

Generalizing Vision-Language Models to Novel Domains: A Comprehensive Survey

Xinyao Li, Jingjing Li, Fengling Li, Lei Zhu, Yang Yang, and Heng Tao Shen, *Fellow, IEEE*

Abstract—Recently, vision-language pretraining has emerged as a transformative technique that integrates the strengths of both visual and textual modalities, resulting in powerful vision-language models (VLMs). Leveraging web-scale pretraining data, these models exhibit strong zero-shot capabilities. However, their performance often deteriorates when confronted with domain-specific or specialized generalization tasks. To address this, a growing body of research focuses on transferring or generalizing the rich knowledge embedded in VLMs to various downstream applications. This survey aims to comprehensively summarize the generalization settings, methodologies, benchmarking and results in VLM literatures. Delving into the typical VLM structures, current literatures are categorized into *prompt-based*, *parameter-based* and *feature-based* methods according to the transferred modules. The differences and characteristics in each category are further summarized and discussed by revisiting the typical *transfer learning (TL) settings*, providing novel interpretations for TL in the era of VLMs. Popular benchmarks for VLM generalization are further introduced with thorough performance comparisons among the reviewed methods. Following the advances in large-scale generalizable pretraining, this survey also discusses the relations and differences between VLMs and up-to-date multimodal large language models (MLLM), e.g., DeepSeek-VL. By systematically reviewing the surging literatures in vision-language research from a novel and practical generalization prospective, this survey contributes to a clear landscape of current and future multimodal researches.

Index Terms—Vision-language models, transfer learning, prompt tuning, robust fine-tuning, domain generalization, test-time adaptation, unsupervised domain adaptation, multimodal large language model

1 INTRODUCTION

DEEP neural networks have achieved remarkable success across a wide range of practical applications. Taking vision models as an example, the progression from AlexNet [1] to ResNet [2] and Vision Transformers [3] has significantly advanced both model scale and representational power. However, training such large-scale models effectively demands substantial labeled data and computational resources. To address this, the concept of *foundation models* has emerged—models pretrained on massive datasets to acquire general-purpose knowledge, which can then be transferred to various downstream tasks [4]. For instance, the ResNet family pretrained on ImageNet [5] has served as a cornerstone in a wide array of vision tasks, including classification [2] and object recognition [6]. Parallel advancements have occurred in natural language processing, with the development of influential models such as the Transformer [7], BERT [8], GPT-2 [9], and GPT-3 [10]. While these models excel in their respective single-modality domains, they inherently lack the ability to perceive and reason over multimodal information.

As shown in Fig. 1, the emergence of the contrastive language-image pretraining paradigm [11] has fundamentally reshaped the landscape of vision-language learning. Leveraging 400M web-crawled image-text pairs, Radford *et al.* introduced the first vision-language foundation model, CLIP [11], which learns by pulling together semantically aligned image-text pairs and pushing apart mismatched

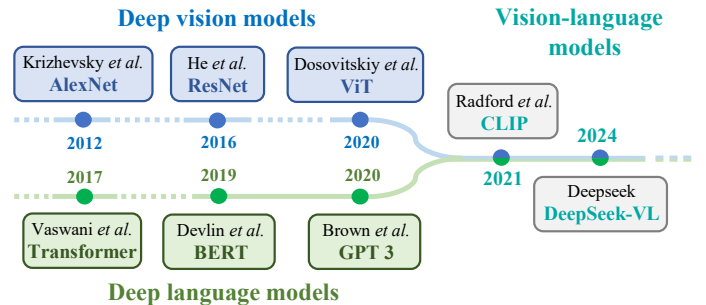


Fig. 1. Examples of vision, language and vision-language deep models.

ones. This large-scale contrastive pretraining equips CLIP with impressive zero-shot capabilities across a variety of tasks, including image classification [11], object detection [12], and video-text retrieval [13]. Subsequent works have further extended the capabilities of VLMs by enlarging and denoising pretraining datasets [14], [15], [16], exploring diverse pretraining strategies [17], [18], and incorporating multilingual data [19], [20], [21].

