and the caption then is combined with the question to generate image-guided questions and the answer accordingly. On the second stage, the generated answer with its corresponding confidence score is used as the answer heuristic to guide the final output. Tiong et al. (2022) introduce plug-and-play VQA (PNP-VQA). It contains an image-question matching module to generate the attention map on the image guided by question, and adopts the attention-enhanced image to generate a set of captions. The most related captions are combined with the question into the question-answering module for answer generation. Liu et al. (2024a) propose ZVQAF. ZVQAF finetunes a caption model for external knowledge. During the training stage, it uses a frozen image-text matching module to select relevant image regions, and generates a caption based on the regions. While the caption and the original question are sent to the frozen LLM to generate the answer, the captioning model is updated via reinforcement learning strategy. Wu et al. (2023c) utilize multiple external knowledge to generate answer. The proposed method identifies the "relation" and "entity" part from the text and embeds them into separate spaces, and aggregates the feature in different spaces with a weighted score for answer prediction. Other than focusing on image-based VQA tasks, some knowledge-based methods help improve the model's performance on the video question

understanding, instead of the image. Wang et al. (2022) propose a few-shot learning method for video question answering. Given the video input, it extracts the visual semantics from the token level, the frame level, and the video level by converting them into the textual representation, and the textual description is fed into the language model with task-specific instruction to generate answer. Pan et al. (2023), on the other hand, adopt the CLIP model to retrieve texts from the text corpus that are relevant to the video frames, then input both the retrieved text and the question, along with video visual features, into the LLM.

Moreover, the knowledge-based methods are not constrained to the visual question answering tasks. Dunlap et al. (2023) propose a knowledge-based method for unseen image domain adaptation, namely Latent Augmentation using Domain Descriptions (LADS). Given an image from the source domain and a text description from both the source and the target domain, the method introduces an augmentation network to transform image embeddings from the source to the target domain using a domain alignment loss and a class consistency loss between the texts from two domains to ensure that the shifted image feature is in a new domain while it remains in the same class. Lin et al. (2023b) develop a knowledge-based method for image classification. While image classification is a uni-modal task, the method includes additional modalities as inputs, e.g., text and audio. These additional modalities are seen as the additional few-shot examples. By mapping inputs from different modalities into the same representation space, the knowledge from the additional modalities can serve as the supplementary information for the image modality and help the classification task. Cai et al. (2025) propose a training-time negation data generation strategy for CLIP that dynamically constructs negated captions without relying on large LLM-generated datasets. By exposing models to diverse, non-stationary negation examples during contrastive training, it improves negation awareness several tasks. Yang et al. (2022c) propose Z-LAVI that adopts external visual knowledge for uni-modal text tasks. Given a text input, either the corpus or the labels, Z-LAVI retrieves an image by recalling existing images from a search engine and generates synthesized images using visual generative models. The result of the LLM and the VLM are assembled to predict the final answer. Lastly, (Khattak et al., 2024) propose a knowledge-based prompt learning method, *protext*, for recognition task. Instead of directly freezing the model and train the text prompt, *protext* first collects the detailed descriptions for training classes via LLM, and then use contrastive learning to align the image caption with corresponding descriptions to enrich the text encoder with diverse contextual knowledge. All the methods above are summarized into Table 4.

5 Application Areas of Transfer Learning for Vision & Language Models

While efficient transfer learning methods significantly reduce the computational and data requirements for adapting large VLMs, their true impact is best demonstrated through real-world applications. In this section, we explore how these transfer learning methods enable practical and scalable implementations of VLMs, showcasing their role in solving real-world vision-language challenges, in medical, robotic, and human understanding domain, while maintaining efficiency and adaptability.

5.1 Medical & Health Care

Pre-trained VLMs, such as Flamingo, LLaVA, CLIP, etc., have been adopted in medical field by finetuned on medical data(Chambon et al., 2022; Chen et al., 2022c; Moor et al., 2023; Li et al., 2024a). However, since medical data is often scarce, compared with the general domain data, more advanced transfer learning plays a significant role in building effective medical VLMs.

A group of methods adopt prompt-based techniques for classification of medical image (Chang et al., 2024; Wang et al., 2024; Ye et al., 2024; Denner et al., 2024). Those methods explore the zero-shot and the few-shot learning using VLMs on radiology and X-ray images. Lu et al. (2023), on the other hand, introduce a multiple-instance learning technique to solve zero-shot image classification on gigapixel images, enhancing the capacity of CLIP capacity on large-scale medical image. Shakeri et al. (2024) introduce a few-shot medical image classification benchmark that can be used for validation and model improvement.

VLMs have also been used on medical image segmentation tasks. Poudel et al. (2024) finetune a general-purpose VLM with segmentation capability and evaluate its performance on 11 diverse medical datasets. Wang et al. (2023c) adopt a prompt-based technique for medical image segmentation by introducing Fourier

visual prompts to enhance the model's performance on X-ray images. Jiang et al. (2024) explore zero-shot lesion segmentation on 3D medical images with cross-modal interaction using VLMs.

