How Vision-Language Tasks Benefit from Large Pre-trained Models: A Survey

Yayun Qi, Hongxi Li, Yiqi Song, Xinxiao Wu, *Member, IEEE,* Jiebo Luo, *Fellow, IEEE*

***Abstract*—The exploration of various vision-language tasks, such as visual captioning, visual question answering, and visual commonsense reasoning, is an important area in artificial intelligence and continuously attracts the attention of the research community. Despite improvements in overall performance, classic challenges still exist in vision-language tasks and hinder the development of this area. In recent years, the rise of pre-trained models is driving the research on vision-language tasks. Because of the massive scale of training data and model parameters, pre-trained models have exhibited excellent performance in numerous downstream tasks. Inspired by the powerful capabilities of pre-trained models, new paradigms have emerged to solve the classic challenges. Such methods have become mainstream in current research with increasing attention and rapid advances. In this paper, we present a comprehensive overview of how vision-language tasks benefit from pre-trained models. First, we review several main challenges in vision-language tasks and discuss the limitations of previous solutions before the era of pre-training. Next, we summarize the recent advances in incorporating pre-trained models to address the challenges in vision-language tasks. Finally, we analyze the potential risks associated with the inherent limitations of pre-trained models, discuss possible solutions, and attempt to provide future research directions.**

***Index Terms*—Vision-language tasks, Pre-trained model, Vision-language model, Large language model.**

I. Introduction

As an intersection of computer vision and natural language processing, vision-language tasks have attracted significant attention in recent years [1]. Research in this area aims to bridge the gap between visual and textual modalities, offering promising enhancements for downstream applications [2]. The scope of visual-language tasks is quite broad and encompasses classic tasks such as visual captioning [3], [4], visual question answering [5], [6], visual-text retrieval [7], [8], etc. In addition, some traditional visual recognition tasks [9]–[12] (*e.g.,* image classification) have begun to emphasize the semantic meaning behind language supervision rather than simply treating these signals as one-hot vectors. By incorporating semantic information as an external language modality, these tasks can also be viewed as visual-language tasks [13]. As research in the field of vision-language continues to advance, tasks now address a broader scope of visual data, expanding from images [5], [14]–[16] to videos [17]–[20]. Furthermore, the capabilities required by these tasks have shifted from basic perception [21], [22] to more advanced reasoning [23]–[27]. These advancements have resulted in significant improvements in performance across various aspects.

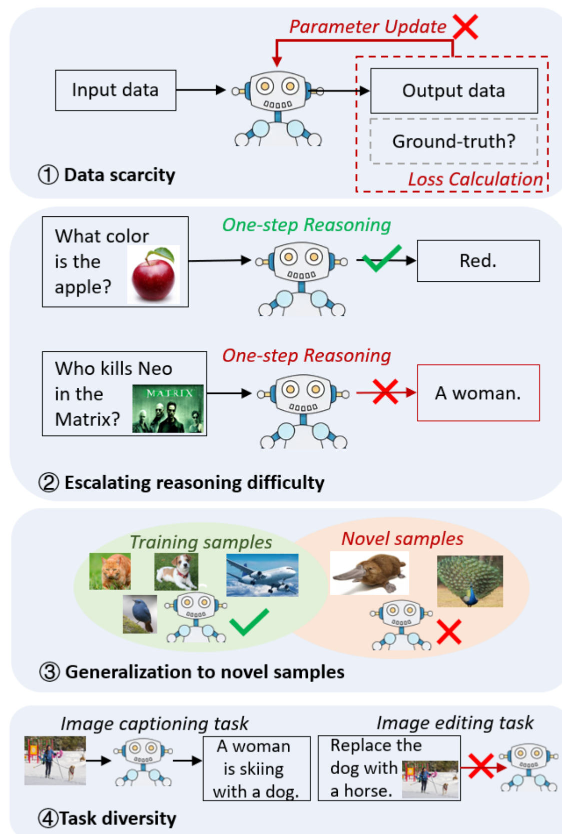

Fig. 1. An illustration of four classic challenges in vision-language tasks.

Despite the remarkable advances in many vision-language tasks, some classic challenges continue to hinder further improvement in this area and negatively impact practical applications. In Fig. 1, we present the four main challenges faced by vision-language tasks. From the perspective of training data, manually annotating data for vision-language tasks is labor-intensive, and it is challenging to automate this by masking data to construct self-supervised training data as in unimodal

Yayun Qi, Hongxi Li, and Yiqi Song are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China. Emails: {qiyayun,lihongxi,yiqis}@bit.edu.cn
Xinxiao Wu is with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China, and also with the Guangdong Provincial Laboratory of Machine Perception and Intelligent Computing, Shenzhen MSU-BIT University, Shenzhen 518172, China. Email: wuxinxiao@bit.edu.cn
Jiebo Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627, USA. Email: jiebo.luo@gmail.com
Xinxiao Wu is the corresponding author.

0000–0000/00$00.00 © 2024 IEEE

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3632653

tasks, which results in a significant scarcity of annotated training data. From the perspective of reasoning, the difficulties of both understanding the visual content and answering the questions have increased dramatically, requiring models to perform multi-hop reasoning rather than directly obtaining answers from visual information, which poses challenges to complex reasoning. From the perspective of generalization, when new samples that exceed the model's capabilities in real-world scenarios are encountered, the limited cross-modal knowledge in the training set is not sufficient to handle these samples effectively. From the perspective of task diversity, different tasks involve different reasoning processes and input-output workflows and require different models, one of which is tailored to a specific training process and a unique set of parameters, resulting in poor model universality. Researchers continuously propose methods to address or alleviate these challenges, and new solutions to these challenges have emerged due to the rapid development of large-scale pre-trained models.

With the rise of pre-trained Large Language Models (LLMs), the approach to performing downstream tasks has gradually shifted from neural network-based fully supervised learning to the "pre-train & fine-tune" paradigm and the "pre-train, prompt, predict" paradigm [28]. In this process, numerous LLMs have emerged (*e.g.,* LLaMA [29], Vicuna [30], and QWEN [31]), and the scale of the training data and model parameters continues to increase. Consequently, these models have exhibited intelligent emergence and performed well in many downstream tasks of NLP. Inspired by the success of LLMs, the vision-language field has also seen the emergence of discriminative pre-trained Vison-Language Models (VLMs) represented by CLIP [32] (*e.g.,* ALIGN [33], VLMo [34], and FLIP [35]), and generative VLMs represented by BLIP [36] (*e.g.,* MiniGPT-4 [37], LLaVA [38], and Video-LLaMA [39]). Compared with LLMs, VLMs incorporate additional comprehensive knowledge of vision-language correspondence.

Because of the outstanding few-shot learning capabilities, extensive knowledge storage, and robust generalization of large pre-trained models, an increasing number of methods for vision-language tasks are integrating pre-trained models to benefit from these models. The motivations and strategies for using pre-trained models vary significantly among these methods. However, the research community lacks a comprehensive survey that systematically summarizes these works. Such a survey is crucial for reviewing and categorizing existing methods, as it not only offers researchers a coherent overview but also provides valuable insights for future research.

In this paper, we present a comprehensive overview of how vision-language tasks benefit from large pre-trained models. To the best of our knowledge, this survey is pioneering in its focus on this topic and in categorizing methods according to the challenges that they tackle. We believe this survey can provide valuable guidance for researchers to deepen their understanding of this field and promote further developments in the community.

To ensure the systematic nature of this paper, we first review several main challenges encountered in vision-language tasks. In addition, we discuss the limitations of solutions before the era of pre-trained models. Subsequently, we summarize existing methods that use the capabilities of pre-trained models to address the challenges in vision-language tasks. These methods are categorized according to the specific challenges that they target and are further subdivided into detailed classifications according to their paradigms. Notably, we encapsulate these methodological paradigms into illustrations for a more intuitive understanding. The tasks involved include visual classification, visual captioning, visual question answering, image editing, video localization, visual data generation, open-vocabulary object detection, etc. Finally, we explore the potential risks associated with the use of pre-trained models in vision-language tasks due to the inherent limitations of pre-trained models, and discuss potential solutions to guide future research directions.

The remainder of the paper is organized as follows. Section III introduces the main challenges encountered in vision-language tasks. Section V reviews the solutions to these challenges using pre-trained models. Section VI discusses the potential risks associated with integrating pre-trained models and explores promising future directions to mitigate these risks.

II. Related Surveys and Differences

With the advancement of the research on pre-trained models, several surveys [40]–[48] have focused on summarizing the fundamental information of pre-trained models (*i.e.,* how they work), such as their training data, model structures, evaluation practices, and pre-training objectives and strategies. Among them, VLM-related surveys [43]–[48] aim to cover current VLMs and classify them with different taxonomies, in order to comprehensively review the unique features of each VLM and inspire future research on the design of novel VLMs.

Another trend in the surveys [49]–[51] on pre-trained models is to summarize the methods that combine pre-trained models to tackle specific tasks (*i.e.,* how to use). These surveys typically categorize methods based on their paradigms or target tasks to inspire future research to apply or improve existing paradigms on various tasks. However, surveys on how to use pre-trained models are scarce and focus mostly on limited pre-trained models and modalities. The research community lacks a survey that covers the integration of both LLMs and VLMs to address vision-language tasks involving images, videos, and text. To this end, we present a survey that focuses on the novel paradigms of incorporating pre-trained models to address classic challenges in vision-language tasks. Table I identifies the distinctions between our survey and other surveys in terms of content and coverage.

III. Challenges in Vision-Language Tasks

In this section, we introduce the four major challenges faced by all models in vision-language tasks, including data scarcity, escalating reasoning complexity, the generalization to novel samples, and task diversity. For each challenge, we discuss the corresponding methods and their limitations before the era of pre-training.

TABLE I
COMPARISON BETWEEN OUR SURVEY AND OTHER SURVEYS. WE DIVIDE THESE SURVEYS INTO TWO CLASSES ACCORDING TO THEIR MAIN CONTENT.

Main Content	Survey	Pre-trained Model		Modality			Taxonomy
		LLM	VLM	Image	Video	Text	
Summarize fundamental information of pre-trained models	Foundational Models Defining a New Era in Vision: A Survey and Outlook [44]		√	√	√	√	Categorize VLMs based on the architectures
	Video Understanding with Large Language Models: A Survey [48]	√	√		√	√	Categorize VLMs based on the strategies to integrate LLMs
	Exploring the Reasoning Abilities of Multimodal Large Language Models (MLLMs): A Comprehensive Survey on Emerging Trends in Multimodal Reasoning [47]		√	√	√	√	Categorize VLMs based on the strategies to improve reasoning abilities
	Large-scale Multi-Modal Pre-trained Models: A Comprehensive Survey [43]		√	√	√	√	Categorize VLMs based on the pre-training strategies
	MM-LLMs: Recent Advances in MultiModal Large Language Models [46]		√	√	√	√	Categorize VLMs from both functional and design perspectives
Summarize the integration of pre-trained models in specific tasks	Vision-Language Models for Vision Tasks: A Survey [49]		√	√	√		Categorize methods based on the paradigms of incorporating VLMs
	LLMs Meet Multimodal Generation and Editing: A Survey [50]	√		√	√	√	Categorize methods based on the target tasks
	Generalized Out-of-Distribution Detection and Beyond in Vision Language Model Era: A Survey [51]		√	√			Categorize methods based on the target tasks
	How Vision-Language Tasks Benefit from Large Pre-trained Models: A Survey (Ours)	√	√	√	√	√	Categorize methods based on the target challenges

A. Data Scarcity

For supervised learning methods of vision-language tasks, such as image captioning and visual question answering, annotated data play a crucial role in their training process since the annotated data constitute the foundation for updating parameters. However, manually annotating these large datasets is quite laborious, and collecting annotated data for certain tasks, such as counterfactual image editing and counterintuitive reasoning from real-world scenarios, is particularly challenging. On the other hand, the training data for vision-language tasks involve two different modalities, which makes it impractical to construct self-supervised data using unlabeled unimodal data via proxy tasks, as is commonly done in natural language understanding and generation [52]–[54]. These circumstances lead to the challenge of data scarcity in training, and the primary concern is how to accomplish vision-language tasks without available annotated data.

To address this challenge in vision-language tasks, researchers have developed semi-supervised learning methods [55], [56] that train models using a small set of labeled visual-textual data while simultaneously using a large set of unlabeled text and visual data to enhance generalization. Moreover, weakly supervised learning methods [57], [58] have been proposed to learn from incomplete visual-textual annotations, noisy labels, or weak supervision signals. For example, a visual question answering (VQA) model can be trained with only image-caption pairs as supervision. In addition, unsupervised learning methods [59], [60] have been designed to discover the patterns and relationships between visual and textual data from unpaired visual and text inputs via techniques such as contrastive and adversarial learning. Despite these advances, as the amount of annotated visual-textual data decreases, these methods often struggle to achieve performance comparable to that of fully supervised learning. In addition, in semi-supervised and weakly supervised settings, limited annotations often cause models to overfit the observed visual-textual patterns, and this overfitting is subsequently propagated when assigning pseudo labels. Furthermore, the quality of annotations plays a crucial role, as inaccurate or incomplete annotations may lead to the incorrect learning of visual-textual correlations.

B. Escalating Reasoning Complexity

In the early stages of development, vision-language tasks primarily require models to perceive and summarize visual information from given visual data. For example, tasks may involve querying either the location or attributes of certain entities and identifying the relationships between different entities. Solutions to these tasks can often be derived directly from the visual content without complex reasoning.

As research progresses, an increasing number of vision-language tasks require models to perform additional reasoning to complete. The complexity of reasoning continues to escalate since both the visual content and the associated questions have become more complex than before. Visual scenes have evolved from simple combinations of geometric shapes [129] to real scenes [130] with various objects and relationships, which requires relationship reasoning. In addition, visual data now include not only static images [131] but also activity videos [132], [133] and even longer movie clips [6], [134] that require temporal reasoning. Furthermore, counterintuitive images [26] and humorous videos [135] pose challenges to models in terms of compositional and commonsense reasoning. The related questions shift from recognition to cognition, which involves exploring commonsense information and causal relationships in visual content. This requires additional reasoning processes such as commonsense reasoning [23], abductive reasoning [25], [136], and counterintuitive reason-

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3632653

TABLE II

TAXONOMY OF THE METHODS THAT INCORPORATE PRE-TRAINED MODELS TO TACKLE CLASSIC CHALLENGES IN VISION-LANGUAGE TASKS. EACH METHOD CLASSIFICATION FOLLOWS A SUMMARY OF METHODS COVERED IN THIS SURVEY AND A BRIEF DESCRIPTION OF THEIR PARADIGMS.