Despite their outstanding performance on general tasks, generalizing their pretrained knowledge to more specialized downstream tasks is non-trivial. Without proper transfer, pretrained VLMs struggle in handling out-of-distribution (OOD) data like satellite images [22] and diverse fine-grained images [23], [24]. While the typical pretraining - fine-tuning paradigm for knowledge transfer is still applicable, appropriate full-tuning for VLMs has proved tricky. Direct tuning with insufficient target data may disturb the aligned vision-language representations and lead to degraded performances [25], [26], [27]. Therefore, how to

- Xinyao Li, Jingjing Li, Yang Yang and Heng Tao Shen are with University of Electronic Science and Technology of China, Chengdu 610054, China.
- Fengling Li is with University of Technology Sydney. Lei Zhu is with Tongji University, Shanghai 200070, China.

Manuscript received ; revised .

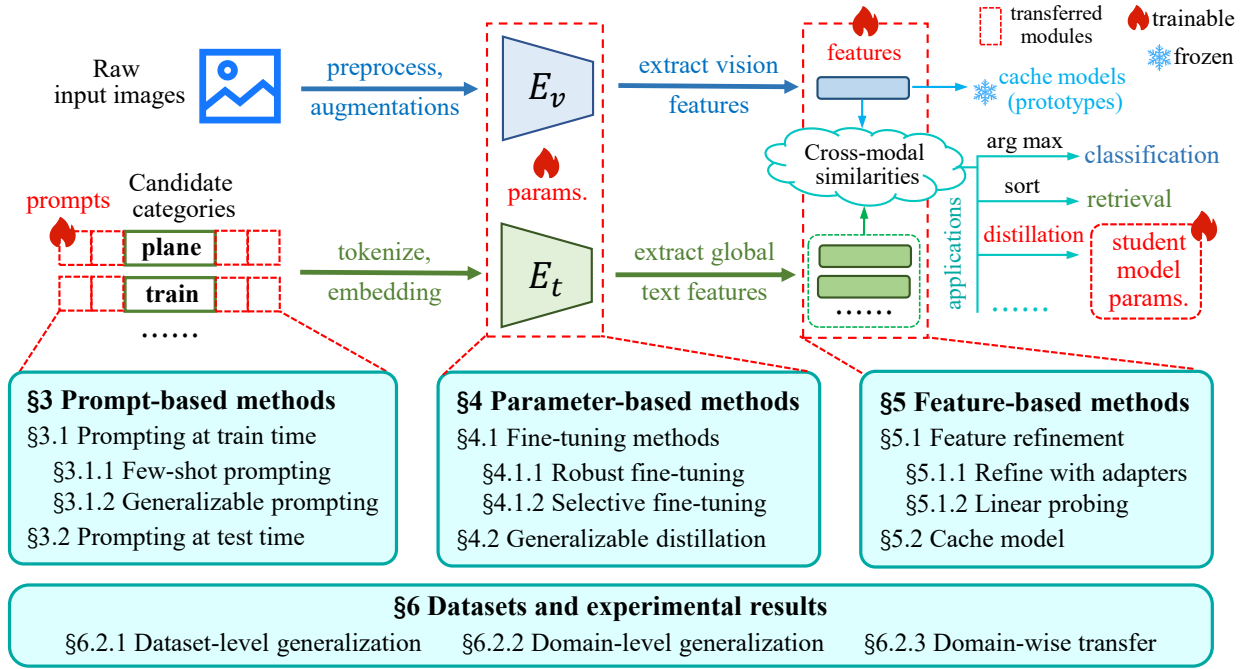


Fig. 2. This survey is organized based on the typical and most investigated dual-branch VLM structure, where generalization methods are categorized according to the transferred modules in the VLM.

elegantly generalize the pretrained knowledge in VLMs to downstream tasks with minimum computation and annotation expenses has become a research hotspot. Noticing VLMs' multimodal nature, researchers have dedicated to adapt well-developed single-modality transfer methods, e.g., prompt tuning [28], adapters [29], distillation [30], to VLMs [26], [31], [32], [33]. Powered by vast general knowledge, VLMs are becoming task-agnostic solvers by setting new strong baselines on various transfer learning (TL) scenarios with their strong zero-shot abilities, e.g., unsupervised domain adaptation (UDA) [34], [35], [36], domain generalization (DG) [37], [38], [39], test-time adaptation (TTA) [40], [41], [42], etc. With such universal success, we ask: *How is knowledge transfer different in the era of VLMs?*