Another group of methods explores text generation in the medical domain (Wu et al., 2023d; Li et al., 2024b). Wu et al. (2023d) adopt adapters and medical knowledge enhancement loss to improve medical report generation based on medical images. Li et al. (2024b) improve the quality of multimodal medical dialogue by extracting the relevant region of an input image in a zero-shot manner. Yu et al. (2025a); Gu et al. (2024), on the other hand, propose propose the prompt-based method to adopt VLMs for medical-VQA tasks for organ-level disease recognition and radiology analysis.

5.2 Robotic System

VLMs have been used to enable robots to better understand and interact with their environments through multimodal reasoning. Kuzmenko & Shvai (2024) examine variant VLMs on zero-shot navigation tasks and study their efficiency for planning and navigation. Chen et al. (2024a) propose a method to adopt commonsense knowledge of VLMs for a legged robot to deal with difficult and ambiguous real-world navigation situations. Narasimhan et al. (2024) introduce a lifelong learning method for VLMs to enhance navigation performance of service robots in human-centered environments. Unlu et al. (2024) utilize GLIP and instructBLIP for zero-shot object goal creation, by letting GLIP to target the object and then use instructBLIP to confirm it. Venkatesh & Min (2024), on the other hand, introduce a VLM for multi-robot pattern formation by using the knowledge from the pre-trained model to translate natural language instructions into the actionable robot configurations.

In addition to robotic navigation, Liu et al. (2024b) apply mark-based visual prompts on VLMs to achieve open-world robotic manipulation via free-form natural language information. Khorramshahi et al. (2021), on the other hand, adopt CLIP's zero shot capability to build a vehicle re-identification retrieval system to identify the make, the model, the color, and the production year of a given vehicle.

5.3 Human Understanding

Human understanding tasks include a variety of machine learning challenges focused on interpreting and analyzing human-related attributes, behaviors, and interactions from visual and multimodal data. With the adoption of VLMs and the corresponding transfer learning strategies, human understanding tasks can be tackled in a more efficient and effective way.

Wang et al. (2023a) finetune the CLIP model with a video contrastive loss, and use a temporal transformer to refine the support videos with the textual information guide for few shot learning in action recognition. Shi et al. (2024) also adopt the CLIP model, and collect a corpus of an action-related text database from an external source to guide the few-shot recognition with commonsense knowledge. Xing et al. (2024b) enhance the CLIP vision encoder with a global temporal adapter and local multimodal adapter, and freeze the rest of the model during finetuning for efficient few-shot action recognition adaptation. Lu et al. (2024), on the other hand, focus on the skeleton action recognition. They use the VLM to help the pre-training of the skeleton encoder, and train a cosine classifier as the support set. Wang et al. (2023b) adopt a text caption model, e.g., BLIP, to align the support and query action recognition videos in a cross-modal approach. Meanwhile, Lin et al. (2023a) and Phan et al. (2024) propose methods for zero-shot action recognition based on VLMs for improving the model's generalizablity on unseen domain.

The efficient VLM transfer learning has also been applied for the facial expression recognition tasks. Foteinopoulou & Patras (2024) and Zhao et al. (2024) introduce the variants of CLIP by training CLIP with the sample-level text descriptions and videos, and inference using the class name, (e.g., happiness and sadness), for the zero-shot facial expression recognition task. Zhao & Patras (2024) also utilize CLIP by adding the learnable context tokens to the text encoder and attach temporal module after vision encoder. Zhang et al. (2024d) propose Set-of-Vision prompting which attaches spatial information such as bounding boxes and facial landmarks to VLM's visual input to enable zero-shot emotion recognition. Chen et al. (2025b), on the other hand, design a mixture-of-expert (MoE) adapter architecture on top of the VLM. It