Challenge	Paradigm	Methods
Data Scarcity	Direct inference on test samples	An LLM provides language priors, and CLIP calculates image-text similarity to introduce visual constraints. [61]–[67]
		A VLM converts visual information into texts and then an LLM processes these texts to complete the task. [68]–[75]
	Learning from unlabeled unimodal data	Use unlabeled data from the target modality for training and then replace the input features with source modality features from the common space of CLIP during testing. [76]–[83]
	Generating pseudo paired data	Employ generative pre-trained models to automatically generate pseudo annotations for training or evaluation purposes. [26], [84]–[87]
Escalating Reasoning Complexity	Divide-and-conquer	LLMs decompose the main question in a complex visual-language reasoning task into sub-questions. [88]–[95]
	Chain-of-Thought	Decompose the direct prediction process of pre-trained models into a series of intermediate reasoning steps. [96]–[106]
Generalization to Novel Samples	Extracting semantic context from an LLM	Exploit the semantic context extracted from an LLM as additional cues for processing novel samples. [9]–[12], [107]–[109]
	Distilling teacher knowledge from a VLM	Regard a VLM as the teacher model of a close-set trained student model. [110]–[117]
Task Diversity	Continual learning	Use prompt learning or instruction tuning to enable a single VLM to continuously learn new vision-language tasks, while maintaining performance on learned tasks. [118]–[120]
	Planning with natural language	Treat an LLM as a planner that, given a task-related instruction, infers to call other pre-trained models for execution. The plan comes in the form of natural language. [121]–[124]
	Planning with code statements	Treat an LLM as a planner that, given a task-related instruction, infers to call other pre-trained models for execution. The plan comes in the form of a code statement. [125]–[128]

ing [26]. Moreover, visual entailment [24] explores the logical relationships between images and sentences, requiring logical reasoning to obtain accurate answers.

Previous methods [5], [137] typically design different model architectures to process visual and optional text inputs and optimize them to compute the probability distribution of the answer vocabulary based on visual information. Since such single-step visual reasoning is oversimplified, these methods have difficulty extracting relevant visual features and external knowledge simultaneously, resulting in low-accuracy answers. Subsequently, some multi-step visual reasoning methods [138]–[140] adopt a structure known as "Retriever, Reasoner (or Reader) and Answer Predictor". Specifically, these methods extract question-relevant visual features and knowledge, iteratively perform spatial and optional temporal reasoning on the original or decomposed questions, and predict answers. However, these methods usually limit the visual reasoning steps to two hops or require the number of hops to be set as a hyper-parameter, which lacks the flexibility and robustness to solve more complex visual reasoning problems.

C. Generalization to Novel Samples

In vision-language tasks, models typically rely on the cross-modal knowledge learned from training data to make inferences on test visual samples. However, the knowledge contained in the training data is task-specific and limited in content, failing to cover all of the information that the model may encounter in real-world scenarios. Therefore, when the trained models encounter novel visual samples that are not covered in the training data, they lack the knowledge required to make accurate visual recognition or reasoning, resulting in incorrect predictions for these novel samples.

In these circumstances, external knowledge plays a crucial role in supporting the generalization to novel samples. Some early methods [141]–[146] use existing knowledge bases such as WordNet [147], Wikidata [148], or collected knowledge as external knowledge sources. By searching for relevant facts in these bases or exploring adjacent nodes in knowledge graphs, these methods extract external knowledge (typically through textual descriptions) to provide a foundational understanding of the unfamiliar visual elements in novel samples. However, these methods often suffer from complex and inflexible knowledge extraction processes and a limited richness of available knowledge content. This hinders these methods from effectively processing novel visual samples, as the complexity of knowledge extraction and the limitations of knowledge content and modality may lead to insufficient or inaccurate contextual information.

D. Task Diversity

The variety of vision-language tasks leads to great differences in the reasoning process and input-output workflow between different tasks. For example, the image editing task focuses on modifying specific visual elements within an image, where the reasoning involves identifying specific visual elements for alteration based on input instructions and generating precise image modifications as the output. In contrast, the visual entailment task focuses on reasoning about the logical relationships between an image and a textual hypothesis, with outputs restricted to predefined labels such as "entailment", "contradictory", or "neutral".

Constrained by this diversity, most supervised methods specify a single model to handle a specific vision-language task. When handling multiple tasks, methods usually follow the paradigm of training multiple target models. Each model has its own set of parameters that are updated during a specific training process. However, this paradigm imposes additional complexities in practical applications and lacks interpretability.

As a potential solution to this problem, multi-task learning [149]–[151] typically involves an encoder shared between various tasks, such as image captioning, visual question answering, and visual entailment, along with multiple task-specific decoders. This architecture allows the model to benefit from shared visual features and visual-textual patterns learned from one task to improve the performance of other tasks. However, successful multi-task learning for vision-language tasks requires carefully designed model architectures and balanced parameter update strategies, which can be challenging in practice. Moreover, when adapting a multi-task model to a new vision-language task, the visual-textual patterns learned by the model and stored in its parameters may not be optimal for the new task, and continuously updating these parameters may suffer from the catastrophic forgetting issue.

IV. Recent Advances in Pre-trained Models

A. Large Language Models

Through the new paradigm of "pre-train, prompt, and predict", LLMs have demonstrated remarkable capabilities that scale with increasing amounts of training data and model parameters. LLMs trained in an unsupervised manner can perform a variety of tasks through appropriate prompt engineering.

Specifically, LLMs can handle unseen samples via few-shot prompting [54] at inference time, where the task pattern is learned from a few input-output examples provided in the context of user input. This flexibility is a significant advantage, as it enables LLMs to adapt to a variety of tasks without extensive retraining.

In addition to few-shot prompting, instruction fine-tuning [152], [153] also plays a crucial role in improving the performance of LLMs. This technique enhances the ability of LLMs to follow instructions and enables them to perform new tasks based on the requirements of the input instructions. By training models on a diverse set of instructions, instruction fine-tuning equips LLMs with the ability to perform a wide range of tasks while being closely aligned with user intent, thus improving their effectiveness in real-world applications.

A significant breakthrough in the development of LLMs is reinforcement learning with human feedback (RLHF) [154], [155]. This technique is essential for aligning LLMs with human preferences and ensuring that LLMs are helpful, honest, and harmless. For example, InstructGPT [154] employs an effective RLHF-tuning approach that enables LLMs to follow the expected instructions, which integrates human feedback into the training process. RLHF has been widely adopted in existing LLMs, such as ChatGPT, Claude, and Bard, significantly improving their overall utility and safety in various applications.

B. Vision-Language Models

Inspired by the success of LLMs in natural language processing, VLMs have gained significant attention in the field of multimodal learning. Unlike LLMs, which are limited to processing a single data modality, VLMs excel in understanding and generating both visual content and text.

As a representative contrastive VLM, CLIP breaks the traditional supervised learning strategy of vision-language models that requires labeled datasets. Instead, CLIP directly learns from raw text about images via a simple pre-training task that predicts which caption matches which image. It encodes the input images and text with an image encoder and a text encoder, respectively, and calculates the similarities of the encoded image and text embeddings. Through contrastive pre-training on the WebImageText dataset, which contains 400M image-text pairs collected from the internet, CLIP learns a shared semantic space for visual and textual modalities. This insight has inspired subsequent research on discriminative VLMs, such as ALIGN [33], ALBEF [156], Llip [157], VideoCLIP [158], ActionCLIP [159], etc.

In addition to the above contrastive VLMs, generative VLMs also play an important role in vision-language tasks. From the perspective of pre-training techniques, BLIP [36] introduces a new model architecture for effective multi-task pre-training, called Multimodal mixture of Encoder-Decoder (MED), which contains an image-grounded text encoder and an image-grounded decoder in addition to unimodal encoders. Due to MED, BLIP is pre-trained with three vision-language objectives: image-text contrastive learning, image-text matching, and image-conditioned language modeling. Since the creation of BLIP, many new techniques have emerged, resulting in various powerful VLMs, such as Q-Former proposed in BLIP2 [160], the diffusion process in Stable Diffusion [161], factually augmented RLHF in LLaVA-RLHF [162], Direct Preference Optimization (DPO) in Diffusion-DPO [163], etc.

From the perspective of pre-training data, existing vision-language datasets, such as COCO [130], Visual Genome [164], Conceptual Captions [165], and Webvid-2M [166], are suitable choices for VLM pre-training. In addition to these datasets, many novel datasets have been constructed to improve the performance of VLMs. Specifically, Liu et al. [167] present a data reformation pipeline to convert image-text pairs into an appropriate instruction-following format using ChatGPT/GPT-4. LLaVA uses 158,000 collected language-image instruction-following samples as data, pioneering visual instruction tuning for VLMs and showing satisfying instruction-following performance. Similarly, Li et al. [168] build a novel video-centric multimodal instruction-tuning dataset. This dataset comprises thousands of videos associated with detailed descriptions and conversations, which offers a valuable resource for training VideoChat. Subsequently, Li et al. [169] present the MIMIC-IT dataset, which includes multimodal in-context information, such as multiple instruction-response pairs and multiple images or videos. Training on this dataset enables Otter [170] to exhibit remarkable proficiency in in-context learning.

C. Benchmarks

Benchmarks are essential for evaluating the capabilities of VLMs and providing a reliable data source to compare the performance of different VLMs. Benchmarks enable researchers to systematically evaluate and validate models, ensuring that advancements in VLMs are both measurable and meaningful in various tasks.

Standard vision-language benchmarks cover evaluations of a variety of common vision-language tasks, such as the COCO [130] and Nocaps [171] benchmarks for image captioning, the OK-VQA [172] and ViZWiZ VQA [173] benchmarks for visual question answering, the RefCOCO [174] and RefCOCO+ [174] benchmarks for refer expression comprehension, and the SEED-Bench [175] and MMBench [176] benchmarks for instruction-following.

In addition to these benchmarks that assess the fundamental performance of VLMs, several novel benchmarks are designed to explore the specific capabilities and limitations of VLMs. By using these benchmarks to challenge VLMs, researchers can gain deeper insights into the performance of VLMs and identify areas for further improvement.

For CLIP-like contrastive VLMs, the most commonly used evaluation task in benchmarks is image-text retrieval, where the design of samples varies. Specifically, CountBench [177] is an image-text counting benchmark for evaluating VLMs' understanding of object counting. The MMVP benchmark [178] exposes nine basic visual patterns and is specifically designed to inquire about differences in CLIP-blind pairs and to evaluate the visual abilities of VLMs using straightforward questions. The Compun benchmark [179] aims to examine the ability of VLMs to interpret compound nouns. Given a text prompt with a compound noun, the task for VLMs is to select the correct image that shows the compound noun among a pair of distractor images that show the constituent nouns that make up the compound noun. Similarly, the Cola benchmark [180] evaluates the compositional reasoning ability of VLMs by requiring them to retrieve images with the correct configuration of attributes and objects while avoiding distractor images that contain the same objects and attributes but in the wrong configuration. Moreover, CREPE [181], ARO [182], and SUGARCREPE [183] are compositionality benchmarks that explore the ability of VLMs to identify the relevant captions for a given image in a set of compositional distractors.

For BLIP-like generative VLMs, the corresponding benchmarks focus mainly on investigating the hallucinations of VLMs in different situations. For example, the collected samples in HALLUSIONBENCH [184] include unedited images such as charts, maps, and posters, as well as hand-crafted or edited images, covering topics such as math, counting, culture, cartoons, sports, and geography. In addition to the benchmarks that focus on hallucinations, an increasing number of interesting benchmarks have been developed. For example, MultipanelVQA [185] tests the multi-panel visual reasoning abilities of VLMs, while SOK-Bench [186] evaluates the commonsense reasoning abilities of VLMs in open-world videos. The WHOOPS benchmark [26] assesses counterintuitive reasoning abilities, and C-VQA [187] examines counterfactual reasoning abilities.

V. Solutions to Challenges

In this section, we introduce the existing methods for vision-language tasks that integrate pre-trained models. To provide a structured overview, we categorize these methods according to the challenges that they address, which is consistent with the four challenges outlined in Section III. Table II summarizes the target challenges, the followed paradigms, and the overall introductions of the corresponding methods covered by this survey.

A. Solutions to Data Scarcity

Benefiting from the image-text common space learned by VLMs, the language priors captured by LLMs from large-scale corpora, and the multimodal transfer capabilities of generative pre-trained models, a series of methods [26], [61]–[87] have recently emerged to address the data scarcity challenge for vision-language tasks. These methods eliminate the reliance on manually annotated data or require only a small number of paired samples to support inference, which can be broadly categorized into the following three categories. (1) Some methods [61]–[75] directly perform an inference on a single test sample, as shown in Fig. 2(a) and (b). (2) Some methods [76]–[83] use unlabeled data from the target modality for training and then replace the input features with source modality features from the common space during testing, as shown in Fig. 2(c). (3) Other methods [26], [84]–[87] employ generative pre-trained models to automatically generate pseudo annotations for training or evaluation purposes, as illustrated in Fig. 2(d). In the following sections, we will review these methods in more detail.

1) Direct inference on test samples: In the absence of available annotated data, a number of methods [61]–[75] integrate an LLM with CLIP to perform direct inferences on test samples without paired training data. As shown in Fig. 2(a), these methods employ an LLM to provide language priors during pre-training, while using CLIP to calculate image-text similarity, thereby introducing visual constraints into the test process.

Specifically, to achieve zero-shot image captioning, Tewel et al. [61] propose ZeroCap, which uses GPT-2 [188] to progressively generate image captions. Considering that GPT-2 cannot directly perceive images, this method encodes the captions generated by GPT-2 and the images into the CLIP common space and minimizes the cosine distance between these embeddings to adjust the context cache of GPT-2. This operation ensures that the caption matches the visual content. Moreover, to maintain grammatical accuracy, ZeroCap uses cross-entropy loss to constrain the generated content to be consistent with the language-prior knowledge in GPT-2. Similarly, Su et al. [62] offer a new decoding strategy that incorporates CLIP similarity scores into the decoding probability distributions of GPT-2. This strategy ensures that the obtained token probability distributions consider the similarity between the image and

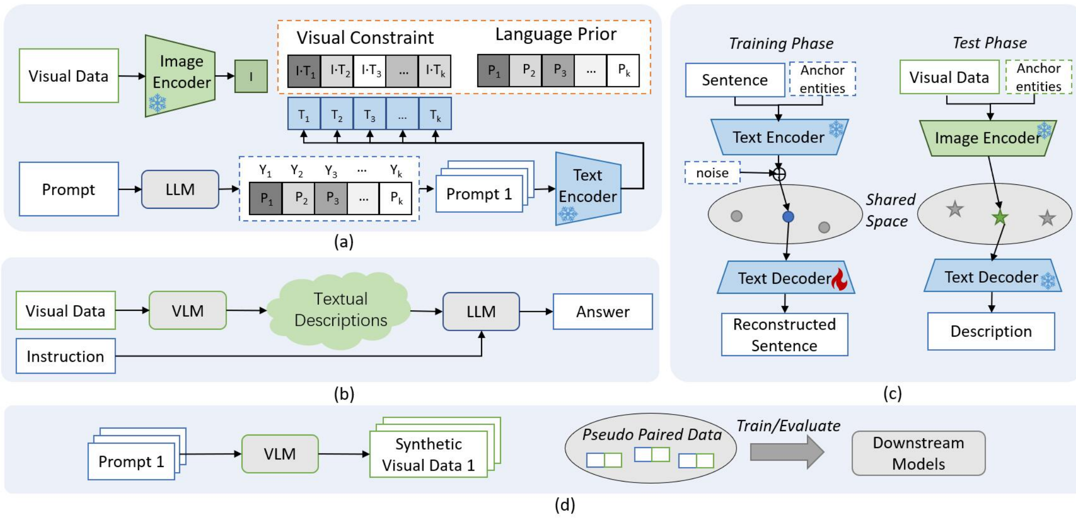

Fig. 2. Paradigms for addressing the data scarcity challenge in vision-language tasks with the help of pre-trained models. (a) shows the paradigm of integrating an LLM with CLIP to perform a direct inference on the test sample. (b) shows the paradigm of converting visual information into texts by a VLM to perform direct inferences on test samples. (c) shows the paradigm of learning from unlabeled unimodal data. (d) shows the paradigm of using a VLM to generate pseudo-paired data for training or evaluation.

the text. Zeng et al. [63] enhance ZeroCap [61] by introducing Gibbs-BERT, which adopts a sampling-based search instead of the auto-regressive generation used in ZeroCap. In addition, they incorporate a task-specific discriminator to identify the captions that align with the user control signal. To generate captions rich in world knowledge, Zeng et al. [64] design a retrieve-then-filter module to obtain key concepts relevant to the image. Using these concepts, they employ a pre-trained keywords-to-sentence LLM called CBART [189] to generate corresponding image captions under the guidance of visual relevance fusion scores. For visual storytelling, Wang et al. [65] formulate this task as a visual-conditioned generation problem, where a visual condition planner aggregates the visual conditions computed by CLIP into GPT-2.