In response to the question, this work conducts a comprehensive literature review on the generalization of pre-trained VLMs. Existing surveys [43], [44], [45] mainly focus on the *pretraining* of VLMs, including model structures, pre-training objectives and datasets. While survey [43] mentions TL for VLMs, only a part of few-shot adaptation methods are included, and the differences in transfer setting are not considered. This paper presents the first survey that concentrates on the knowledge *transfer and generalization* of VLMs. As shown in Fig. 2, our analysis base on the most researched dual-branch VLM, e.g., CLIP [11]. We identify and categorize the key components in the VLM and summarize corresponding transfer methods. (1) Prompt-based methods only tunes the textual prompt embeddings to control the VLM behavior [31], [32], [40]. (2) Parameter-based methods either strategically update the pretrained parameters [46], [47], [48], or learn new parameters by distillation [33], [38], [39]. (3) Feature-based methods either update the extracted features with additional modules [26], [35], or build trainable cache modules with feature prototypes [27], [41], [49]. From the view of transfer learning and heavily-studied

transfer settings [4], [50], [51], we revisit these VLM-based methods and analyze the nuances derived from setting characteristics. Finally, we introduce mainstream benchmarks for different transfer settings and provide a rigorous performance comparison among VLM-based methods. As the strong abilities of VLMs bring significant boosts compared with single-modality methods, a benchmarking overview for VLM-based transfer methods is urgently needed.

The revolutionary breakthroughs in modern large language models (LLM) [52], [53], [54], [55] have further pushed the limits of VLMs. By connecting language-aligned vision encoders (e.g., CLIP's vision encoder) with LLMs and train with vast multimodal instruction-following data, the resultant vision-LLM, termed multimodal large language model (MLLM) in this survey, exhibits strong and vast generalization to a wider range of general visual-language tasks [18], [56], [57], [58], including video understanding, visual question answering, image captioning, segmentation and recognition, etc. The development of general-purpose MLLMs is a fast-evolving field involving multiple top AI institutions, with new-generation products being released monthly. As MLLMs are another means of important vision-language model generalizable to various tasks, this survey includes the most typical and most up-to-date MLLMs to show practical and scaled applications of vision-language research. This survey summarizes general frameworks for building MLLMs, and discusses the model types, used training data, optimization objectives, and features of different MLLMs, aiming to provide readers with a comprehensive understanding of frontier advances in the field of vision-language research. An overview of recent advances in general vision-language research is presented in Fig. 3. The contributions of this survey are summarized as follows:

- 1) This survey systematically reviews the research progress of knowledge transfer and generalize of