Category	Dataset	Training	Testing	Domain
Image Classification	MNIST(Lecun et al., 1998) [link]	60,000	10,000	handwriting number
	ImageNet-1K (Deng et al., 2009) [link]	1,281,167	50,000	general
	CIFAR-10 (Krizhevsky et al., 2009) [link]	50,000	10,000	animal, transportation
	CIFAR-100 (Krizhevsky et al., 2009) [link]	50,000	10,000	real-world objects
	SUN397 (Xiao et al., 2014) [link]	19,850	19,850	general
	Stanford-Cars (Krause et al., 2013) [link]	8,144	8,041	vehicles
	Oxford 102 Flowers (Nilsback & Zisserman, 2008)[link]	2,040	6,149	flowers
	Food-101 (Bossard et al., 2014) [link]	75,750	25,250	food
	Stanford Dogs Khosla et al. (2011) [link]	12,000	8,580	dogs
	Fashion-MNIST (Xiao et al., 2017) [link]	60,000	10,000	fashion
	SVHN (Xiao et al., 2023) [link]	73,275	26,032	house numbers
	Caltech-101 (Li et al., 2022a) [link]	3,060	6,085	general
	FGVC-Aircraft (Maji et al., 2013) [link]	6,667	3,333	aircraft
	Birdsnap (Berg et al., 2014) [link]	42,283	2,149	birds
	Places (Zhou et al., 2018a) [link]	1,800,000	328,500	city scene
	CelebA (Liu et al., 2015) [link]	162,770	19,962	face attributes
Image Captioning	Flickr8K (Plummer et al., 2015) [link]	7,000	1,000	general
	Flickr30k (Plummer et al., 2015) [link]	31,728	-	general
	COCO Caption (Chen et al., 2015) [link]	82,783	5,000	general
	Visual Genome (Krishna et al., 2016) [link]	108,000	-	general
	AI2D (Kembhavi et al., 2016) [link]	4,000	1,000	scientific diagram
	CUB-200 (Reed et al., 2016) [link]	5,994	5,794	birds
	Fashion Cap. (Yang et al., 2022b) [link]	993,000	-	fashion
	CC3M (Sharma et al., 2018) [link]	3,300,000	-	general
	CC12M (Changpinyo et al., 2021) [link]	12,400,000	-	general
	TextCaps. (Sidorov et al., 2020) [link]	21,953	3,289	general
Image-Text Retrieval	Flickr30K (Plummer et al., 2015) [link]	31,728	-	general
	Visual Genome (Krishna et al., 2016) [link]	108,000	-	general
	MSCOCO (Lin et al., 2014)	82,783	5,000	general
	WIT (Srinivasan et al., 2021) [link]	37,600,000	-	Wikipedia
	Open Image (Kuznetsova et al., 2020) [link]	8,850,000	125,000	general
Visual Question Answering	VQAv2 (Goyal et al., 2017) [link]	83,000	81,000	general
	CLiMB (Srinivasan et al., 2022) [link]	83,000	81000	general
	GQA (Hudson & Manning, 2019) [link]	113,000	-	scene graph
	OKVQA (Marino et al., 2019) [link]	14,000	-	external knowledge
	COCOQA (Ren et al., 2015) [link]	123,000	78,000	general
	TextQA (Singh et al., 2019) [link]	18,408	-	text material
	TDIUC (Kafle & Kanan, 2017) [link]	167,000	-	general
	KVQA (Shah et al., 2019) [link]	183,000	-	knowledge graph
Visual Entailment	SNLI-VE (Xie et al., 2019) [link]	529,000	17,000	general
	NLVR2 (Suhr et al., 2019) [link]	86,000	10,200	general
	Flickr30k Entities (Plummer et al., 2016) [link]	244,000	-	general
	Hatefull Memes Dataset (Kiela et al., 2021) [link]	10,000	-	hateful content detection
Semantic Segmentation	Cityscapes (Cordts et al., 2016) [link]	3,475	1,525	urban scene
	PASCAL VOC (Everingham et al., 2010) [link]	11,540	-	general
	ADE20K (Zhou et al., 2018b) [link]	20,210	3,000	general
	COCO-stuff (Caesar et al., 2018) [link]	164,000	-	general
	Mapillary Vistas (Neuhold et al., 2017) [link]	25,000	-	street scene
	CamVid (Brostow et al., 2008) [link]	700	-	road scene
	BDD100K (Yu et al., 2020) [link]	100,000	-	driving

Table 5: Summary of commonly used datasets

creates one adapter for each modality of the input to captures and fuses emotional movements from different data.

As the efficient VLM transfer learning methods continue to demonstrate strong performance across real-world application domains, it becomes increasingly clear that progress in this area depends not only on architectural innovation but also on the quality of the data used to train and evaluate these models. High-quality, diverse, and well-curated datasets are essential for capturing the complexity of vision-language interactions and for ensuring that improvements translate beyond controlled settings. In the following section, we introduce the key datasets and benchmarks for the VLM researches and provide the foundation for measuring progress in vision-language learning.

6 Vision Language Datasets & Benchmarks

Vision-language datasets play a pivotal role as the foundational backbone for training and evaluating VLMs. These datasets are designed to cover a wide range of challenges, from basic detection and classification tasks to visual–linguistic reasoning tasks that mimic complex human cognitive abilities. Beyond enabling model

training, dataset diversity and coverage critically influence model generalization, as insufficient or biased data distributions can lead to phenomena such as mode collapse Hu et al. (2025b), where models overfit to dominant patterns while failing to represent rare or diverse multimodal concepts. Vision-language datasets therefore not only enable training VLMs but also shape how effectively they understand and interact with multimodal inputs. In this section, we categorize vision-language datasets into six major task types, with a summary of all datasets presented in Table 5.

Image Classification task aims to assign predefined labels to images based on their visual content. Data in image classification dataset are organized as image-label pairs:

$$\mathcal{D}_i = \{I_i, Y_i\}, \tag{7}$$

Where $\mathcal{D}_i$ denotes the arbitrary sample of an image classification dataset, and I_i, Y_i represent the image and label in the sample. Image classification datasets are different in terms of size, number of classes, domains, etc. Image classification tasks are mainly evaluated via accuracy metric.

Image Captioning task generates descriptive textual sentences for images by understanding their visual content and context. Data pairs are presented in the format of:

$$\mathcal{D}_i = \{I_i, C_i\}, \tag{8}$$

where C_i is the image caption. Different from image classification task that can be simply measured by accuracy, image captioning are evaluated through more complext benchmarks such as BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016), etc.