This paradigm has been extended to video captioning tasks. For example, Tewel et al. [66] modify the token-level optimization in ZeroCap to sentence-level optimization, as token-level optimization tends to force each token to describe all frames in the video. This method introduces a cross-entropy loss similar to ZeroCap and updates the pseudo tokens in the GPT-2 prompt using total frame-caption matching scores from CLIP, rather than modifying the GPT-2 context cache. For dense video captioning, Jo et al. [67] use GPT-2 and CLIP to localize and describe the events in videos. In their method, a soft moment mask is introduced to represent video temporal segments. This method jointly optimizes the soft moment mask and the textual prefix context, which aims to accurately align the generated text with the corresponding events in the video.

Similarly, in the VQA task, Song et al. [68] convert questions into masked templates and then fill these templates by using inherent commonsense knowledge from T5 [190] to obtain multiple candidate answers. The candidate answer with the highest CLIP similarity value to the image is selected as the final answer. Yu et al. [69] employ StyleGAN [191] as a generator for counterfactual image editing in which CLIP is used to ensure the alignment between the generated images and the target counterfactual text.

In addition to combining the language priors from an LLM and the visual constraints from CLIP, another feasible idea is to use a VLM to convert visual information into texts and then process these texts by an LLM to complete the task, as shown in Fig. 2(b).

Yang et al. [70] verbalize visual information in an image by using its caption. They then prompt GPT-3 [54] to address the target VQA task by inputting the concatenation of the image caption and the question, with the caption as contextual information. This allows the generation of answers to questions in an open-ended text generation manner, which is applicable to both zero-shot and few-shot settings. Similarly, Tiong et al. [71] propose a method that generates captions for the top-k image patches most relevant to the question. A question-answering module is then designed to generate an answer based on the question and the generated captions.

Since videos contain more information than images do, video-oriented methods [72]–[75] employ more diverse tools to describe videos rather than just captions. Wang et al. [72] extract video information at both the visual-token level and the frame level to generate comprehensive video descriptions. In addition, Wang et al. [72] and Hanu et al. [73] both use automatic speech recognition to encode audio information in videos. Bhattacharya et al. [74] further enhance video descriptions by extracting subtitles from video frames using optical character recognition. For a given video, Pan et al. [75] retrieve a set of semantically similar texts from an external corpus. By using these retrieved texts along with the question, an LLM can then generate the final answer.

2) Learning from unlabeled unimodal data: CLIP learns to align image and text embeddings during pre-training. Inspired by this insight, an intuitive idea to address the data scarcity challenge is to learn from unlabeled unimodal data as an approximation of supervised training on paired vision-language data. When only text data are available, the paired

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT IMAGE CAPTIONING METHODS PROPOSED TO ADDRESS THE DATA SCARCITY CHALLENGE.

Methods	COCO				Flickr30k			
	B@4	M	C	S	B@4	M	C	S
	Supervised training							
BUTD [3]	36.2	27.0	113.5	20.3	27.3	21.7	56.5	16.0
ClipCap [192]	33.5	27.5	113.1	23.2	21.7	22.1	53.5	-
	Direct inference on test samples							
ZeroCap [61]	2.6	11.5	14.6	5.5	-	-	-	-
MAGIC [62]	12.9	17.4	49.3	11.3	6.4	13.1	20.4	7.1
ConZIC [63]	1.3	11.2	13.3	5.0	-	-	-	-
MeaCap [64]	7.1	16.6	42.5	11.8	7.2	17.8	36.5	13.1
CEPT [66]	2.2	12.7	17.2	7.3	-	-	-	-
	Text-only training							
CapDec [76]	26.4	25.1	91.8	11.9	17.7	20.0	39.1	9.9
CLOSE [77]	28.6	25.2	95.4	18.1	-	-	-	-
DeCap [78]	24.7	25.0	91.2	18.7	21.2	21.8	56.7	15.2
Knight [79]	27.8	26.4	98.9	19.6	22.6	24.0	56.3	16.3
CLMs [80]	15.0	18.7	55.7	10.9	16.8	16.2	22.5	9.8
ViECap [81]	27.2	24.8	92.9	18.2	21.4	20.1	47.9	13.6
	Generate pseudo paired data							
ICSD [84]	29.9	25.4	96.6	-	25.2	20.6	54.3	-
SynTIC [85]	29.9	25.8	101.1	19.3	22.3	22.4	56.6	16.6

visual data required for training can be replaced by the CLIP embeddings of text data. These text embeddings, together with the corresponding text data, form paired data for supervised training on the target task. During the test phase, the text embeddings are replaced by visual embeddings encoded by CLIP. The opposite is true if only visual data are available. Fig. 2(c) shows this idea of learning from unlabeled unimodal data. However, the domain gap between the CLIP embeddings of the two modalities presents a challenge to the success of this ideal replacement.

Nukrai et al. [76] make the first attempt to reduce this gap. They hypothesize that in CLIP's visual-textual common space, visual embedding lies within a small sphere around the corresponding text embedding. By injecting Gaussian noise into the input text embedding during training, all embeddings within this sphere are mapped to the same caption. On the basis of this assumption, the corresponding image embedding should also fall into this sphere and can be decoded to the correct caption. Similarly, Gu et al. [77] find that adding Gaussian noise is effective, and they explore the use of mean shift, linear adapters, and structured noise as additional methods to bridge the gap. Inspired by these findings, many efforts have been made to solve the gap problem. Li et al. [78] and Wang et al. [79] both replace image embedding with a weighted combination of similar text embeddings during the test phase. Moreover, Wang et al. [79] replace the input text embeddings with similar weighted combinations during training to further narrow the training-test gap. Some methods use entities in sentences and objects in images as anchors [80] or additional prompts [81] for the decoder to make the training input and test input more similar.

In addition to the scenarios with only text data available, several other methods [82], [83], [193] focus on using solely visual data to achieve a text-free training process. Kim et al. [82] propose a zero-shot video localization method that selects a frame and encodes it using CLIP as a pseudo-language query representation. The frame is selected from an event proposal generated based on the visual similarity of frames. During the test phase, the pseudo-language query representation is replaced by the real language query embeddings from CLIP. Similarly, CLIP-GEN [83] requires only a set of unlabeled images to train a text-to-image generator. Specifically, it trains a Transformer to convert image CLIP embeddings into discrete image tokens in the VQGAN [193] codebook space. After training, the Transformer can generate coherent image tokens from the input text CLIP embeddings, which can then be decoded into images by VQGAN.

3) Generating pseudo-paired data: When no paired data are available for training, another feasible solution is to leverage the multimodal conversion capabilities of pre-trained models to automatically generate pseudo annotations for training or evaluation, as illustrated in Fig. 2(d).

Using tools such as Stable Diffusion [161], synthetic images can be generated to create pseudo-image-sentence pairs for training an image captioning model. Specifically, Ma et al. [84] propose a method to generate multi-context synthetic images by summarizing captions describing the same image from different perspectives. Liu et al. [85] further refine pseudo image features to make them closer to natural image features. Yang et al. [86] synthesize a variety of new images based on semantic masks in the target dataset to form new training pairs with the synthetic images and conditional masks for semantic segmentation. Instead of generating data, Zhou et al. [87] use pseudo-text features derived from images to learn the text-to-image generation, eliminating the need for a general image captioning model. This paradigm is also applicable when the expected data are difficult to collect from the real world, such as counterfactual images [26] containing co-occurring elements that violate commonsense knowledge.

Accordingly, by integrating powerful pre-trained models, the aforementioned methods better bridge the gap between visual and textual modalities than the methods used before the pre-trained model era. Specifically, the previous methods typically connect two modalities from scratch solely on the basis of their shared objects or relationships. Their core components or ideas include applying object-based rewards [59], treating objects as anchors to learn a shared manifold [60], representing data in both modalities by objects [194], constructing pseudo-paired data based on object relationships [195], etc. Compared with the extensive knowledge encapsulated in pre-trained models, the information available in objects or relationships is limited, which hinders the previous methods from effectively bridging the modality gap and often results in noisy outputs.

For the above three categories of methods used to solve the challenge of data scarcity, we take the visual captioning task with the largest number of related works as an example. We summarize the performance of these methods [61]–[64], [66], [76]–[81], [84], [85] and compare them with two classic fully supervised methods [3], [192]. As representative benchmarks for the visual captioning task, COCO [130] and Flickr30K [131] are also widely used to evaluate the methods that address data scarcity in this task. The COCO dataset consists of 123,287 images, including 5,000 images

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3632653

for validation and 5,000 images for testing. Each image in COCO has approximately five crowdsourced captions. The Flickr30K dataset contains 31,783 images, including 1,000 images for validation and 1,000 images for testing. Each image in Flickr30K is annotated with five crowdsourced captions. BLEU4 [196], METEOR [197], CIDEr [198], and SPICE [199] are commonly used evaluation metrics for evaluating the quality of the generated text. BLEU4 measures precision by calculating the overlap of four-grams between the generated captions and the reference captions. METEOR considers precision and recall by aligning the generated captions with the reference captions based on stemming and synonymy. It provides a more human-like evaluation by accounting for variations in wording. As a specialized metric for visual captioning, CIDEr highlights important words by weighting n-grams according to their frequency in reference captions. SPICE evaluates the semantic content of generated captions by focusing on the mentioned objects, attributes, and relationships. Table III shows the comparison results of the methods on the COCO and Flickr30K datasets using the four aforementioned evaluation metrics.

As shown in Table III, the methods that generate pseudo-paired data achieve the best performance, which are comparable to the fully supervised methods. The second-best methods use unimodal data for training, while the methods that directly perform inferences on test samples perform slightly worse. Direct inference methods are superior in their training-free characteristics and are easy to apply in practice. These methods require only a test visual sample as input data and do not require additional unimodal training data such as image datasets or sentence corpora. However, because of the absence of a sentence corpus as a reference for the language style, these methods have difficulty generating content consistent with the ground-truth caption language style, resulting in lower scores in n-gram-based evaluation metrics. In addition, some of them use CLIP to calculate image-text similarity and sequentially select words to complete the masked sentence based on the calculated similarity scores. Such methods may encounter the risk of concept association bias caused by CLIP acting as a bag-of-words model, which will be discussed in Section VI-C. The idea of learning from unlabeled unimodal data may suffer from a discrepancy between training and inference, where text embeddings are used for training while visual embeddings are used for inferences. Fortunately, methods based on this idea continue to explore ways to narrow the modality gap between the visual and textual embeddings of CLIP, which not only results in consistent performance improvements but also provides valuable insights for subsequent studies in vision-language tasks. For the methods that generate pseudo-paired data, the strong multimodal conversion capabilities of pre-trained models enable them to effectively establish connections between visual and textual modalities. Their main drawback lies in the error accumulation during the construction of pseudo-paired data and the subsequent training process. Moreover, these methods need to develop strategies to enrich synthetic training data so that the trained model is sufficiently robust during inferences. However, the cost of constructing pseudo-paired data also needs to be considered when these methods are applied in practice, especially for expensive visual data generation.

B. Solutions to Escalating Reasoning Complexity

As the reasoning complexity in vision-language tasks, such as visual question answering, visual entailment, and visual commonsense reasoning, continues to increase, it becomes progressively challenging for models to accurately predict answers through a single reasoning step. To address this challenge, divide-and-conquer paradigms [88]–[95] and chain-of-thought (CoT) paradigms [96]–[106] have been proposed to perform multi-step reasoning by decomposing the original question and the prediction process, respectively, thus enhancing the accuracy of answers to complex questions, as shown in Fig. 3.

1) Divide-and-conquer: Because of the synergy between LLMs and VLMs, the main question in a complex visual-language reasoning task can be decomposed into a series of sub-questions about visual details and sub-questions about the factual knowledge related to the main question. Since the sub-questions are easier for models to answer than the original question and the information in their answers helps reasoning, the divide-and-conquer paradigm reduces the reasoning complexity by answering the divided sub-questions and integrating their answers to predict the final answer. This divide-and-conquer paradigm is shown in Fig. 3 (a).

A series of studies [88]–[92] have focused on exploring more visual information to help models reason about a given task. Some of these studies [88]–[90] refine the original practice of using general visual captions to represent visual content and feed it into an LLM for reasoning. For example, Chen et al. [88] design a "see, think and confirm" procedure to extract detailed object information from images. However, this method uses a fixed perception process across different questions and images. In contrast, Zhou et al. [89] develop a novel framework to automatically reveal crucial visual details related to the reasoning question, which is achieved by querying a VLM with sub-questions generated by an LLM. This paradigm is also applicable to handling information-dense images [90], such as charts or plots.

Furthermore, to ensure that the LLM gathers sufficient information, You et al. [91] use ChatGPT as the selected LLM to iteratively decompose the reasoning task. The process continues until the reasoner (*e.g.,* LLM) feels confident that it can solve the original problem based on the collected information. Similarly, Yang et al. [92] use the strong language priors of ChatGPT to generate detailed pre-questions on the basis of the original question. These pre-questions are then directly concatenated with the original question to form the input prompt for VLM reasoning.