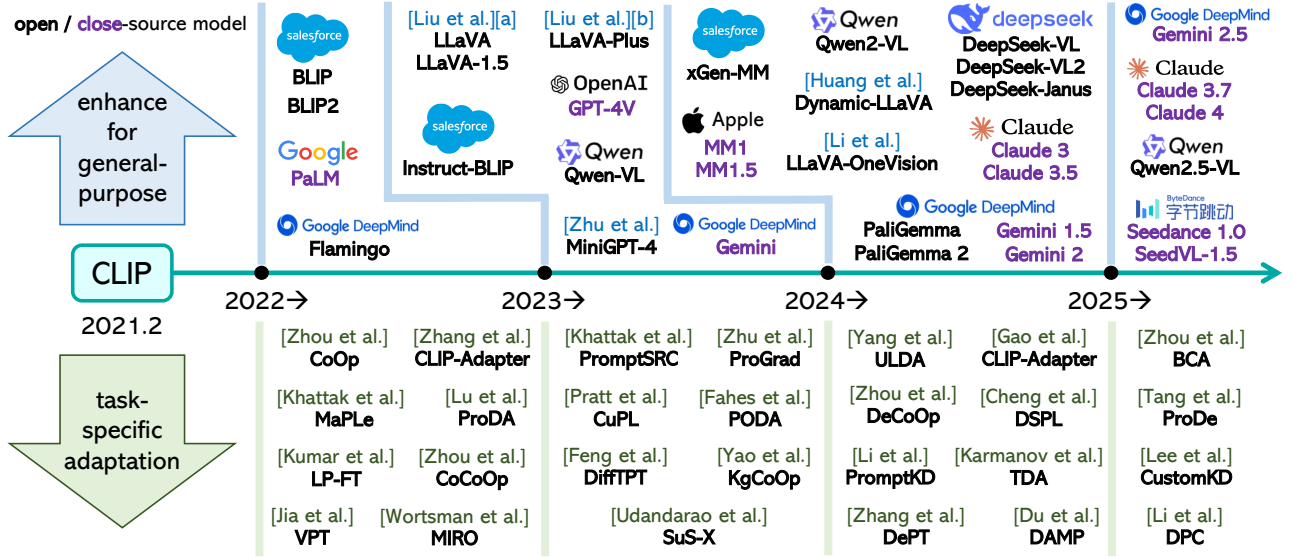


Fig. 3. An overview of recent advances in vision-language systems, including adapting VLMs (lower part) and building general MLLMs (upper part).

VLMs, covering popular transfer settings including unsupervised domain adaptation, domain generalization, few-shot adaptation, test-time adaptation, etc. To the best of our knowledge, this paper is the first to investigate VLM-based methods from the perspective of model generalization.

- 2) This survey identifies three key components within VLM architectures that are critical for knowledge transfer: prompt-based, parameter-based, and feature-based methods. A more fine-grained analysis is conducted to elucidate the specific techniques and adaptations employed within each transfer setting.
- 3) This survey collects mainstream benchmarks for evaluating VLM-based generalization methods. A detailed performance comparison considering the generalization settings, model backbones and method design is provided, contributing a fair and comprehensive evaluation of current research advances.
- 4) This survey includes VLMs enhanced and generalized by modern LLMs, termed MLLMs. A comprehensive summarization of the typical and cutting-edge MLLMs regarding their structures, used unimodal models, generalization to vast vision-language tasks, training data and objectives, is provided.
- 5) This survey analyzes the key challenges that persist in current vision-language research and discusses potential directions for future exploration.

The rest of this survey is organized as follows. Section 2 introduces the preliminaries of the studied VLMs, as well as the typical transfer learning settings involved in this survey. Section 3 introduces prompt-based methods divided by the prompting scenarios: Train-time prompting (Section 3.1) and test-time prompting (Section 3.2). Section 4 introduces parameter-based methods, including the robust fine-tuning of VLMs' parameters (Section 4.1) and distilling the pretrained knowledge to student model parameters (Section 4.2). Section 5 introduces feature-level knowledge transfer, where the features are updated with learnable adapters (Section 5.1) or compressed into train-free key-value caches (Section 5.2). Section 6 describes benchmarks

and evaluation results of the reviewed methods. Section 7 introduces how pretrained VLMs, e.g., CLIP, are generalized and enhanced with modern LLMs. Section 8 summarizes current research advances and discuss promising future directions in the vision-language field.

2 BACKGROUND

2.1 Preliminaries of Vision-Language Models

Aligning with current research focus, this survey mainly discusses the *two-tower* VLM as categorized in [43], exemplified by CLIP [11]. We start by introducing the pretraining objective, model structure and inference procedure of VLMs.

Pretraining objective. As shown in Fig. 4, VLMs are trained with abundant image-text pairs $\{(x_i, t_i)\}_{i=1}^n$, where x_i are input images and t_i are texts describing the image, e.g., *A photo of a [object in the image]* [11]. The VLM combines a vision and language encoder E_v and E_t to handle inputs from both modalities, outputting corresponding vision and text representations: $v_i = E_v(x_i)$, $\mu_i = E_t(t_i)$. The vision encoders are from the ResNet family [2] or the vision transformer family [3]. Employing contrastive learning [59], representations of matched vision and text pairs are pulled closer, while other pairs are pushed away. Given a batch of B image-text pairs, the pretraining loss is defined as:

$$\mathcal{L}_{\text{I2T}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(v_i \cdot \mu_i) / \tau}{\sum_{j=1}^B \exp(v_i \cdot \mu_j) / \tau}, \quad (1)$$

$$\mathcal{L}_{\text{T2I}} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\mu_i \cdot v_i) / \tau}{\sum_{j=1}^B \exp(\mu_i \cdot v_j) / \tau}, \quad (2)$$

$$\mathcal{L}_{\text{pre}} = \frac{1}{2} (\mathcal{L}_{\text{I2T}} + \mathcal{L}_{\text{T2I}}), \quad (3)$$

where τ is a learnable temperature parameter.