Image-Text Retrieval datasets are deployed to align visual and textual modalities to enable the retrieval of the relevant image given the textual query (image retrieval) or the relevant text given an image query (text retrieval). For image-text retrieval datasets, data is organized similar to image captioning:

$$\mathcal{D}_i = \{I_i, C_i\}. \tag{9}$$

However, a given model needs to learn a reliable matching pattern between the two modalities. The evaluation metric for image-text retrieval is primarily *Recall@K*, which measures the proportion of relevant items in the top-K results of the ranking given the query.

Visual Question Answering task combines computer vision and natural language processing to answer questions about the content of an image. It requires models to understand both the image and the question, and to reason about objects, relationships, and contextual details. The data for VQA is in a triplet form:

$$\mathcal{D}_i = \{I_i, Q_i, A_i\}, \tag{10}$$

where Q_i represents the question and A_i represents its answer. Visual question answering tasks can also be evaluated using accuracy, BLEU and F1 score.

Visual Entailment task is the multimodal challenge that involves determining whether a given text, denoted as hypothesis, logically entails, contradicts, or is neutral to an image. In this task, a model is presented with an image and a text statement, and it must tell the relationship in one of the three categories. Similar to the visual question answering data sample, the visual entailment data is in a triplet form:

$$\mathcal{D}_i = \{I_i, H_i, Y_i\}, \tag{11}$$

where H_i means the hypothesis. The model takes the image and the hypothesis as inputs and predicts the label within entailment, neutral and contradiction. The most common metric for visual entailment task is accuracy.

Semantic Segmentation is a computer vision task that involves classifying each pixel in an image into a predefined set of categories, and effectively partitioning the image into segments that correspond to different objects or regions. Semantic segmentation datasets are organized in the form of:

$$\mathcal{D}_i = \{I_i, M_i\}, \tag{12}$$

where M_i denotes the semantic mask that associates each pixel value with a class label. The common metric for semantic segmentation task are pixel accuracy and intersection over union (IoU).

7 Conclusion

In this paper, we explored the landscape of efficient transfer learning methods for vision-language tasks, highlighting the state-of-the-art approaches that leverage adapters, prompts, innovative model and external knowledge. These methods have demonstrated significant advancements in enabling pre-trained models to adapt effectively to a wide range of vision-language applications, from image captioning to visual question answering, with improved the VLM efficiency in terms of computational resources and data usage. Transfer learning methods are essential for advancing VLMs, enabling bridging the gap between visual and textual understanding in a cost-effective manner. As the field evolves, continued innovation in transfer learning techniques will be instrumental in unlocking new possibilities for VLMs across diverse domains.

References

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.

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.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation, 2016. URL `https://arxiv.org/abs/1607.08822`.

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 (ICCV)*, December 2015.

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022.

Thomas Berg, Jiongxin Liu, Seung Woo Lee, Michelle L Alexander, David W Jacobs, and Peter N Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2011–2018, 2014.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Gabriel J. Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV (1)*, pp. 44–57, 2008.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

Yuliang Cai and Mohammad Rostami. Dynamic transformer architecture for continual learning of multimodal tasks. *arXiv preprint arXiv:2401.15275*, 2024a.

Yuliang Cai and Mohammad Rostami. Clumo: Cluster-based modality fusion prompt for continual learning in visual question answering. *arXiv preprint arXiv:2408.11742*, 2024b.

Yuliang Cai, Jesse Thomason, and Mohammad Rostami. Task-attentive transformer architecture for continual learning of vision-and-language tasks using knowledge distillation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yuliang Cai, Jesse Thomason, and Mohammad Rostami. Tng-clip: Training-time negation data generation for negation awareness of clip. *arXiv preprint arXiv:2505.18434*, 2025.

Rui Cao and Jing Jiang. Knowledge generation for zero-shot knowledge-based vqa. *arXiv preprint arXiv:2402.02541*, 2024.

Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022.

Tianyou Chang, Shizhan Chen, Guodong Fan, and Zhiyong Feng. A vision-language model based on prompt learner for few-shot medical images diagnosis. In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 1455–1460, 2024. doi: 10.1109/CSCWD61410.2024.10580842.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.

Annie S. Chen, Alec M. Lessing, Andy Tang, Govind Chada, Laura Smith, Sergey Levine, and Chelsea Finn. Commonsense reasoning for legged robot adaptation with vision-language models, 2024a. URL `https://arxiv.org/abs/2407.02666`.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. 2022a.

Haoran Chen, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Multi-prompt alignment for multi-source unsupervised domain adaptation. *Advances in Neural Information Processing Systems*, 36:74127–74139, 2023.

Haoxing Chen, Zizheng Huang, Yan Hong, Yanshuo Wang, Zhongcai Lyu, Zhuoer Xu, Jun Lan, and Zhangxuan Gu. Efficient transfer learning for video-language foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29129–29138, June 2025a.

Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang, and Yanbin Hao. Mixture of multimodal adapters for sentiment analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1822–1833, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.90. URL `https://aclanthology.org/2025.naacl-long.90/`.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. URL `https://arxiv.org/abs/1504.00325`.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020. URL `https://arxiv.org/abs/1909.11740`.