In addition to visual information, many methods [93]–[95] emphasize acquiring factual knowledge related to the question and visual content. For example, Qi et al. [93] propose a method that recursively decomposes a question into simpler factual and visual sub-questions until they are answerable. The factual sub-questions are answered by GPT-3 [54] to provide factual knowledge, such as the release date of a movie

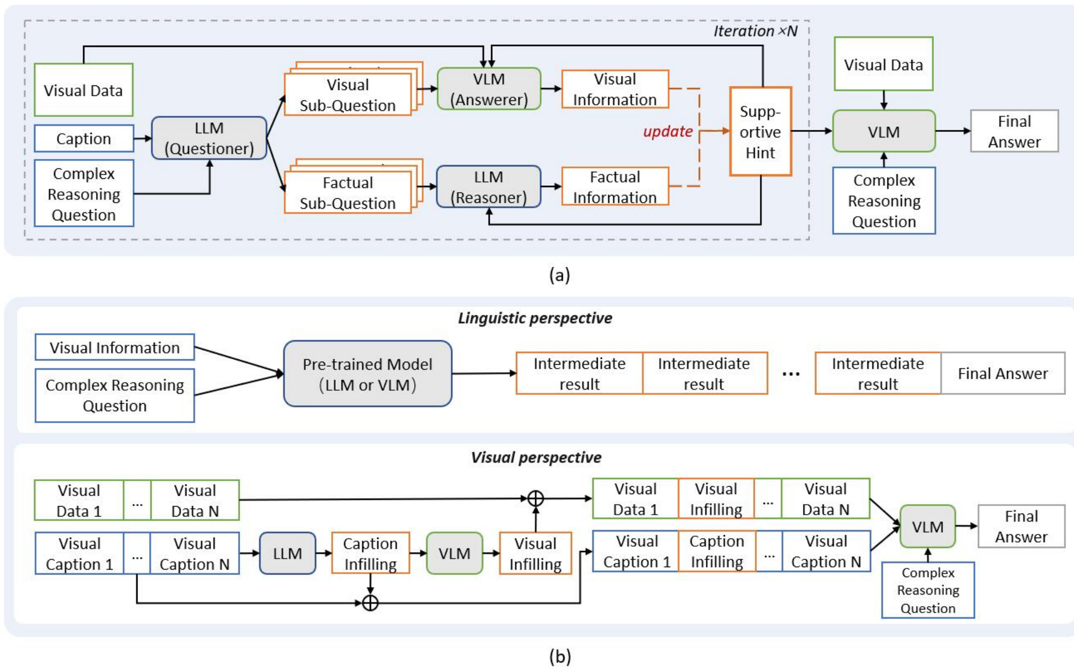

Fig. 3. Paradigms of using pre-trained models to overcome the challenge of escalating reasoning complexity in vision-language tasks. (a) shows the basic idea of the divide-and-conquer solution. (b) shows the chain-of-thought solution, which includes two different pipelines for the linguistic and visual perspectives.

TABLE IV
PERFORMANCE COMPARISON OF DIVIDE-AND-CONQUER METHODS AND CHAIN-OF-THOUGHT METHODS IN SOLVING THE ESCALATING REASONING COMPLEXITY CHALLENGE.

Methods	OK-VQA	A-OKVQA	VCR	SNLI-VE	ScienceQA
	One-step reasoning				
MiniGPT-4 [37]	37.5	58.2	40.6	35.1	47.4
	Divide-and-conquer				
IPVR [88]	44.6	46.0	-	-	-
FIIG [94]	59.3	59.8	-	-	-
ViCoR [89]	-	70.9	55.4	-	-
IdealGPT [91]	-	-	50.7	55.3	
Qvix [92]	-	-	-	50.1	55.0
	Chain-of-thought				
mm-CoT [97]	-	-	-	-	84.9
T-SciQ [98]	-	-	-	-	91.8
KAM-CoT [99]	-	-	-	-	92.5

depicted in an image, which cannot be directly observed from the visual content alone. After gathering visual and factual hints, Wang et al. [94] introduce a refinement module to filter out irrelevant hints and summarize useful ones. Rajabzadeh et al. [95] employ a divide-and-conquer strategy that uses tool interactions supplemented by a web search to provide additional supporting hints.

2) Chain-of-thought: The chain-of-thought paradigm [200] decomposes the direct prediction process of an LLM into a series of intermediate reasoning steps. Inspired by its success in addressing complex natural language problems, many researchers [96]–[106] have adopted this paradigm to solve complex visual-language reasoning problems.

Early methods explore the use of LLMs for multimodal chain-of-thought reasoning [96]–[99], where the LLM generates rationales or step-by-step thought processes as intermediate outputs to improve the accuracy of the final answers. The main concern of this approach lies in how to effectively inject visual information into an LLM. Lu et al. [96] represent visual content through corresponding captions and use this approach to evaluate several LLM baselines on a newly proposed benchmark ScienceQA. This benchmark includes data examples that consist of multimodal question-answering information and grounded lectures and explanations. Zhang et al. [97] later introduce a two-stage framework to sequentially perform rationale generation and rationale-based answer reasoning by fine-tuning an LLM to accept the fused image and language features as inputs. Given the challenge of collecting a high-quality CoT corpus, Wang et al. [98] follow a plan-and-solve prompting process to automatically generate CoT rationales in a zero-shot manner. These rationales serve as teaching signals to fine-tune a student VLM. In addition, Mondal et al. [99] incorporate external knowledge from a knowledge graph during reasoning to provide supplementary contextual information to an LLM, thereby improving reasoning performance.

Subsequently, the research focus has shifted toward exploring the unique chain-of-thought reasoning process for visual-language tasks [100]–[103], starting with evaluating the chain-of-thought capabilities of various VLMs and then enhancing them. Chen et al. [100] establish the CURE benchmark and evaluate existing VLMs. They observe that most VLMs have difficulty with CoT reasoning and propose a method for VLMs to learn the CoT reasoning ability from an LLM. This method involves training the VLM to generate rationales under the guidance and feedback of an LLM. For visual document understanding, Zhu et al. [101] integrate outputs from OCR tools, an LLM, and a multimodal-verifier to form teacher

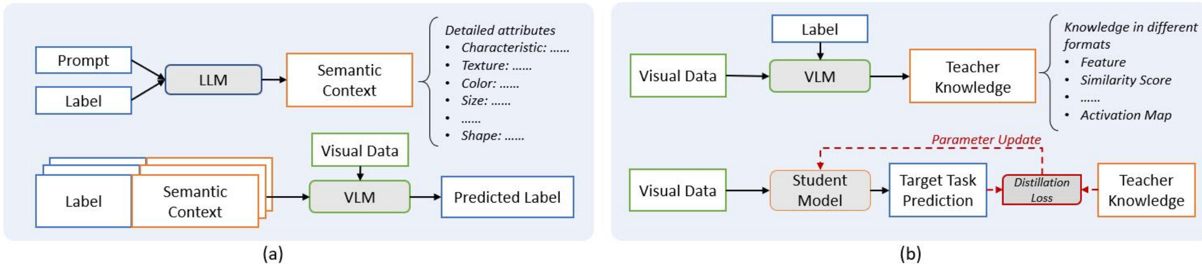

Fig. 4. Paradigms of using pre-trained models to solve the generalization to novel samples challenge in vision-language tasks. (a) shows the basic idea of extracting semantic context from an LLM, while (b) shows the basic idea of distilling teacher knowledge from a VLM.

rationales. These rationales are then used to train a small student VLM to predict both rationales and answers for input questions. The aforementioned methods still require additional training and support from other models. Mitra et al. [102] and Zhang et al. [103] propose CCoT and CoCoT, respectively, which directly obtain rationales from the VLM itself through prompting without extra training. Specifically, CCoT [102] regards scene graphs generated by the VLM as intermediate rationales for better compositional reasoning. In contrast, CoCoT [103] prompts the VLM to derive comparison results across multiple input images as intermediate rationales to enhance the VLM's reasoning capabilities on tasks involving multiple images.

All of the above methods embody the thinking process from a linguistic perspective, as shown at the top of Fig. 3(b). However, as depicted at the bottom of Fig. 3(b), another group of methods [104]–[106] seeks to embody this process from a visual perspective by integrating visual augmentation into reasoning. Meng et al. [104] suggest that language-based reasoning is often too complex and abstract. To address this problem, they propose CoI that transforms the conventional textual step-by-step reasoning into a sequence of generated images. This visual augmentation technique is also effective for processing more complex sequential visual data. Benefiting from the multimodal generative capabilities of VLMs, both Rose et al. [105] and Himakunthala et al. [106] aim to bridge the existing logical gaps in sequential data by generating multi-modal fillings. This strategy enriches the temporal and causal information derived from visual content, thereby improving reasoning performance.

The following three complex visual reasoning tasks are commonly used to evaluate the reasoning capabilities of different methods: knowledge-based visual question answering, visual commonsense reasoning, and visual entailment. For knowledge-based visual question answering, OK-VQA [172], A-OKVQA [201], and ScienceQA [96] are representative benchmarks. OK-VQA is an open-domain dataset that comprises 14,055 image-question pairs that span various knowledge domains, including transportation, food, and weather, with each question compared with ten ground-truth answers. This benchmark challenges methods to reason using commonsense and domain-specific knowledge. A-OKVQA extends OK-VQA by scaling up the dataset and task complexity, providing 24,903 samples with justifications for answers. Unlike OK-VQA and A-OKVQA, ScienceQA focuses on scientific reasoning and covers questions from natural science, social science, and language science. Each question in ScienceQA is paired with detailed annotations, including lectures and explanations of answers. For visual commonsense reasoning, VCR [23] consists of 290,000 samples derived from 110,000 unique movie scenes, and it emphasizes multi-step reasoning in human-centric scenarios. For visual entailment, SNLI-VE [24] is a typical benchmark that requires methods to determine whether an image semantically entails a textual hypothesis. This benchmark contains approximately 570,000 image-hypothesis-answer triplets. For these benchmarks, the accuracy of the generated answers serves as the evaluation metric. We summarize the performance of these methods [88], [89], [91], [92], [94], [97]–[99] on addressing the escalating reasoning complexity challenge and compare them with the one-step reasoning results from the pre-trained MiniGPT-4 model [37]. Table IV shows the comparison results of these methods on three complex visual reasoning tasks and uses accuracy as the evaluation metric.

As shown in Table IV, the divide-and-conquer and chain-of-thought methods outperform the one-step reasoning results of MiniGPT-4 on all datasets. Notably, the chain-of-thought methods [97]–[99] also surpass Qvix [92] in ScienceQA since the chain-of-thought methods involve fine-tuning with teacher data from an LLM. These comparisons demonstrate that decomposing the original question or the reasoning process is an effective paradigm for solving complex visual reasoning problems. Specifically, divide-and-conquer methods enrich the input information that supports pre-trained models to reason about complex questions. The enriched input contains detailed visual information and factual knowledge that are useful for reasoning, which is obtained by answering sub-questions derived from the original reasoning question. This paradigm is suitable for addressing complex reasoning questions that require a thorough comprehension of visual information and invisible background knowledge. However, as the original question is divided into multiple sub-questions, these methods suffer from the risk of error accumulation in the process of collecting the answers to the sub-questions. Chain-of-thought methods enhance the reasoning process of pre-trained models by encouraging them to perform step-by-step reasoning and output thinking processes instead of just providing an isolated answer. This paradigm not only improves reasoning accuracy but also increases the interpretability of integrating pre-trained models to complete vision-language tasks. Moreover, chain-

TABLE V
PERFORMANCE COMPARISON OF DIFFERENT OPEN-VOCABULARY IMAGE CLASSIFICATION METHODS PROPOSED TO ADDRESS THE GENERALIZATION TO NOVEL SAMPLES CHALLENGE.

Methods	ImageNet	CUB	Food101	Place365	Oxford Pets	Describable Texture	Mean
CLIP [32]	58.46	51.95	79.31	37.37	79.94	41.38	58.07
VCD [9]	62.97	52.57	83.63	39.90	83.46	44.26	61.13
LCDAtt [10]	-	64.05	81.85	-	85.91	-	-
CPHC [12]	63.88	54.18	83.02	40.73	82.69	48.19	62.12

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT OPEN-VOCABULARY OBJECT DETECTION METHODS PROPOSED TO ADDRESS THE GENERALIZATION TO NOVEL SAMPLES CHALLENGE.

Methods	LVIS		COCO		
	AP_r	AP	Novel AP	Base AP	Overall AP
Supervised-RFS [202]	12.3	24.3	-	-	-
OVR-CNN [203]	-	-	22.8	46.0	39.9
ViLD [110]	16.6	25.5	27.6	59.5	51.3
PB-OVD [113]	-	-	30.8	46.1	42.1
FVLM [114]	18.6	24.2	28.0	43.7	39.6
BARON [111]	22.6	27.6	42.7	54.9	51.7
OADP [112]	21.9	28.7	30.0	53.3	47.0

of-thought methods that embody the thinking process from a visual perspective are effective in circumstances where reasoning over complex sequential visual data or step-by-step visual reasoning is needed, since these methods can use more complementary visual information for reasoning. Despite the aforementioned strengths, most chain-of-thought methods involving VLMs require fine-tuning on manually annotated multimodal chain-of-thought data, and the annotation of such data is expensive and requires expert knowledge.

C. Solutions to the Generalization to Novel Samples

When a model encounters novel visual samples unseen during the training phase, it cannot make accurate inferences because of the lack of knowledge about the visual appearance and semantic meaning of these samples. Pre-trained models that have a large number of parameters are trained extensively on massive datasets, and their parameters contain a wealth of world knowledge. Therefore, from both the visual and semantic perspectives, pre-trained models are suitable choices to provide external knowledge for novel samples. As shown in Fig. 4, the solutions to the generalization to novel samples challenge can be roughly divided into two categories, namely, extracting semantic context from an LLM [9]–[12], [107]–[109] and distilling teacher knowledge from a VLM [110]–[114].

1) Extracting semantic context from an LLM: To better generalize to novel samples, a series of solutions exploit semantic context as additional cues for processing these samples. In these solutions, the semantic context enriches the model's comprehension of the novel samples by providing visual semantic information that aids in inference.

Considering the extensive world knowledge embedded in LLMs and the convenience of querying, several methods [9]–[12], [107]–[109] use LLMs as a source of semantic context. Menon and Vondrick [9] make the first attempt to use language descriptions as external semantic contexts for visual recognition. Specifically, they query an LLM with rich world knowledge to obtain the descriptive features of classes and use these class descriptors for classification via CLIP. This modification not only improves classification accuracy but also enhances the interpretability of CLIP decisions.

Nevertheless, descriptions generated by an LLM often contain low discriminative information and noise. Strategies for selecting or generating concise descriptions have become a hot topic of subsequent research [10]–[12]. Yan et al. [10] propose a novel learning-to-search method that discovers concise sets of attributes from the original generation while maintaining classification performance. Dai et al. [11] address the problem of unfaithful generated descriptors caused by LLM hallucinations. Their method determines when to trust the outputs of an LLM by estimating confidence scores using consistency-based uncertainty calibration. Moreover, Ren et al. [12] focus on the problem that descriptions of similar but different categories are indistinguishable when generating results for each category separately. They use ChatGPT to compare and group categories, constructing a category hierarchy through hierarchical comparison, which makes the decision boundary of each category more compact.