Inference. With vision and text encoders trained by Eq. (3), VLMs can make zero-shot inferences. Take the most general image classification as an example, the inference procedure is shown in Fig. 2. Given a set of C candidate category names $\{\text{class}_c\}_{c=1}^C$, we can construct naive prompt inputs t_c as in [11]: A photo of a class_c . The text encoder

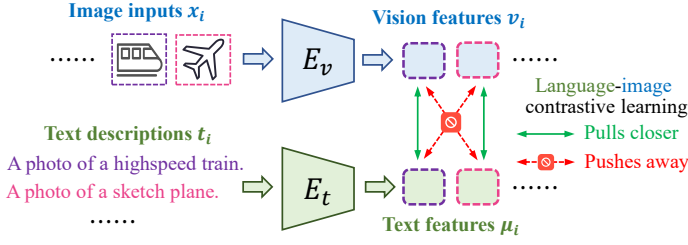


Fig. 4. Illustration of the language-image contrastive learning procedure. E_v , E_t are vision and language encoders, respectively.

then generates text representations $\{\mu_c\}_{c=1}^C$ corresponding to the class names. Given an image x and its feature representation $v = E_v(x)$, the probability that image x belongs to class c is obtained by:

$$P_c(x) = \frac{\exp(\cos \langle v, \mu_c \rangle / \tau)}{\sum_{i=1}^C \exp(\cos \langle v, \mu_i \rangle / \tau)}, \quad (4)$$

where $\cos \langle \cdot, \cdot \rangle$ computes the cosine similarities. Note that based on the vision-language representation similarities, VLMs can be applied to various downstream tasks including object detection [12], text retrieval [13], etc.

Transferable modules. As depicted in Fig. 2, this survey concludes the following three key VLM modules for knowledge transfer and generalization.

(1) *Prompts* are the text inputs t_i for the language encoder E_t , describing the details of an image. However, the pretraining texts usually include more visual details [11] while the naive prompts for downstream applications only include class names (Eq. (4)), resulting in poor performance on more specialized tasks [31]. Therefore, Zhou *et al.* [31] propose to learn the input prompts. As shown in Fig. 2, the prefix and suffix of prompts are learnable:

$$t_i = [U]_0, \dots, [U]_k[\text{class}]_i[U]_{k+1}, \dots, [U]_m, \quad (5)$$

where $[U]_k$ are learnable prompt embeddings that replace the original texts. The prompts can either be appended to the start and end of a sentence. In addition to the prompts in Eq. (5), there are various prompting paradigm and optimization objectives, which are elaborated in Section 3.

(2) *Parameters* in encoders E_v, E_t are pretrained with Eq. (3) and frozen for zero-shot predictions. However, with proper regularization the parameters can be updated for generalization to downstream tasks. A more practical scenario is to distill the VLM knowledge to smaller student models with parameters θ_{stu} for efficient deployments. Typical knowledge distillation optimizes student parameters by:

$$\theta_{\text{stu}} = \arg \min_{\theta_{\text{stu}}} \sum_x \mathcal{L}_d([P_0(x), \dots, P_C(x)], \text{stu}(x)), \quad (6)$$

where $\text{stu}(x)$ is the student output, and \mathcal{L}_d is distance measurement, e.g., KL divergence. Section 4 introduces more variants of fine-tuning and distillation methods.

(3) *Features* refer to the high-dimension vectors μ_i, v_i generated by the encoders. Inspired by the adapters for parameter-efficient fine-tuning [29], the multimodal features can be efficiently refined with additional adapter modules. Take vision features as an example, the adapter Φ works by:

$$v_i^* = v_i + \alpha \cdot \Phi(v_i), \quad (7)$$

where v_i^* is the updated feature, α is a hyperparameter. As a more efficient alternative, key-value cache models can be built with the extracted features and corresponding model outputs, encompassing the model knowledge. Detailed discussions are in Section 5.

2.2 Transfer Learning

Transfer learning (TL) has been attracting research interests since more than a decade ago [60]. TL enables deep models to share and reuse knowledge for similar tasks, enhancing annotation and computation efficiency in deep learning. This survey focuses on the following most frequently investigated TL settings in VLM-based methods.