Yutong Chen, Fangyun Wei, Xiao Sun, Zhirong Wu, and Stephen Lin. A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5120–5130, 2022b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5152–5161, 2022c.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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, 2016. URL https://arxiv.org/abs/1604.01685.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Stefan Denner, Markus Bujotzek, Dimitrios Bounias, David Zimmerer, Raphael Stock, Paul F. Jäger, and Klaus Maier-Hein. Visual prompt engineering for medical vision language models in radiology, 2024. URL https://arxiv.org/abs/2408.15802.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

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

Lisa Dunlap, Clara Mohri, Devin Guillory, Han Zhang, Trevor Darrell, Joseph E Gonzalez, Aditi Raghunathan, and Anna Rohrbach. Using language to extend to unseen domains. International Conference on Learning Representations (ICLR), 2023.

Deniz Engin and Yannis Avrithis. Zero-shot and few-shot video question answering with multi-modal prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2804–2810, 2023.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. doi: 10.1007/s11263-009-0275-4.

Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18623–18633, 2023.

Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

Niki Maria Foteinopoulou and Ioannis Patras. Emoclip: A vision-language method for zero-shot video facial expression recognition, 2024. URL https://arxiv.org/abs/2310.16640.

Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41 (4):1–13, 2022.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

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.

Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Tiancheng Gu, Kaicheng Yang, Dongnan Liu, and Weidong Cai. Lapa: Latent prompt assist model for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4971–4980, June 2024.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10867–10877, 2023.

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. URL `https://arxiv.org/abs/1902.00751`.

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.

Naihao Hu, Xiaodan Zhang, Qiyuan Zhang, Wei Huo, and Shaojie You. Zpvqa: Visual question answering of images based on zero-shot prompt learning. *IEEE Access*, 13:50849–50859, 2025a. doi: 10.1109/ACCESS.2025.3550942.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2963–2975, 2023.

Zizhao Hu, Mohammad Rostami, and Jesse Thomason. Multimodal synthetic data finetuning and model collapse: Insights from vlms and diffusion models. In *Proceedings of the 27th International Conference on Multimodal Interaction*, pp. 588–599, 2025b.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023.

Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

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*, pp. 6700–6709, 2019.

Saurav Jha, Dong Gong, and Lina Yao. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *arXiv preprint arXiv:2403.19137*, 2024.

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*, pp. 4904–4916. PMLR, 2021.

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*, pp. 709–727. Springer, 2022.

Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Shiji Song, and Gao Huang. Cross-modal adapter for vision–language retrieval. *Pattern Recognition*, 159:111144, 2025. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2024.111144. URL https://www.sciencedirect.com/science/article/pii/S0031320324008951.

Yankai Jiang, Wenhui Lei, Xiaofan Zhang, and Shaoting Zhang. Unleashing the potential of vision-language pre-training for 3d zero-shot lesion segmentation via mask-attribute alignment, 2024. URL https://arxiv.org/abs/2410.15744.

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*, 2021a.

Xisen Jin, Bill Yuchen Lin, Mohammad Rostami, and Xiang Ren. Learn continually, generalize rapidly: Lifelong knowledge accumulation for few-shot learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 714–729, 2021b.

Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms, 2017. URL https://arxiv.org/abs/1703.09684.

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*, pp. 14162–14171, 2024.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. URL https://arxiv.org/abs/1603.07396.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2, 2019.

Aditya Khamparia, Babita Pandey, Shrasti Tiwari, Deepak Gupta, Ashish Khanna, and Joel J. P. C. Rodrigues. An integrated hybrid cnn–rnn model for visual description and generation of captions. *Circuits Syst. Signal Process.*, 39(2):776–788, February 2020. ISSN 0278-081X. doi: 10.1007/s00034-019-01306-8. URL https://doi.org/10.1007/s00034-019-01306-8.

Zaid Khan, Vijay Kumar BG, Samuel Schulter, Manmohan Chandraker, and Yun Fu. Exploring question decomposition for zero-shot vqa. *Advances in Neural Information Processing Systems*, 36, 2024.

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*, pp. 19113–19122, 2023.

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, 2024. URL https://arxiv.org/abs/2401.02418.

Pirazh Khorramshahi, Sai Saketh Rambhatla, and Rama Chellappa. Towards accurate visual and natural language-based vehicle retrieval systems. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4178–4187, 2021. doi: 10.1109/CVPRW53098.2021.00472.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes, 2021. URL https://arxiv.org/abs/2005.04790.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. URL https://arxiv.org/abs/1602.07332.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report TR-2009, 2009. Accessed: [Insert the date you accessed this paper].

Dmytro Kuzmenko and Nadiya Shvai. Balancing performance and efficiency in zero-shot robotic navigation, 2024. URL https://arxiv.org/abs/2406.03015.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16155–16165, 2023.

Yunshi Lan, Xiang Li, Xin Liu, Yang Li, Wei Qin, and Weining Qian. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4389–4400, 2023.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

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, 2024a.

Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, Apr 2022a.

Jingzheng Li and Hailong Sun. Lift: Transfer learning in vision-language models for downstream adaptation and generalization. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4678–4687, 2023.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

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*, pp. 12888–12900. PMLR, 2022b.

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*, pp. 19730–19742. PMLR, 2023.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020. URL https://arxiv.org/abs/2004.06165.