In addition to image classification, the paradigm of deriving semantic context from an LLM has been extended to relationship detection [107] and action recognition [108], [109], [204]. Li et al. [107] design a procedure to decompose each predicate category into subject, object, and spatial components and then query ChatGPT to generate descriptions of each component to construct composite descriptions as new prompts. This helps CLIP distinguish different fine-grained relation types. In the video domain, Jia et al. [108] identify 12 pivotal attributes related to the scene, actor, and body aspects. Based on these attributes, GPT-4 hierarchically generates knowledge-rich descriptions to form action-conditioned prompts. Shi et al. [204] enhance the semantics of action categories by using BERT [52] to collect text proposals that contain language descriptions of actions. These proposals form an action knowledge base that allows CLIP to extract action semantics by calculating the similarities between text proposals and video frames. In addition to enriching the semantic context for class label representations, Yousaf et al. [109] propose converting videos into captions and then using ChatGPT to convert these captions into language attributes and descriptions. These descriptive visual cues are integrated with visual embeddings to provide additional semantic context for various downstream tasks.

For the above methods that extract semantic context from an LLM to address the generalization challenge, we summarize their performance on the open-vocabulary image classification task and compare them with the method that uses only CLIP for classification. Six image classification benchmarks are commonly used to evaluate the effectiveness of the methods in open-vocabulary image classification: ImageNet [205], CUB [206], Food101 [207], Place365 [208], Oxford Pets [209]

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2025.3632653

and Describable Textures [210]. These benchmarks differ in their class coverage. Specifically, ImageNet contains over 14 million images across approximately 220,000 daily object classes; CUB focuses on fine-grained classification with 11,788 images across 200 bird species; Food101 contains 101,000 images of 101 food classes; Place365 has more than 180,000 images spanning 365 scene classes; Oxford Pets contains 3,680 images of 37 cat and dog breeds; and Describable Texture contains 5,640 images across 47 texture and pattern classes. In these datasets, the classification accuracy is used as the evaluation metric. Table V shows the comparison results of the methods [9], [10], [12], [203] using ViT-B/32 [211] as the backbone model on the aforementioned six image classification benchmarks. The VCD [9], LCDAtt [10], and CPHC [12] methods all outperform the base CLIP model, demonstrating the advantage of extracting the semantic context from an LLM in improving the model's generalization to novel samples.

2) Distilling teacher knowledge from a VLM: Since the knowledge contained in a VLM is much more extensive than that contained in a close-set model trained on downstream datasets, the VLM can be regarded as the teacher model for the close-set trained student model. By distilling the useful teacher knowledge in the VLM into the student model, the student model can acquire the ability to handle novel samples. The main focus of this knowledge distillation process is how to effectively represent and distill the teacher knowledge.

Gu et al. [110] make the first attempt to distill the teacher knowledge from an open-vocabulary image classification model [32] to enable open-vocabulary object detection. This is achieved by aligning the region embeddings of the detected boxes with the text and image embeddings inferred by the teacher classification model. In addition to distilling individual region embeddings, Wu et al. [111] propose aligning the embeddings of a bag of regions to encourage the student model to understand the scene comprehensively rather than just focusing on isolated objects. This method treats the bag of regions as a bag of words to obtain the bag-of-regions embeddings, which are then aligned with corresponding visual features from the CLIP image encoder. Similarly, Wang et al. [112] introduce a distillation pyramid mechanism to supplement the missing relational information in object distillation.

In addition to distilling knowledge by aligning embeddings, Gao et al. [113] exploit the localization capability of the teacher model ALBEF [212] to generate pseudo bounding-box labels from image-caption pairs, which can be used as training data for the student detector. Kuo et al. [114] observe that CLIP preserves the local sensitivity for detection and thus they adopt a more straightforward approach for distillation by directly incorporating the teacher model into a new student model. Specifically, they train a detector head on top of the frozen teacher CLIP and obtain the final inference results from a combination of detection scores and CLIP predictions.

For open-set recognition, Liao et al. [115] first build a common and scalable semantic hierarchy with all downstream datasets aligned. They prompt-tune a VLM to first recognize unknown classes and then prompt the fine-tuned VLM with handcrafted prompts for zero-shot classification. To mitigate label bias, Liao et al. [116] propose a prompt tuning method that introduces open words from WordNet, thereby extending the prompt texts from close-set label words to more. In this way, the prompts are tuned in a simulated open-set scenario. Qu et al. [117] propose to simulate additional virtual open-set classes and use VLMs to generate images for both the closed-set and simulated open-set classes. These images serve as references in the inference phase of VLMs, in order to reduce the models' reliance on spurious-discriminative features.

For the above methods that distill teacher knowledge from a VLM to address the generalization challenge, we summarize their performance on the open-vocabulary object detection task and compare them with two baselines: Supervised-RFS [202], which is trained using both novel and base labels, and OVR-CNN [203], a caption-enhanced baseline that does not use any VLM as a knowledge provider. For open vocabulary object detection methods, COCO [130] and LVIS [202] are widely used benchmarks. Following the common setting [213], the training set of COCO is divided into a base set with 48 base classes and a target set with 17 novel classes. In total, COCO contains 107,761 training images with 665,387 bounding box annotations for the base classes and 4,836 test images with 28,538 bounding box annotations for both the base and novel categories. In LVIS, 866 frequent and common categories are treated as base categories, while 366 rare categories serve as novel categories. To evaluate a method's ability to detect novel objects, COCO uses the average precision of novel categories (*i.e.,* Novel AP) as the main evaluation metric, while LVIS uses the average precision of rare categories (*i.e.* Ap_r). Table VI shows the comparison results of the methods [110]–[114] on COCO and LVIS when ResNet-50 [214] is used as the backbone model.

As shown in Table VI, the ViLD [110], PB-OVD [113], FVLM [114] BARON [111], and OADP [112] methods all outperform the baselines on the novel classes while maintaining stable or even higher performance on the base classes and overall performance. These results demonstrate the advantage of distilling knowledge from a VLM in enhancing the generalization to novel samples.

For the methods that extract semantic context from LLMs, their inference processes on novel samples are highly interpretable because the information in the extracted context is easy to understand. Therefore, if errors occur during the inference process, the interpretability of the semantic context can effectively help researchers trace back the inference process and identify where the model's understanding differs from the expected outcome. This allows researchers to figure out the issue and optimize the method accordingly. However, because of VLMs' confusion on compositional concepts, some fine-grained contexts that vary in detailed visual attributes may still be ambiguous to VLMs, which hinders VLMs from distinguishing visually similar samples. How to better utilize fine-grained contexts remains an open question. On the other hand, distilling knowledge from VLMs allows close-set student models to benefit from the rich visual and textual understanding that VLMs have learned, thereby establishing a more generalized connection between visual and textual modalities. This can significantly improve the ability of student models to handle open-vocabulary tasks without requiring extensive task-specific training data. Nevertheless, knowledge distillation inevitably leads to a loss of teacher knowledge, thereby reducing the generalization capability of the student models. This solution can potentially be improved by extracting additional semantic context from LLMs to compensate for the knowledge loss. The integration of these two types of methods may provide new insights for mitigating the generalization to novel samples challenge.

D. Solutions to Task Diversity

Because of the diversity of vision-language tasks in terms of input-output formats and reasoning processes, designing a general model that can handle multiple tasks is challenging. In this context, designing specific models for different tasks and updating the parameters through training place a great burden on practical applications.

Considering the adaptability of VLMs, a new trend has emerged that focuses on the continual learning of VLMs. Novel methods [118]–[120] of prompt learning and instruction tuning have been developed to incrementally update VLM knowledge. These methods allow a VLM to adapt to new vision-language tasks while retaining previously learned knowledge. Moreover, they are highly parameter efficient and require only minor updates to a subset of the model's parameters or the integration of lightweight modules. This advancement represents a significant step toward enhancing the generality of VLMs in addressing the growing diversity of vision-language tasks. In addition to the continual learning of a single VLM, another new trend has emerged to extend traditional task-specific models to a general modular system [121]–[128]. The core of this system is the LLM as a planner. Given a task-related instruction, the planner infers a plan that consists of a sequence of operations to be executed by various tools. Following the plan, the system sequentially calls upon different tools, including pre-trained models and pre-defined modules, to generate the final response to the input instruction. As shown in Fig. 5, the plan can be natural language [121]–[124] or code statements [125]–[128]. Since the pre-trained models have strong in-context learning capabilities, the general system does not require an additional training process. With flexible planners and plug-and-play tools, such a system is not affected by task diversity and can address various visual-language tasks simultaneously, such as visual captioning, natural language image editing, factual knowledge object tagging, visual mathematical reasoning, and multi-image reasoning.

1) Continual Learning: Given their extensive knowledge storage, VLMs serve as strong starting points for adapting to various vision-language tasks. Moreover, continual learning techniques can alleviate the issue of catastrophic forgetting during VLM adaptation, thereby allowing VLMs to adapt to new tasks while maintaining proficiency in previous tasks. To enable VLMs to handle various types of VQA tasks, Qian et al. [118] propose a multimodal prompt learning method for VLMs that consists of decoupled prompts and prompt interaction strategies to capture the complex interactions between visual and textual modalities. When new images or new question types that require different reasoning processes appear, the proposed method updates its decoupled prompts to learn how to perform new tasks while maintaining the ability to solve previous tasks. Chen et al. [119] introduce COIN, a novel benchmark designed to evaluate VLMs in continual instruction tuning. COIN consists of 10 datasets that span eight different tasks, including referring expression comprehension, classification, knowledge-grounded image question answering, etc. Experimental results on COIN identify that catastrophic forgetting primarily arises from VLMs' failures in intention alignment, which can be mitigated by introducing MoELoRA to use experts to acquire task-specific knowledge. In addition to catastrophic forgetting, Zheng et al. [120] discover a negative forward transfer in VLMs, where learning new tasks leads to performance degradation on unseen tasks. To achieve positive forward transfer, they suggest reusing pre-trained knowledge and projecting prompt gradients to the pre-trained space. Moreover, they allocate different subspaces for each task and project prompt gradients to the residual space to achieve anti-forgetting. These advancements highlight the effectiveness of continual learning in addressing the growing diversity of vision-language tasks, enabling a VLM to maintain adaptability and robustness across a wide range of tasks.

2) Planning with natural language: Inspired by ChatGPT's ability to handle NLP problems across multiple domains, several works [121], [122] integrate visual experts with ChatGPT to create a system that can perform numerous visual-language tasks. In this system, ChatGPT acts as a planner and outputs a plan in a natural language format to call upon other tools.

Specifically, Yang et al. [121] inject the usage knowledge of visual experts into ChatGPT by adding instructions about the capabilities of each expert and providing some contextual examples for each expert within the prompts. Moreover, to facilitate ChatGPT in calling various expert models, it is instructed to generate special watchwords when a visual expert is needed to interpret the visual input. The output of the expert is serialized and combined with the historical message to further activate ChatGPT until no more experts are needed. The system then returns the final answer to the user. Following this paradigm, the proposed system can address challenging visual understanding tasks such as visual mathematical and textual reasoning, open-world concept understanding, visual planning and prediction, etc. Similarly, Wu et al. [122] introduce a new prompt manager that converts visual content, historical intermediates, and information from visual experts into understandable prompts for ChatGPT.

In addition to incorporating off-the-shelf vision models as plug-and-play modules, Lu et al. [123] integrate web search engines and several heuristic-based modules, such as Bing Search and OpenAI tools, into their system, to enhance the reasoning capabilities of the system, including knowledge retrieval, web search, mathematical reasoning, and table understanding. As a result, the range of applications that can be achieved by this system has been expanded.

The success of the above methods in the image domain encourages the extension of this paradigm to the video domain. Lin et al. [124] design a system that takes a video file as input and outputs a script describing the video content. This script

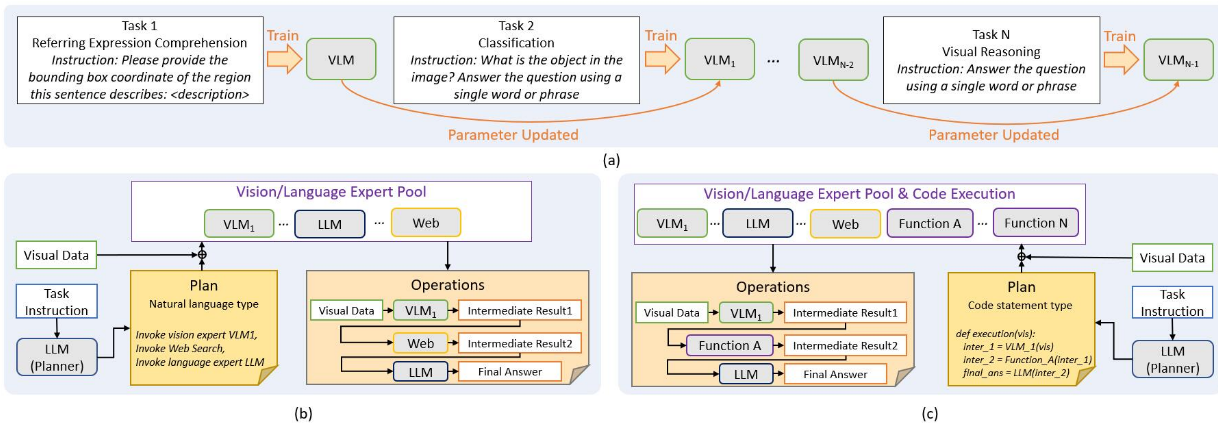

Fig. 5. Paradigm of using pre-trained models to conquer the task diversity challenge in vision-language tasks. (a) shows the basic idea of applying continual learning strategies in a single VLM to enable it to address different vision-language tasks. (b) and (c) show the basic idea of building a general modular system that plans with natural language and code statements, respectively, to address different vision-language tasks.

TABLE VII
A COMPARISON OF THE GENERAL SYSTEMS PROPOSED TO ADDRESS THE TASK DIVERSITY CHALLENGE.

Methods	Visual Data		Planner		Tool			
	Image	Video	Planner Name	Planning Format	Experts	Web Search	Python	Tool Size
MM-REACT [121]	√		ChatGPT	natural lang	√	√		10
Visual ChatGPT [122]	√		ChatGPT	natural lang	√			22
Chameleon [123]	√		GPT-4	natural lang	√	√	√	13
VisProg [125]	√		GPT-3	Code	√		√	20
ViperGPT [126]	√	√	GPT-3 Codex	Code	√		√	15
MM-VID [124]		√	GPT-4	natural lang	√			4
ProViQ [127]		√	ChatGPT	Code	√		√	9
ZS-CVR [128]	√	√	PaLM2 Code-bison	Code	√		√	14

enables an LLM to gain a generalized understanding of the video and perform various video tasks based on the user's questions.

3) Planning with code statements: As a highly structured language, code is more precise than natural language. Code statements can be executed directly by machines in sequence, including conditional statements, arithmetic operations, and defined functions. Furthermore, the result returned by the previous function can be used as the input of the subsequent function to achieve information transmission. Given this insight, Gupta et al. [125] and Suris et al. [126] propose the use of Python programs generated from text queries to call other visual experts to process visual inputs. These methods benefit from the latest code generation LLMs (*e.g.,* GPT-3 Codex [215]) or LLMs (*e.g.,* GPT-3 [54]) and enable efficient program generation beyond manually created programs. The inclusion of built-in Python logic and mathematical operators further improves the logic of the system.