(1) **Unsupervised domain adaptation** [61], [62], [63], [64] (UDA) aims to transfer knowledge from a labeled source domain to an unlabeled target domain with shifted distribution. Specifically, the source domain is denoted as $D_S = \{(x_i^S, y_i^S)\}_{i=1}^{n_S}$, and the target domain contains only unlabeled data $D_T = \{x_i^T\}_{i=1}^{n_T}$. The source and target domain follows different distributions: $P(X_S) \neq P(X_T)$. The goal is to train a domain-invariant model φ with D_S and D_T that minimizes prediction error $\ell(\varphi(x^T), y^T)$ on the target domain. Typical single-modality methods either minimizes explicit measurements of domain gap [65], [66] or implicitly aligns source and target features with adversarial training [61], [67].

(2) **Domain generalization** [68], [69] (DG) considers a more practical scenario where target data are *unseen* during training, i.e., the model is only trained with labeled D_S . The model needs to learn source knowledge while retaining the ability to generalize to out-of-distributions (OOD). Single-modality methods aim to learn domain-invariant representations by metric learning [70], [71], adversarial training [72], [73], meta learning [74], [75], and so on.

(3) **Test-time adaptation** [76] (TTA) adjusts model behavior at inference time with high efficiency. The source pretrained source model φ is expected to produce accurate target predictions with minimum training. Since the target domain is unlabeled, typical methods adopt unsupervised measurements, e.g., entropy [77] to assess or improve the prediction confidences. A similar setting is source-free domain adaptation [78], [79] (SFDA). The difference is that SFDA assumes availability of the whole target domain for adaptation, while TTA only has access to data streams.

(4) **Few-shot learning** [80], [81] (FSL) allows limited annotation on target samples. Specifically, an N -way- K -shot FSL problem includes N -category target training data with K labeled samples for each class. FSL provides a data-efficient way for downstream adaptation, especially suitable for VLMs with vast base knowledge. The following method introductions are categorized according to the transferred modules and applied transfer settings introduced above.

3 PROMPT-BASED METHODS

As introduced in Section 2.1, prompts are in essence learnable parameters injected to the embeddings or features. In addition to the *text prompts* appended to the word embeddings [31], researchers have proposed *visual prompts* [82] and *context prompts* [83]. As depicted in Fig. 5, the differences between these prompts are the applied locations. Text prompts

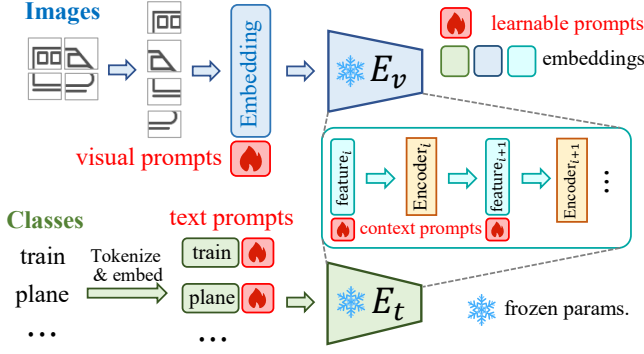


Fig. 5. Illustrations of text, visual and context prompts.

complement the textual descriptions of the language branch. Visual prompts are appended to the image patches of vision transformer [3]. Context prompts refine the intermediate features in both encoders. These prompting techniques exhibit distinct behaviors depending on the target tasks and optimization goals, which is elaborated as follows.

3.1 Prompting at Train Time

Proposed in [31], train-time prompting has become a representative methodology for fine-tuning VLMs. An overview of prompt-based tuning methods covering the solved problem setting, prompt type and method description is presented in Table 1. The major difference in the problem setting comes from the accessibility of labels, where FSL provides limited annotation for each class, while the rest settings are target-unlabeled.

3.1.1 Few-Shot Prompting

With labeled samples in FSL, prompt-tuning methods can adapt the VLMs to downstream tasks efficiently. Inspired by CoOp [31], there have been extensive explorations on the prompting style. While CoOp appends learnable prompts for text embeddings (type T), VPT [82] explores visual prompts to refine the tokenized input image patches for the vision transformer of CLIP (type V). Follow-up works [84], [85], [86] combine the strength of both modalities by simultaneously learning text and vision prompts. Rather than prompting *before* encoders, MaPLe [83] proposes to prompt intermediate features *in* the encoders (type C). Different from the three types of prompts that applies to the representation space, Bahng *et al.* [87] propose to learn *visible* prompts at the pixel level, an approach relevant to unadversarial training [88].