Zhangpu Li, Changhong Zou, Suxue Ma, Zhicheng Yang, Chen Du, Youbao Tang, Zhenjie Cao, Ning Zhang, Jui-Hsin Lai, Ruei-Sung Lin, Yuan Ni, Xingzhi Sun, Jing Xiao, Jieke Hou, Kai Zhang, and Mei Han. Zalm3: Zero-shot enhancement of vision-language alignment via in-context information in multi-turn multimodal medical dialogue, 2024b. URL https://arxiv.org/abs/2409.17610.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, CAICE '24, pp. 405–409, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716942. doi: 10.1145/3672758.3672824. URL https://doi.org/10.1145/3672758.3672824.

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*, pp. 740–755. Springer, 2014.

Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2851–2862, 2023a.

Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19325–19337, 2023b.

Cheng Liu, Chao Wang, Yan Peng, and Zhixu Li. Zvqaf: Zero-shot visual question answering with feedback from large language models. *Neurocomputing*, 580:127505, 2024a.

Fangchen Liu, Kuan Fang, Pieter Abbeel, and Sergey Levine. Moka: Open-world robotic manipulation through mark-based visual prompting, 2024b. URL https://arxiv.org/abs/2403.03174.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Jinxing Liu, Junjin Xiao, Haokai Ma, Xiangxian Li, Zhuang Qi, Xiangxu Meng, and Lei Meng. Prompt learning with cross-modal feature alignment for visual domain adaptation. In *CAAI International Conference on Artificial Intelligence*, pp. 416–428. Springer, 2022.

Jun Liu, Ziqian Lu, Hao Luo, Zheming Lu, and Yangming Zheng. Progressive multi-prompt learning for vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(10):9562–9574, 2025. doi: 10.1109/TCSVT.2025.3557474.

Xiangyu Liu, Yanlei Shang, and Yong Chen. Trimpl: Masked multi-prompt learning with knowledge mixing for vision-language few-shot learning. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 552–560, 2024c.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images, 2023. URL https://arxiv.org/abs/2306.07831.

Mingqi Lu, Siyuan Yang, Xiaobo Lu, and Jun Liu. Cross-modal contrastive pre-training for few-shot skeleton action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9798–9807, 2024. doi: 10.1109/TCSVT.2024.3402952.

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*, pp. 5206–5215, 2022.

Han Ma, Baoyu Fan, Benjamin K Ng, and Chan-Tong Lam. Vl-few: Vision language alignment for multimodal few-shot meta learning. *Applied Sciences*, 14(3):1169, 2024.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. *arXiv preprint arXiv:2210.07179*, 2022.

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*, pp. 3195–3204, 2019.

Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pp. 353–367. PMLR, 2023.

Siddarth Narasimhan, Aaron Hao Tan, Daniel Choi, and Goldie Nejat. Olivia-nav: An online lifelong vision language approach for mobile robot social navigation, 2024. URL https://arxiv.org/abs/2409.13675.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5000–5009, 2017. URL https://api.semanticscholar.org/CorpusID:5753855.

Bolin Ni, Hongbo Zhao, Chenghao Zhang, Ke Hu, Gaofeng Meng, Zhaoxiang Zhang, and Shiming Xiang. Enhancing visual continual learning with language-guided supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24068–24077, 2024.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*. ICVGIP, Dec 2008.

Junting Pan, Ziyi Lin, Yuying Ge, Xiatian Zhu, Renrui Zhang, Yi Wang, Yu Qiao, and Hongsheng Li. Retrieving-to-answer: Zero-shot video question answering with frozen large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 272–283, 2023.

Omiros Pantazis, Gabriel Brostow, Kate Jones, and Oisin Mac Aodha. Svl-adapter: Self-supervised adapter for vision-language pretrained models. *arXiv preprint arXiv:2210.03794*, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Fang Peng, Xiaoshan Yang, Linhui Xiao, Yaowei Wang, and Changsheng Xu. Sgva-clip: Semantic-guided visual adapting of vision-language models for few-shot image classification. *IEEE Transactions on Multimedia*, 2023.

Thinh Phan, Khoa Vo, Duy Le, Gianfranco Doretto, Donald Adjeroh, and Ngan Le. Zeetad: Adapting pretrained vision-language model for zero-shot end-to-end temporal action detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7046–7055, 2024.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016. URL https://arxiv.org/abs/1505.04870.

Kanchan Poudel, Manish Dhakal, Prasiddha Bhandari, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Exploring transfer learning in medical image segmentation using vision-language models, 2024. URL https://arxiv.org/abs/2308.07706.

Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, and Wenwu Zhu. Decouple before interact: Multi-modal prompt learning for continual visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2953–2962, 2023.

Longtian Qiu, Renrui Zhang, Ziyu Guo, Ziyao Zeng, Zilu Guo, Yafeng Li, and Guangnan Zhang. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021.

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*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

Scott Reed, Zeynep Akata, Bernt Schiele, and Honglak Lee. Learning deep representations of fine-grained visual descriptions, 2016. URL https://arxiv.org/abs/1605.05395.

Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering, 2015. URL https://arxiv.org/abs/1505.02074.