Choudhury et al. [127] also adopt the paradigm of invoking other modules through Python programs to design their system. For the input video data, the system performs reasoning at multiple semantic levels by incorporating information from individual frames, disjoint video clips, and the entire video. This capability is supported by image-based modules such as object detection and image QA and video-based modules such as video retrieval, video captioning, speech transcription, tracking and video summarization. Considering that video reasoning tasks require strong spatial and temporal reasoning capabilities, Stanić et al. [128] propose dividing abstract routines into spatial and temporal categories. Then, they predefine certain operations according to prior knowledge, such as spatial routines for retrieving relations relative to a patch as well as temporal routines for event localization.

The above methods address the challenge of task diversity by designing general systems in three aspects: the types of visual data processed by the system, the characteristics of the planner, and the tools used. This summary is shown in Table VII.

A general modular system that integrates multiple pre-trained models can perform a wide range of vision-language tasks in a zero-shot or few-shot manner. This characteristic is particularly advantageous, as it minimizes the need for extensive retraining or fine-tuning on new tasks. In comparison, even though continual learning avoids the massive cost of retraining VLMs, it still requires parameter optimization and storage when adapting to new tasks. Consequently, a general modular system is more efficient and scalable than continual learning. Moreover, by invoking corresponding APIs to execute different pre-trained models, this system further eliminates the need for locally storing model parameters and simplifies deployment in resource-constrained environments. In addition to the computational benefits, general modular system planning with code statements benefits from the direct code execution of arithmetic operations and conditional statements, which are typically challenging for pre-trained models to perform accurately on their own. Moreover, the solution of designing a general system provides high interpretability through user-friendly intermediate outputs and a transparent solving process. Additionally, the plug-and-play configuration of both the planner and tools increases the flexibility of the system. When a well-performing tool is newly released, it can easily replace the previous tool by simply adjusting the prompts and in-context examples. However, attempts at video data are still ongoing, and adjustments for video characteristics require further exploration. Video data contain much more information than a single image does. For a general system, it is critical to find feasible ways to ensure that useful information is extracted without being affected by redundant data. Furthermore, for extremely long videos (*e.g.,* movies), an ideal system needs to represent complex video information in a more compact form to avoid exceeding the maximum acceptable length of the LLM planner.

VI. Potential Risks

As discussed in the previous sections, vision-language tasks benefit greatly from pre-trained models because of their powerful advantages, such as zero-shot inference ability, knowledge implicitly stored in the parameters, learned common vision-language space, etc. Despite these well-known strengths and their remarkable performance on downstream tasks, pre-trained models present weaknesses in certain areas, including hallucinations, outdated knowledge, concept association bias, and compositional concept confusion. These issues pose potential risks when pre-trained models are introduced to vision-language tasks. In this section, we summarize these typical issues and further discuss potential research directions where future progress can be made.

A. Hallucinations

When pre-trained models are used to solve the challenges in visual-language tasks, the hallucination problem of pre-trained models is introduced. Specifically, the methods [70]–[75] that use VLMs to verbalize visual content are easily affected by the visual illusion problem of VLMs [216], resulting in hallucinations in visual descriptions. These hallucinations arise from the architecture of generative VLMs, which typically combines a pre-trained visual encoder, an LLM, and a trainable Q-Former-like connection module. The visual component is often weaker than the language component, leading to the dominance of the inherent bias of the language module. As a result, VLMs tend to overly rely on language priors to answer visual questions, especially when the visual content contains out-of-domain information [217]. Similarly, the methods [88]–[95] that use LLMs for reasoning are prone to the negative impacts of the hallucination problem in LLMs [218], including input-conflicting hallucinations, context-conflicting hallucinations and fact-conflicting hallucinations, leading to unfaithful or nonsensical reasoning results. Moreover, the hallucination problem becomes more severe if the method involves chain-like interactions between pre-trained models. In this case, the hallucinations generated at any step are accepted as credible inputs of another model in subsequent steps, which causes a snowball-like accumulation of hallucinations [219]. For example, a complex reasoning VQA sample contains an image of a girl holding cookie cutters in preparation for baking, along with the following question: "What might the girl complete after a while?" Divide-and-conquer methods [88]–[95] typically start by describing the visual information using a VLM. However, because of the hallucination issue, the model may incorrectly describe the scene as the girl playing with toy blocks, which diverges from the actual context. This hallucinated visual description can mislead subsequent reasoning steps and lead to cumulative errors. Specifically, these methods rely on toy blocks as visual hints and may incorrectly infer that the girl will build a castle instead of correctly predicting that she will bake cookies.

Despite the aforementioned negative impacts of hallucinations, the current methods that introduce pre-trained models into vision-language tasks often overlook this issue and fail to implement strategies to mitigate it. As a suggestion for future work, two potential strategies can be incorporated to address the hallucinations that arise from the use of pre-trained models in vision-language tasks.

The first strategy is to design hallucination mitigation techniques for pre-trained models, especially those that do not require additional training, such as general decoding strategies and prompting techniques. Most of these techniques are plug-and-play and can be easily implemented through various methods at minimal cost. For example, prompt engineering that explicitly instructs VLMs or LLMs not to spread false or unverifiable information, such as "If you don't know the answer to a question, please don't share false information", which is proposed by Touvron et al. [29]. In addition, simple logical consistency checks of objects in VLM responses can effectively reduce hallucinations. The second strategy involves designing appropriate verification mechanisms during the interaction between pre-trained models. Potential verification mechanisms include cross-referencing outputs from different models or integrating a feedback loop in which the outputs are assessed against the original visual inputs. Moreover, incorporating human-in-the-loop approaches, where human reviewers assess the accuracy of model predictions, can further refine the verification process. This would allow for a dynamic learning environment where the model can adjust based on real-time feedback, thereby reducing the incidence of hallucinations. These mechanisms ensure that credible results are passed to subsequent models, thus avoiding the accumulation of hallucinations.

B. Outdated Knowledge

Real-world knowledge is constantly changing and updating, but pre-trained models struggle to keep up-to-date knowledge because they implicitly store knowledge in static parameters [220]. Since the knowledge in the pre-trained models is limited by the timeliness of training data, extracting knowledge from the pre-trained models to support complex reasoning in vision-language tasks may lead to the use of outdated knowledge. If reasoning about complex vision-language questions requires specific knowledge and the knowledge is updated after the pre-trained model is trained, then the pre-trained model can either provide outdated information or not provide relevant information at all. Using outdated knowledge containing incorrect or empty information as the basis for reasoning leads to deviations in the subsequent reasoning process, thereby ultimately resulting in incorrect conclusions. For example, a knowledge-based VQA sample contains an image of a crowd celebrating a recently awarded Nobel Prize winner, paired with a question: "What impact has the winner's achievement had on society?" As Nobel Prize events evolve over time, pre-trained models are unable to update their internal knowledge to include newly awarded winners. Consequently, VLMs struggle to recognize these new winners, while LLMs fail to provide background information regarding the winners' achievements. Without access to up-to-date information, the methods that rely solely on static knowledge in pre-trained models struggle to associate visual content with the relevant information needed to answer the question. Suffering from the outdated knowledge issue of pre-trained models, these methods fail to correctly reason about this sample.

As solutions to the outdated knowledge issue arising from integrating LLMs into NLP tasks, knowledge editing and retrieval-augmented generation (RAG) are practicable for integrating pre-trained models into vision-language tasks. Specifically, knowledge editing updates and refines the internal knowledge of pre-trained models to ensure that it incorporates the latest information relevant to the task. This approach either introduces an auxiliary network or modifies a subset of the model's parameters to embed new knowledge, thereby aligning the model's output with the latest information. On the other hand, RAG complements pre-trained models with up-to-date knowledge retrieved from the internet or other real-time resources. Both visual content [221], [222] and text [223], [224] can serve as references for retrieval, and the retrieval results are injected into the input of pre-trained models or fused with the models' raw outputs. This integration of external knowledge enriches a model's context and thus enhances its ability to generate accurate and trustworthy responses.

C. Concept Association Bias

VLMs pre-trained using image-text contrastive objectives such as CLIP are widely used to calculate the similarities between images and sentences, which serve as a common constraint to ensure the consistency between generated sentences and visual content. However, many studies [182], [225], [226] highlight that such VLMs tend to treat an input sentence as a bag of concepts and ignore the syntactic structure of the sentence. As a result, relying on similarities calculated by VLMs is prone to concept association bias, especially for the visual captioning and VQA methods that perform direct inference on test samples, which is introduced in Sec V-A1. Specifically, these methods use a VLM to fill in a masked sentence by maximizing the similarity between the filled sentence and the image, where the masked sentence is a partially generated image caption or a template derived from a visual question. Because of the bag-of-words nature of VLMs, they tend to assign higher similarity scores to sentences that encompass a wider range of elements depicted in the image. Since the masked sentences usually focus on partial concepts of the image, the filling content will ignore the sentence semantics and bias toward the missing concepts. For example, given an image with a purple eggplant and a yellow lemon, CLIP needs to complete the masked sentence, "In this picture, the color of the lemon is [mask]", by selecting a word that maximizes the visual-textual similarity score between the image and the full sentence. Since the image contains two different visual elements but the masked sentence mentions only the lemon, CLIP tends to assign a higher similarity score to "purple" than to "yellow" when filling the masked part, as it relates to the unmentioned eggplant. Therefore, the unreliable guidance of similarity scores may cause zero-shot image captioning methods to fail to describe the image. Similarly, the negative impact of concept association bias may cause zero-shot VQA methods to incorrectly answer questions such as "What color is the lemon in the image?"

The negative impact of concept association bias can be mitigated from two different perspectives. From the perspective of improving VLMs, it is feasible to extend current CLIP-like VLMs by adding additional architectures that perform deeper modality interactions. This strategy goes beyond the previous shallow interaction (*e.g.,* a simple inner product) and enables VLMs to learn more accurate correspondences between visual and textual elements. In addition, fine-tuning on augmented data helps CLIP-like VLMs capture fine-grained differences in scenes more effectively. For example, one feasible data augmentation strategy involves constructing hard negative samples by changing the original word order of the paired image caption. From the perspective of improving the integration of VLMs, it is recommended to avoid relying solely on the visual-textual similarity calculated by a VLM as a visual constraint for sentence completion, especially when the current sentence covers only a limited number of visual elements in the image or video. Instead, as proposed by Tewel et al. [66], the entire sentence should be considered as a whole and refined progressively using similarity. Alternatively, recent advances in zero-shot image captioning [65], [76]–[81] learn from unimodal data and decode sentences directly from the transformed image embeddings, thereby eliminating the reliance on similarity scores. By using the learned common vision-language space instead of similarity scores, these methods effectively mitigate bias and outperform similarity-based methods.

D. Confusion on Compositional Concepts

Since a VLM often struggles with understanding compositional concepts [177], [180], [181], [227], [228], relying on a VLM to bridge textual and visual modalities may lead to confusion in connecting compositional concepts in texts with corresponding visual elements. Specifically, the differences in detailed compositional concepts among similar images or videos are easily ignored by a VLM [177], [228]. On the other hand, when given text as queries to retrieve visual content, a VLM attends to a sparse set of the input text [227] where adjectives, numbers, and prepositions are often ignored. This makes it difficult for a VLM to accurately locate the specified visual content strictly based on the descriptive details given in the query, potentially leading to incorrect extractions of visual information. For example, consider a scenario where a VLM is tasked with open-vocabulary image classification on an image of a white and gray cat, likely to be a Ragdoll or Birman cat. To enrich the VLMs' comprehension of novel samples, many methods extract the semantic context of these cat categories from an LLM. This context typically contains detailed descriptions of visual appearances, such as "Ragdoll has large and striking blue eyes" and "Birman has slightly tilted almond-shaped eyes", which are the key evidence to distinguish between the two categories. However, when encoding these descriptions, VLMs often overlook the differences in adjectives. This drawback hinders the ability of VLMs to classify images correctly based on the generated context. In addition, CLIP tends to ignore fine-grained visual differences, such as color, count, and orientation, thereby encoding some different images into similar visual embeddings. For generative VLMs that take CLIP's visual encoder as their visual component, this ambiguity in visual encoding hinders them from correctly describing or answering questions about these visual attributes. For example, VLMs may not be able to determine whether a fruit in an image is cut or uncut, due to confusion over the subtle differences between visual embeddings of cut and uncut fruits. This drawback causes VLMs to provide inaccurate visual cues when solving complex visual questions and further misleads subsequent reasoning.

The negative impact of confusion on compositional concepts can be mitigated by improving the VLM and its integration. One feasible solution is to develop more powerful visual encoders for generative VLMs. By employing a mixture-of-features (MoF) strategy, we can combine CLIP's visual features with those from vision-only models trained in a self-supervised manner, thus improving the ability of CLIP to capture visual attributes. Moreover, some compositional concept-aware objectives [177], [227] have been proposed to fine-tune VLMs, where positive and negative samples are designed on the basis of the differences in compositional concepts in images. For the methods that integrate pre-trained models, it is more effective to design a unified system to comprehensively consider the results from multiple models and specialized detectors rather than relying on a single VLM. This modular architecture enables different models to address specific tasks, such as object detection or optical character recognition. Furthermore, some tasks that are challenging for a VLM, such as counting objects or determining their relative positions, can be effectively represented as code statements. For example, a question such as "how many seagulls are in the image" can be translated into a sequence of operations: detect the seagulls and then apply a code function to sum the detected regions. By integrating code execution with the output of a VLM, such tasks can be completed more accurately and efficiently, and thus benefit from the strengths of both VLM's visual understanding and precise computational coding.

VII. Conclusion

In this survey, we systematically review the methods of integrating pre-trained models to solve challenges in vision-language tasks. To ensure the systematic nature of this survey, we start with an introduction to the main challenges and then categorize the existing methods according to the challenges that they address, including data scarcity, escalating reasoning complexity, the generalization to novel samples, and task diversity. Specifically, we summarize the paradigms into illustrations of the pipeline to provide a more intuitive understanding of the various processes involved. In addition, we discuss the potential risks associated with the inherent limitations of pre-trained models. We hope that this survey can provide researchers with a comprehensive overview of the current status of the field and provide insights into future research directions.

References

[1] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, and A. Zadeh, "Multimodal research in vision and language: A review of current and emerging trends," *Information Fusion*, vol. 77, pp. 149–171, 2022.

[2] Y. Zhou and N. Shimada, "Vision+ language applications: A survey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 826–842.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[4] K. Lin, L. Li, C.-C. Lin, F. Ahmed, Z. Gan, Z. Liu, Y. Lu, and L. Wang, "Swinbert: End-to-end transformers with sparse attention for video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 949–17 958.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[6] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4631–4640.

[7] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3536–3545.

[8] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 638–10 647.

[9] S. Menon and C. Vondrick, "Visual classification via description from large language models," in *The Eleventh International Conference on Learning Representations*, 2022.

[10] A. Yan, Y. Wang, Y. Zhong, C. Dong, Z. He, Y. Lu, W. Y. Wang, J. Shang, and J. McAuley, "Learning concise and descriptive attributes for visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3090–3100.