Jintao Rong, Hao Chen, Tianxiao Chen, Linlin Ou, Xinyi Yu, and Yifan Liu. Retrieval-enhanced visual prompt learning for few-shot classification. *arXiv preprint arXiv:2306.02243*, 2023.

Mohammad Rostami, Soheil Kolouri, and Praveen K Pilly. Complementary learning for overcoming catastrophic forgetting using experience replay. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 3339–3345, 2019.

Mohammad Rostami, Soheil Kolouri, Zak Murez, Yuri Owechko, Eric Eaton, and Kuyngnam Kim. Zero-shot image classification using coupled dictionary embedding. *Machine Learning with Applications*, 8:100278, 2022.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: knowledge-aware visual question answering. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33018876. URL https://doi.org/10.1609/aaai.v33i01.33018876.

Fereshteh Shakeri, Yunshi Huang, Julio Silva-Rodríguez, Houda Bahig, An Tang, Jose Dolz, and Ismail Ben Ayed. Few-shot adaptation of medical vision-language models, 2024. URL https://arxiv.org/abs/2409.03868.

Zeyu Shangguan, Daniel Seita, and Mohammad Rostami. Cross-domain multi-modal few-shot object detection via rich text. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6570–6580. IEEE, 2025a.

Zeyu Shangguan, Daniel Seita, and Mohammad Rostami. Cross-domain few-shot object detection with multi-modal textual enrichment. *arXiv preprint arXiv:2502.16469*, 2025b.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E Gonzalez, Kurt Keutzer, and Trevor Darrell. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5656–5667, 2024.

Yuheng Shi, Xinxiao Wu, Hanxi Lin, and Jiebo Luo. Commonsense knowledge prompting for few-shot action recognition in videos. *IEEE Transactions on Multimedia*, 26:8395–8405, 2024. doi: 10.1109/TMM.2024.3361157.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension, 2020. URL https://arxiv.org/abs/2003.12462.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. URL https://arxiv.org/abs/1904.08920.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems*, 26, 2013.

Lin Song, Ruoyi Xue, Hang Wang, Hongbin Sun, Yixiao Ge, Ying Shan, et al. Meta-adapter: An online few-shot learner for vision-language model. *Advances in Neural Information Processing Systems*, 36: 55361–55374, 2023.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2443–2449, 2021.

Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 35:29440–29453, 2022.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs, 2019. URL https://arxiv.org/abs/1811.00491.

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*, pp. 5227–5237, 2022.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pp. 270–279. Springer, 2018.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv preprint arXiv:2210.08773*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Halil Utku Unlu, Shuaihang Yuan, Congcong Wen, Hao Huang, Anthony Tzes, and Yi Fang. Reliable semantic understanding for real world zero-shot object goal navigation, 2024. URL https://arxiv.org/abs/2410.21926.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation, 2015. URL https://arxiv.org/abs/1411.5726.

Vishnunandan L. N. Venkatesh and Byung-Cheol Min. Zerocap: Zero-shot multi-robot context aware pattern formation via large language models, 2024. URL https://arxiv.org/abs/2404.02318.

Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification, 2016. URL https://arxiv.org/abs/1604.04573.

Sicheng Wang, Che Liu, and Rossella Arcucci. How does diverse interpretability of textual prompts impact medical vision-language zero-shot tasks?, 2024. URL https://arxiv.org/abs/2409.00543.

Xiang Wang, Shiwei Zhang, Jun Cen, Changxin Gao, Yingya Zhang, Deli Zhao, and Nong Sang. Clip-guided prototype modulating for few-shot action recognition, 2023a. URL https://arxiv.org/abs/2303.02982.

Xiang Wang, Shiwei Zhang, Hangjie Yuan, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Few-shot action recognition with captioning foundation models, 2023b. URL https://arxiv.org/abs/2310.10125.

Yan Wang, Jian Cheng, Yixin Chen, Shuai Shao, Lanyun Zhu, Zhenzhou Wu, Tao Liu, and Haogang Zhu. Fvp: Fourier visual prompting for source-free unsupervised domain adaptation of medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(12):3738–3751, 2023c. doi: 10.1109/TMI.2023.3306105.

Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, and Tieniu Tan. Improving zero-shot generalization for clip with synthesized prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3032–3042, 2023d.

Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *Advances in Neural Information Processing Systems*, 35:8483–8497, 2022.

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.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

Jingyuan Wen, Yutian Luo, Nanyi Fei, Guoxing Yang, Zhiwu Lu, Hao Jiang, Jie Jiang, and Zhao Cao. Visual prompt tuning for few-shot text classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 5560–5570, 2022.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256, 2023a. doi: 10.1109/BigData59044.2023.10386743.

Qiong Wu, Shubin Huang, Yiyi Zhou, Pingyang Dai, Annan Shu, Guannan Jiang, and Rongrong Ji. Approximated prompt tuning for vision-language pre-trained models. *arXiv preprint arXiv:2306.15706*, 2023b.

Qiong Wu, Wei Yu, Yiyi Zhou, Shubin Huang, Xiaoshuai Sun, and Rongrong Ji. Parameter and computation efficient transfer learning for vision-language pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.

Sen Wu, Guoshuai Zhao, and Xueming Qian. Resolving zero-shot and fact-based visual question answering via enhanced fact retrieval. *IEEE Transactions on Multimedia*, 26:1790–1800, 2023c.