[11] Y. Dai, H. Lang, K. Zeng, F. Huang, and Y. Li, "Exploring large language models for multi-modal out-of-distribution detection," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[12] Z. Ren, Y. Su, and X. Liu, "Chatgpt-powered hierarchical comparisons for image classification," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, J. Gao *et al.*, "Vision-language pre-training: Basics, recent advances, and future trends," *Foundations and Trends® in Computer Graphics and Vision*, vol. 14, no. 3–4, pp. 163–352, 2022.

[14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[15] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3441–3450.

[16] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *International conference on machine learning*. PMLR, 2015, pp. 1462–1471.

[17] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen, and Y.-W. Tai, "Memory-attended recurrent network for video captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8347–8356.

[18] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, "Heterogeneous memory enhanced multimodal attention model for video question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1999–2007.

[19] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "Mocogan: Decomposing motion and content for video generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1526–1535.

[20] L. Chen, C. Lu, S. Tang, J. Xiao, D. Zhang, C. Tan, and X. Li, "Rethinking the bottom-up framework for query-based video localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 551–10 558.

[21] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[22] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5831–5840.

[23] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6720–6731.

[24] N. Xie, F. Lai, D. Doran, and A. Kadav, "Visual entailment: A novel task for fine-grained image understanding," *arXiv preprint arXiv:1901.06706*, 2019.

[25] J. Hessel, J. D. Hwang, J. S. Park, R. Zellers, C. Bhagavatula, A. Rohrbach, K. Saenko, and Y. Choi, "The abduction of sherlock holmes: A dataset for visual abductive reasoning," in *European Conference on Computer Vision*. Springer, 2022, pp. 558–575.

[26] N. Bitton-Guetta, Y. Bitton, J. Hessel, L. Schmidt, Y. Elovici, G. Stanovsky, and R. Schwartz, "Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2616–2627.

[27] S. Zhong, Z. Huang, S. Gao, W. Wen, L. Lin, M. Zitnik, and P. Zhou, "Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[28] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[30] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/

[31] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[33] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[34] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei, "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 897–32 912, 2022.

[35] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu, "Filip: Fine-grained interactive language-image pre-training," *arXiv preprint arXiv:2111.07783*, 2021.

[36] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[37] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[38] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[39] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[40] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[41] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[42] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024.

[43] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *Machine Intelligence Research*, vol. 20, no. 4, pp. 447–482, 2023.

[44] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan, "Foundational models defining a new era in vision: A survey and outlook," *arXiv preprint arXiv:2307.13721*, 2023.

[45] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *arXiv preprint arXiv:2306.13549*, 2023.

[46] D. Zhang, Y. Yu, C. Li, J. Dong, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," *arXiv preprint arXiv:2401.13601*, 2024.

[47] Y. Wang, W. Chen, X. Han, X. Lin, H. Zhao, Y. Liu, B. Zhai, J. Yuan, Q. You, and H. Yang, "Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning," *arXiv preprint arXiv:2401.06805*, 2024.

[48] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu *et al.*, "Video understanding with large language models: A survey," *arXiv preprint arXiv:2312.17432*, 2023.

[49] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[50] Y. He, Z. Liu, J. Chen, Z. Tian, H. Liu, X. Chi, R. Liu, R. Yuan, Y. Xing, W. Wang *et al.*, "Llms meet multimodal generation and editing: A survey," *arXiv preprint arXiv:2405.19334*, 2024.

[51] A. Miyai, J. Yang, J. Zhang, Y. Ming, Y. Lin, Q. Yu, G. Irie, S. Joty, Y. Li, H. Li *et al.*, "Generalized out-of-distribution detection and beyond in vision language model era: A survey," *arXiv preprint arXiv:2407.21794*, 2024.

[52] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[53] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, "Ernie: Enhanced representation through knowledge integration," *arXiv preprint arXiv:1904.09223*, 2019.

[54] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[55] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon, "Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach," *arXiv preprint arXiv:1909.02201*, 2019.

[56] D. Mandal, P. Rao, and S. Biswas, "Semi-supervised cross-modal retrieval with label prediction," *IEEE Transactions on Multimedia*, vol. 22, no. 9, pp. 2345–2353, 2019.

[57] X. Duan, W. Huang, C. Gan, J. Wang, W. Zhu, and J. Huang, "Weakly supervised dense event captioning in videos," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[58] M. Luo, Y. Zeng, P. Banerjee, and C. Baral, "Weakly-supervised visual-retriever-reader for knowledge-based question answering," *arXiv preprint arXiv:2109.04014*, 2021.

[59] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4125–4134.

[60] I. Laina, C. Rupprecht, and N. Navab, "Towards unsupervised image captioning with shared multimodal embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7414–7424.

[61] Y. Tewel, Y. Shalev, I. Schwartz, and L. Wolf, "Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 918–17 928.

[62] Y. Su, T. Lan, Y. Liu, F. Liu, D. Yogatama, Y. Wang, L. Kong, and N. Collier, "Language models can see: Plugging visual controls in text generation," *arXiv preprint arXiv:2205.02655*, 2022.

[63] Z. Zeng, H. Zhang, R. Lu, D. Wang, B. Chen, and Z. Wang, "Conzic: Controllable zero-shot image captioning by sampling-based polishing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 465–23 476.

[64] Z. Zeng, Y. Xie, H. Zhang, C. Chen, Z. Wang, and B. Chen, "Meacap: Memory-augmented zero-shot image captioning," *arXiv preprint arXiv:2403.03715*, 2024.

[65] Y. Wang, W. Zhou, Z. Lu, and H. Li, "Text-only training for visual storytelling," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3686–3695.

[66] Y. Tewel, Y. Shalev, R. Nadler, I. Schwartz, and L. Wolf, "Zero-shot video captioning with evolving pseudo-tokens," *arXiv preprint arXiv:2207.11100*, 2022.

[67] Y. Jo, S. Lee, A. S. Lee, H. Lee, H. Oh, and M. Seo, "Zero-shot dense video captioning by jointly optimizing text and moment," *arXiv preprint arXiv:2307.02682*, 2023.

[68] H. Song, L. Dong, W.-N. Zhang, T. Liu, and F. Wei, "Clip models are few-shot learners: Empirical studies on vqa and visual entailment," *arXiv preprint arXiv:2203.07190*, 2022.

[69] Y. Yu, F. Zhan, R. Wu, J. Zhang, S. Lu, M. Cui, X. Xie, X.-S. Hua, and C. Miao, "Towards counterfactual image manipulation via clip," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3637–3645.

[70] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, and L. Wang, "An empirical study of gpt-3 for few-shot knowledge-based vqa," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3081–3089.

[71] A. M. H. Tiong, J. Li, B. Li, S. Savarese, and S. C. Hoi, "Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training," *arXiv preprint arXiv:2210.08773*, 2022.

[72] Z. Wang, M. Li, R. Xu, L. Zhou, J. Lei, X. Lin, S. Wang, Z. Yang, C. Zhu, D. Hoiem *et al.*, "Language models with image descriptors are strong few-shot video-language learners," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8483–8497, 2022.

[73] L. Hanu, A. L. Verő, and J. Thewlis, "Language as the medium: Multimodal video classification through text only," *arXiv preprint arXiv:2309.10783*, 2023.

[74] A. Bhattacharya, Y. K. Singla, B. Krishnamurthy, R. R. Shah, and C. Chen, "A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot," *arXiv preprint arXiv:2305.09758*, 2023.

[75] J. Pan, Z. Lin, Y. Ge, X. Zhu, R. Zhang, Y. Wang, Y. Qiao, and H. Li, "Retrieving-to-answer: Zero-shot video question answering with frozen large language models," *arXiv preprint arXiv:2306.11732*, 2023.

[76] D. Nukrai, R. Mokady, and A. Globerson, "Text-only training for image captioning using noise-injected clip," *arXiv preprint arXiv:2211.00575*, 2022.

[77] S. Gu, C. Clark, and A. Kembhavi, "I can't believe there's no images! learning visual tasks using only language data," *arXiv preprint arXiv:2211.09778*, 2022.

[78] W. Li, L. Zhu, L. Wen, and Y. Yang, "Decap: Decoding clip latents for zero-shot captioning via text-only training," *arXiv preprint arXiv:2303.03032*, 2023.

[79] J. Wang, M. Yan, and Y. Zhang, "From association to generation: Text-only captioning by unsupervised cross-modal mapping," *arXiv preprint arXiv:2304.13273*, 2023.

[80] J. Wang, Y. Zhang, M. Yan, J. Zhang, and J. Sang, "Zero-shot image captioning by anchor-augmented vision-language space alignment," *arXiv preprint arXiv:2211.07275*, 2022.

[81] J. Fei, T. Wang, J. Zhang, Z. He, C. Wang, and F. Zheng, "Transferable decoding with visual entities for zero-shot image captioning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3136–3146.

[82] D. Kim, J. Park, J. Lee, S. Park, and K. Sohn, "Language-free training for zero-shot video grounding," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2539–2548.

[83] Z. Wang, W. Liu, Q. He, X. Wu, and Z. Yi, "Clip-gen: Language-free training of a text-to-image generator with clip," *arXiv preprint arXiv:2203.00386*, 2022.

[84] F. Ma, Y. Zhou, F. Rao, Y. Zhang, and X. Sun, "Image captioning with multi-context synthetic data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4089–4097.

[85] Z. Liu, J. Liu, and F. Ma, "Improving cross-modal alignment with synthetic pairs for text-only image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3864–3872.

[86] L. Yang, X. Xu, B. Kang, Y. Shi, and H. Zhao, "Freemask: Synthetic images with dense annotations make stronger segmentation models," *arXiv preprint arXiv:2310.15160*, 2023.

[87] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun, "Towards language-free training for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 907–17 917.

[88] Z. Chen, Q. Zhou, Y. Shen, Y. Hong, H. Zhang, and C. Gan, "See, think, confirm: Interactive prompting between vision and language models for knowledge-based visual reasoning," *arXiv preprint arXiv:2301.05226*, 2023.

[89] K. Zhou, K. Lee, T. Misu, and X. E. Wang, "Vicor: Bridging visual understanding and commonsense reasoning with large language models," *arXiv preprint arXiv:2310.05872*, 2023.

[90] P. Wang, O. Golovneva, A. Aghajanyan, X. Ren, M. Chen, A. Celikyilmaz, and M. Fazel-Zarandi, "Domino: A dual-system for multi-step visual language reasoning," *arXiv preprint arXiv:2310.02804*, 2023.

[91] H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. Ayyubi, K.-W. Chang, and S.-F. Chang, "IdealGPT: Iteratively decomposing vision and language reasoning via large language models," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://openreview.net/forum?id=IvwcvJHLpc

[92] K. Yang, T. Shen, X. Tian, X. Geng, C. Tao, D. Tao, and T. Zhou, "Good questions help zero-shot image reasoning," *arXiv preprint arXiv:2312.01598*, 2023.

[93] J. Qi, Z. Xu, Y. Shen, M. Liu, D. Jin, Q. Wang, and L. Huang, "The art of SOCRATIC QUESTIONING: Recursive thinking with large language models," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. [Online]. Available: https://openreview.net/forum?id=jqbhtSDPz7

[94] Z. Wang, C. Chen, P. Li, and Y. Liu, "Filling the image information gap for VQA: Prompting large language models to proactively ask questions," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2874–2890. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.189

[95] H. Rajabzadeh, S. Wang, H. J. Kwon, and B. Liu, "Multimodal multi-hop question answering through a conversation between tools and efficiently finetuned large language models," *arXiv preprint arXiv:2309.08922*, 2023.

[96] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2507–2521, 2022.

[97] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[98] L. Wang, Y. Hu, J. He, X. Xu, N. Liu, H. Liu, and H. T. Shen, "T-sciq: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 162–19 170.

[99] D. Mondal, S. Modi, S. Panda, R. Singh, and G. S. Rao, "Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning," *arXiv preprint arXiv:2401.12863*, 2024.

[100] Y. Chen, K. Sikka, M. Cogswell, H. Ji, and A. Divakaran, "Measuring and improving chain-of-thought reasoning in vision-language models," *arXiv preprint arXiv:2309.04461*, 2023.

[101] W. Zhu, A. Agarwal, M. Joshi, R. Jia, J. Thomason, and K. Toutanova, "Efficient end-to-end visual document understanding with rationale distillation," *arXiv preprint arXiv:2311.09612*, 2023.

[102] C. Mitra, B. Huang, T. Darrell, and R. Herzig, "Compositional chain-of-thought prompting for large multimodal models," *arXiv preprint arXiv:2311.17076*, 2023.

[103] D. Zhang, J. Yang, H. Lyu, Z. Jin, Y. Yao, M. Chen, and J. Luo, "Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs," *arXiv preprint arXiv:2401.02582*, 2024.

[104] F. Meng, H. Yang, Y. Wang, and M. Zhang, "Chain of images for intuitively reasoning," *arXiv preprint arXiv:2311.09241*, 2023.

[105] D. Rose, V. Himakunthala, A. Ouyang, R. He, A. Mei, Y. Lu, M. Saxon, C. Sonar, D. Mirza, and W. Y. Wang, "Visual chain of thought: Bridging logical gaps with multimodal infillings," *arXiv preprint arXiv:2305.02317*, 2023.

[106] V. Himakunthala, A. Ouyang, D. Rose, R. He, A. Mei, Y. Lu, C. Sonar, M. Saxon, and W. Y. Wang, "Let's think frame by frame: Evaluating video chain of thought with video infilling and prediction," *arXiv preprint arXiv:2305.13903*, 2023.

[107] L. Li, J. Xiao, G. Chen, J. Shao, Y. Zhuang, and L. Chen, "Zero-shot visual relation detection via composite visual cues from large language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[108] C. Jia, M. Luo, X. Chang, Z. Dang, M. Han, M. Wang, G. Dai, S. Dang, and J. Wang, "Generating action-conditioned prompts for open-vocabulary video action recognition," *arXiv preprint arXiv:2312.02226*, 2023.

[109] A. Yousaf, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Videoprompter: an ensemble of foundational models for zero-shot video understanding," *arXiv preprint arXiv:2310.15324*, 2023.

[110] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *International Conference on Learning Representations*, 2022.

[111] S. Wu, W. Zhang, S. Jin, W. Liu, and C. C. Loy, "Aligning bag of regions for open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 254–15 264.

[112] L. Wang, Y. Liu, P. Du, Z. Ding, Y. Liao, Q. Qi, B. Chen, and S. Liu, "Object-aware distillation pyramid for open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 186–11 196.

[113] M. Gao, C. Xing, J. C. Niebles, J. Li, R. Xu, W. Liu, and C. Xiong, "Open vocabulary object detection with pseudo bounding-box labels," in *European Conference on Computer Vision*. Springer, 2022, pp. 266–282.

[114] W. Kuo, Y. Cui, X. Gu, A. Piergiovanni, and A. Angelova, "Open-vocabulary object detection upon frozen vision and language models," in *The Eleventh International Conference on Learning Representations*, 2023.

[115] N. Liao, Y. Liu, L. Xiaobo, C. Lei, G. Wang, X.-S. Hua, and J. Yan, "Cohoz: Contrastive multimodal prompt tuning for hierarchical open-set zero-shot recognition," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3262–3271.

[116] N. Liao, X. Zhang, M. Cao, J. Yan, and Q. Tian, "M-tuning: Prompt tuning with mitigated label bias in open-set scenarios," *arXiv preprint arXiv:2303.05122*, 2023.

[117] H. Qu, X. Hui, Y. Cai, and J. Liu, "Lmc: Large model collaboration with cross-assessment for training-free open-set object recognition," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[118] Z. Qian, X. Wang, X. Duan, P. Qin, Y. Li, and W. Zhu, "Decouple before interact: Multi-modal prompt learning for continual visual question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2953–2962.

[119] C. Chen, J. Zhu, X. Luo, H. Shen, L. Gao, and J. Song, "Coin: A benchmark of continual instruction tuning for multimodel large language model," *arXiv preprint arXiv:2403.08350*, 2024.

[120] J. Zheng, Q. Ma, Z. Liu, B. Wu, and H. Feng, "Beyond anti-forgetting: Multimodal continual instruction tuning with positive forward transfer," *arXiv preprint arXiv:2401.09181*, 2024.

[121] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang, "Mm-react: Prompting chatgpt for multimodal reasoning and action," *arXiv preprint arXiv:2303.11381*, 2023.

[122] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang, and N. Duan, "Visual chatgpt: Talking, drawing and editing with visual foundation models," *arXiv preprint arXiv:2303.04671*, 2023.

[123] P. Lu, B. Peng, H. Cheng, M. Galley, K.-W. Chang, Y. N. Wu, S.-C. Zhu, and J. Gao, "Chameleon: Plug-and-play compositional reasoning with large language models," *ArXiv*, vol. abs/2304.09842, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:258212542

[124] K. Lin, F. Ahmed, L. Li, C.-C. Lin, E. Azarnasab, Z. Yang, J. Wang, L. Liang, Z. Liu, Y. Lu *et al.*, "Mm-vid: Advancing video understanding with gpt-4v (ision)," *arXiv preprint arXiv:2310.19773*, 2023.

[125] T. Gupta and A. Kembhavi, "Visual programming: Compositional visual reasoning without training," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[126] D. Sur'is, S. Menon, and C. Vondrick, "Vipergpt: Visual inference via python execution for reasoning," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[127] R. Choudhury, K. Niinuma, K. M. Kitani, and L. A. Jeni, "Zero-shot video question answering with procedural programs," *ArXiv*, vol. abs/2312.00937, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265608945

[128] A. Stani'c, S. Caelles, and M. Tschannen, "Towards truly zero-shot compositional visual reasoning with llms as programmers," *ArXiv*, vol. abs/2401.01974, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266755924

[129] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[130] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. 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*. Springer, 2014, pp. 740–755.

[131] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2641–2649.

[132] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, and W. Y. Wang, "Vatex: A large-scale, high-quality multilingual dataset for video-and-language research," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 4581–4591.

[133] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5288–5296.

[134] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3202–3212.

[135] B. Xie, S. Zhang, Z. Zhou, B. Li, Y. Zhang, J. Hessel, J. Yang, and Z. Liu, "Funqa: Towards surprising video comprehension," *arXiv preprint arXiv:2306.14899*, 2023.

[136] C. Liang, W. Wang, T. Zhou, and Y. Yang, "Visual abductive reasoning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 15 565–15 575.

[137] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 2016, pp. 451–466.

[138] M. Li and M.-F. Moens, "Dynamic key-value memory enhanced multi-step graph reasoning for knowledge-based visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 10 983–10 992.

[139] X. Song, Y. Shi, X. Chen, and Y. Han, "Explore multi-step reasoning in video question answering," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 239–247.

[140] S. Wang, Y. Zhu, H. Liu, Z. Zheng, C. Chen *et al.*, "Knowledge editing for large language models: A survey," *arXiv preprint arXiv:2310.16218*, 2023.

[141] D. M. Vo, H. Chen, A. Sugimoto, and H. Nakayama, "Noc-rek: novel object captioning with retrieved vocabulary from external knowledge," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 17 979–17 987.

[142] M. Narasimhan and A. G. Schwing, "Straight to the facts: Learning knowledge base retrieval for factual visual question answering," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 451–468.

[143] Q. Wu, P. Wang, C. Shen, A. Dick, and A. Van Den Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4622–4630.

[144] G. Li, H. Su, and W. Zhu, "Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks," *arXiv preprint arXiv:1712.00733*, 2017.

[145] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, "Fvqa: Fact-based visual question answering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2413–2427, 2017.

[146] Z. Chen, J. Chen, Y. Geng, J. Z. Pan, Z. Yuan, and H. Chen, "Zero-shot visual question answering using knowledge graph," in *The Semantic Web–ISWC 2021: 20th International Semantic Web Conference, ISWC 2021, Virtual Event, October 24–28, 2021, Proceedings 20*. Springer, 2021, pp. 146–162.

[147] G. A. Miller, "Wordnet: a lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[148] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[149] R. Hu and A. Singh, "Unit: Multimodal multitask learning with a unified transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 1439–1449.

[150] D.-K. Nguyen and T. Okatani, "Multi-task learning of hierarchical vision-language representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 492–10 501.

[151] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 437–10 446.

[152] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations*, 2021.

[153] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu *et al.*, "Instruction tuning for large language models: A survey," *arXiv preprint arXiv:2308.10792*, 2023.

[154] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27 730–27 744, 2022.

[155] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," *Advances in neural information processing systems*, vol. 30, 2017.

[156] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *Advances in neural information processing systems*, vol. 34, pp. 9694–9705, 2021.

[157] S. Lavoie, P. Kirichenko, M. Ibrahim, M. Assran, A. G. Wildon, A. Courville, and N. Ballas, "Modeling caption diversity in contrastive vision-language pretraining," *arXiv preprint arXiv:2405.00740*, 2024.

[158] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," *arXiv preprint arXiv:2109.14084*, 2021.

[159] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," *arXiv preprint arXiv:2109.08472*, 2021.

[160] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[161] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[162] Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang *et al.*, "Aligning large multimodal models with factually augmented rlhf," *arXiv preprint arXiv:2309.14525*, 2023.

[163] K. Yang, J. Tao, J. Lyu, C. Ge, J. Chen, W. Shen, X. Zhu, and X. Li, "Using human feedback to fine-tune diffusion models without any reward model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8941–8951.

[164] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[165] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.

[166] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1728–1738.

[167] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[168] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[169] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," *arXiv preprint arXiv:2306.05425*, 2023.

[170] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *arXiv preprint arXiv:2305.03726*, 2023.

[171] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "Nocaps: Novel object captioning at scale," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8948–8957.

[172] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[173] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3608–3617.

[174] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 69–85.

[175] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," *arXiv preprint arXiv:2307.16125*, 2023.

[176] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu *et al.*, "Mmbench: Is your multi-modal model an all-around player?" in *European Conference on Computer Vision*. Springer, 2025, pp. 216–233.

[177] R. Paiss, A. Ephrat, O. Tov, S. Zada, I. Mosseri, M. Irani, and T. Dekel, "Teaching clip to count to ten," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3170–3180.

[178] S. Tong, Z. Liu, Y. Zhai, Y. Ma, Y. LeCun, and S. Xie, "Eyes wide shut? exploring the visual shortcomings of multimodal llms," *arXiv preprint arXiv:2401.06209*, 2024.

[179] S. Kumar, S. Ghosh, S. Sakshi, U. Tyagi, and D. Manocha, "Do vision-language models understand compound nouns?" *arXiv preprint arXiv:2404.00419*, 2024.

[180] A. Ray, F. Radenovic, A. Dubey, B. Plummer, R. Krishna, and K. Saenko, "cola: A benchmark for compositional text-to-image retrieval," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[181] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna, "Crepe: Can vision-language foundation models reason compositionally?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 910–10 921.

[182] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" in *The Eleventh International Conference on Learning Representations*, 2022.

[183] C.-Y. Hsieh, J. Zhang, Z. Ma, A. Kembhavi, and R. Krishna, "Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality," *Advances in neural information processing systems*, vol. 36, 2024.

[184] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoob *et al.*, "Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 375–14 385.

[185] Y. Fan, J. Gu, K. Zhou, Q. Yan, S. Jiang, C.-C. Kuo, Y. Zhao, X. Guan, and X. Wang, "Muffin or chihuahua? challenging multimodal large language models with multipanel vqa," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 6845–6863.

[186] A. Wang, B. Wu, S. Chen, Z. Chen, H. Guan, W.-N. Lee, L. E. Li, and C. Gan, "Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 384–13 394.

[187] L. Zhang, X. Zhai, Z. Zhao, Y. Zong, X. Wen, and B. Zhao, "What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 853–21 862.

[188] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[189] X. He, "Parallel refinements for lexically constrained text generation with bart," *arXiv preprint arXiv:2109.12487*, 2021.

[190] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[191] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[192] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," *arXiv preprint arXiv:2111.09734*, 2021.

[193] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[194] D. Guo, Y. Wang, P. Song, and M. Wang, "Recurrent relational memory network for unsupervised image captioning," *arXiv preprint arXiv:2006.13611*, 2020.

[195] Y. Qi, W. Zhao, and X. Wu, "Relational distant supervision for image captioning without image-text pairs," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4524–4532.

[196] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[197] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.

[198] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[199] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 382–398.

[200] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[201] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-okvqa: A benchmark for visual question answering using world knowledge," in *European Conference on Computer Vision*. Springer, 2022, pp. 146–162.

[202] A. Gupta, P. Dollar, and R. Girshick, "Lvis: A dataset for large vocabulary instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5356–5364.

[203] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang, "Open-vocabulary object detection using captions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 393–14 402.

[204] Y. Shi, X. Wu, H. Lin, and J. Luo, "Commonsense knowledge prompting for few-shot action recognition in videos," *IEEE Transactions on Multimedia*, 2024.

[205] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[206] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[207] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 2014, pp. 446–461.

[208] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 6, pp. 1452–1464, 2017.

[209] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[210] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[211] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[212] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *NeurIPS*, 2021.

[213] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran, "Zero-shot object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 384–400.

[214] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[215] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman *et al.*, "Evaluating large language models trained on code," *arXiv preprint arXiv:2107.03374*, 2021.

[216] F. Liu, T. Guan, Z. Li, L. Chen, Y. Yacoob, D. Manocha, and T. Zhou, "Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models," *arXiv preprint arXiv:2310.14566*, 2023.

[217] W. Dai, Z. Liu, Z. Ji, D. Su, and P. Fung, "Plausible may not be faithful: Probing object hallucination in vision-language pre-training," *arXiv preprint arXiv:2210.07688*, 2022.

[218] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen *et al.*, "Siren's song in the ai ocean: a survey on hallucination in large language models," *arXiv preprint arXiv:2309.01219*, 2023.

[219] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, "How language model hallucinations can snowball," *arXiv preprint arXiv:2305.13534*, 2023.

[220] T. Wu, L. Luo, Y.-F. Li, S. Pan, T.-T. Vu, and G. Haffari, "Continual learning for large language models: A survey," *arXiv preprint arXiv:2402.01364*, 2024.

[221] S. Yu, C. Tang, B. Xu, J. Cui, J. Ran, Y. Yan, Z. Liu, S. Wang, X. Han, Z. Liu *et al.*, "Visrag: Vision-based retrieval-augmented generation on multi-modality documents," *arXiv preprint arXiv:2410.10594*, 2024.

[222] S. Sharifymoghaddam, S. Upadhyay, W. Chen, and J. Lin, "Unirag: Universal retrieval augmentation for multi-modal large language models," *arXiv preprint arXiv:2405.10311*, 2024.

[223] D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Wiki-llava: Hierarchical retrieval-augmented generation for multimodal llms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1818–1826.

[224] J. Qiu, A. Madotto, Z. Lin, P. A. Crook, Y. E. Xu, X. L. Dong, C. Faloutsos, L. Li, B. Damavandi, and S. Moon, "Snapntell: Enhancing entity-centric visual question answering with retrieval augmented multimodal llm," *arXiv preprint arXiv:2403.04735*, 2024.

[225] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, "Winoground: Probing vision and language models for visio-linguistic compositionality," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5238–5248.

[226] Y. Tang, Y. Yamada, Y. Zhang, and I. Yildirim, "When are lemons purple? the concept association bias of vision-language models," in

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 14 333–14 348.

[227] R. Paiss, H. Chefer, and L. Wolf, "No token left behind: Explainability-aided image classification and generation," in *European Conference on Computer Vision*. Springer, 2022, pp. 334–350.

[228] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8253–8280.

BIOGRAPHY SECTION

Yayun Qi received a B.S. degree from Beijing Institute of Technology (BIT), Beijing, China, in 2020. She will pursue a Ph.D. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology. Her research interests include computer vision, and video analysis and understanding.

Hongxi Li received a B.S. degree from Beijing Wuzi University (BWU), Beijing, China, in 2022. He will pursue an M.S. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology. His research interests include computer vision and machine learning.

Yiqi Song received a B.S. degree from Beijing Institute of Technology (BIT), Beijing, China, in 2024. He will pursue a Ph.D. degree at the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology. His research interests include computer vision and video analysis and understanding.

Xinxiao Wu (Member, IEEE) is a Full Professor with the School of Computer Science, Beijing Institute of Technology. She received a B.S. degree in computer science from the Nanjing University of Information Science and Technology in 2005, and a Ph.D. degree in computer science from the Beijing Institute of Technology in 2010. From 2010 to 2011, she was a Postdoctoral Research Fellow with Nanyang Technological University, Singapore. Her research interests include machine learning, computer vision, and video analysis and understanding.

Jiebo Luo (Fellow, IEEE) is the Albert Arendt Hopeman Professor of Engineering and Professor of Computer Science at the University of Rochester which he joined in 2011 after a prolific career of fifteen years at Kodak Research Laboratories. He has authored nearly 600 technical papers and holds over 90 U.S. patents. His research interests include computer vision, NLP, machine learning, data mining, computational social science, and digital health. He has been involved in numerous technical conferences, including serving as program co-chair of ACM Multimedia 2010, IEEE CVPR 2012, ACM ICMR 2016, and IEEE ICIP 2017, as well as general co-chair of ACM Multimedia 2018 and IEEE ICME 2024. He has served on the editorial boards of the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), IEEE Transactions on Multimedia (TMM), IEEE Transactions on Circuits and Systems for Video Technology (TCSVT), IEEE Transactions on Big Data (TBD), ACM Transactions on Intelligent Systems and Technology (TIST), Pattern Recognition, Knowledge and Information Systems (KAIS), Machine Vision and Applications, and Intelligent Medicine. He was the Editor-in-Chief of the IEEE Transactions on Multimedia (2020-2022). Professor Luo is also a Fellow of NAI, ACM, AAAI, SPIE, and IAPR.