Shibin Wu, Bang Yang, Zhiyu Ye, Haoqian Wang, Hairong Zheng, and Tong Zhang. Improving medical report generation with adapter tuning and knowledge enhancement in vision-language foundation models, 2023d. URL `https://arxiv.org/abs/2312.03970`.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017. URL `https://arxiv.org/abs/1708.07747`.

Jianxiong Xiao, Krista A. Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119:3–22, 2014. URL `https://api.semanticscholar.org/CorpusID:10224573`.

Tim Z. Xiao, Johannes Zenn, and Robert Bamler. The svhn dataset is deceptive for probabilistic generative models due to a distribution mismatch, 2023. URL `https://arxiv.org/abs/2312.02168`.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning, 2019. URL `https://arxiv.org/abs/1811.10582`.

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, 2024a.

Jiazheng Xing, Chao Xu, Mengmeng Wang, Guang Dai, Baigui Sun, Yong Liu, Jingdong Wang, and Jian Zhao. Ma-fsar: Multimodal adaptation of clip for few-shot action recognition, 2024b. URL `https://arxiv.org/abs/2308.01532`.

Yicheng Xu, Yuxin Chen, Jiahao Nie, Yusong Wang, Huiping Zhuang, and Manabu Okumura. Advancing cross-domain discriminability in continual learning of vison-language models. *arXiv preprint arXiv:2406.18868*, 2024.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *Advances in Neural Information Processing Systems*, 35:124–141, 2022a.

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*, pp. 23826–23837, 2024.

Xuewen Yang, Heming Zhang, Di Jin, Yingru Liu, Chi-Hao Wu, Jianchao Tan, Dongliang Xie, Jue Wang, and Xin Wang. Fashion captioning: Towards generating accurate descriptions with semantic rewards, 2022b. URL https://arxiv.org/abs/2008.02693.

Yue Yang, Wenlin Yao, Hongming Zhang, Xiaoyang Wang, Dong Yu, and Jianshu Chen. Z-lavi: Zero-shot language solver fueled by visual imagination. *arXiv preprint arXiv:2210.12261*, 2022c.

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.

Yaoqin Ye, Junjie Zhang, and Hongwei Shi. Pseudo-prompt generating in pre-trained vision-language models for multi-label medical image classification, 2024. URL https://arxiv.org/abs/2405.06468.

Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4187–4195, 2017. doi: 10.1109/CVPR.2017.446.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning, 2020. URL https://arxiv.org/abs/1805.04687.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024.

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*, pp. 10899–10909, 2023.

Ting Yu, Zixuan Tong, Jun Yu, and Ke Zhang. Fine-grained adaptive visual prompt for generative medical visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):9662–9670, Apr. 2025a. doi: 10.1609/aaai.v39i9.33047. URL https://ojs.aaai.org/index.php/AAAI/article/view/33047.

Zhou Yu, Xuecheng Ouyang, Zhenwei Shao, Meng Wang, and Jun Yu. Prophet: Prompting large language models with complementary answer heuristics for knowledge-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6797–6808, 2025b. doi: 10.1109/TPAMI.2025.3562422.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021. URL https://arxiv.org/abs/2111.11432.

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*, pp. 1593–1603, 2024.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a.

Jing Zhang, Xiaoqiang Liu, Mingzhe Chen, and Zhe Wang. Cross-modal feature distribution calibration for few-shot visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7151–7159, 2024a.

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*, 2024b.

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*, 46(8):5625–5644, 2024c.

Qixuan Zhang, Zhifeng Wang, Dylan Zhang, Wenjia Niu, Sabrina Caldwell, Tom Gedeon, Yang Liu, and Zhenyue Qin. Visual prompting in llms for enhancing emotion recognition, 2024d. URL `https://arxiv.org/abs/2410.02244`.

Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023b.

Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15211–15222, 2023c.

Yue Zhang, Hehe Fan, Wei Ji, Yongkang Wong, Roger Zimmermann, and Yi Yang. Prompt-aware adapter: Learning adaptive visual tokens for multimodal large language models. *IEEE Transactions on Artificial Intelligence*, pp. 1–10, 2025. doi: 10.1109/TAI.2025.3596925.

Zengqun Zhao and Ioannis Patras. Prompting visual-language models for dynamic facial expression recognition, 2024. URL `https://arxiv.org/abs/2308.13382`.

Zengqun Zhao, Yu Cao, Shaogang Gong, and Ioannis Patras. Enhancing zero-shot facial expression recognition by llm knowledge transfer, 2024. URL `https://arxiv.org/abs/2405.19100`.

Hao Zheng, Shunzhi Yang, Zhuoxin He, Jinfeng Yang, and Zhenhua Huang. Hierarchical cross-modal prompt learning for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1891–1901, October 2025.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464, 2018a. doi: 10.1109/TPAMI.2017.2723009.

Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset, 2018b. URL `https://arxiv.org/abs/1608.05442`.

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*, pp. 16816–16825, 2022a.

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, 2022b.

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, pp. 13041–13049, 2020.

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*, pp. 15659–15669, 2023.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.