While the methods introduced above focus on the prompting position, there are attempts to improve the generalizability of the learned prompts. For better class-level base-to-new generalizability, CoCoOp [32] enhances the class-agnostic prompts in CoOp by introducing a class-conditioned prompting meta-network, while methods [89], [90], [91] explicitly disentangles learnable prompts for base and new classes. Taking inspirations from robust fine-tuning [92], KgCoOp [93] constrains learned prompts to be similar with default prompts to prevent overfitting. ProGrad [94] prevents forgetting by aligning prompt update gradients. PromptSRC [95] introduces prompt tuning regularizations through mutual agreement, self-ensembling

and textual diversity. Identifying the limited informativeness of the default prompt in Section 2.1 [96], ProText [97] introduces external knowledge from large language models (LLMs) like GPT3 [10] to guide the prompt tuning process.

3.1.2 Generalizable Prompting

Domain generalizable prompting aims to generalize learned prompts to unlabeled out-of-distributions (OOD), including the setting of DG and UDA. One line of research aims to learn *domain-invariant* prompts that is generalizable across different domains, as well as *domain-specific* prompts that encode domain knowledge [34], [36], [98], [99], [99], [100]. By modifying Eq. (5), the domain prompts are presented as:

$$t_c^d = [U]_0, \dots, [U]_k [V]_{k+1}^d, \dots, [V]_m^d [\text{class}]_c, \quad (8)$$

where $[U]_i$ are domain-invariant prompts, $[V]_i^d$ are specialized prompts for domain d , $d \in \{s, t\}$ represent source or target domain, and c is the index of category. The prompts in Eq. (8) are trained with labeled source data, and pseudo-labeled target data in UDA [34], [36]. The pseudo labels are generally obtained via VLMs' high-quality zero-shot predictions, which contributes to the significant performance boosts, as detailed in Section 6. STYLIP [99] trains additional modules to model domain style. DAMP [36] fosters knowledge and alignments between different modality encoders.

As a typical category of DG methods [115], data generation techniques synthesize diverse labeled data from unseen domain to enhance the model's robustness [116], [117], [118]. For VLM-based methods, data generation can be achieved at the feature level with learnable augmenters in a meta-learning style [101], [107], [108], or using more advanced generative techniques like stable diffusion [103]. As a special kind of augmentation, Hao *et al.* [105] observe that random noise can contribute to model robustness, and propose to utilize *quantization error* for regularization.

As a variant of UDA, multi-source UDA (MSDA) [119], [120] introduces multiple labeled source domains with different distribution, and the model is expected to aggregate all the domain knowledge to enhance target performance. Intuitively, MSDA can be solved by training prompts in Eq. (8) for each domain, i.e., $d \in \{s_0, s_1, \dots, s_N, t\}$ [94], [100]. With the pretrained general knowledge, VLM-based UDA methods can be extended for MSDA by viewing all source domains as a whole domain with mixed distribution [36].

3.2 Prompting at Test Time

Prompting for TTA requires on-the-fly adjustments of the VLM in an unsupervised manner. Current prompt-based TTA methods mostly base on TPT [40]. Denote the frozen VLM as $\varphi(x; \mathbf{t})$ that takes image input x and prompts $\mathbf{t} = \{t_i\}_{i=1}^C$ of all C classes. The TPT loss is defined as:

$$\mathcal{L}_{\text{tpt}} = \text{Ent}\left(\frac{1}{B} \sum_{i=1}^B \varphi(x_i; \mathbf{t})\right), \quad (9)$$

where $\{x_i\}_{i=1}^B$ are augmented views with high confidence given every original input x , $\text{Ent}(\mathbf{y}) = -\sum_{c=1}^C y_c \log y_c$ is Shannon entropy [121]. Intuitively, Eq. (9) maximizes the consistency among all reliable augmented views of a test input. Follow-up methods seek to improve this paradigm

TABLE 1
Prompt-based train-time generalization methods for VLMs. T, V, C represent text, visual, context prompt types, respectively.

Method	Venue	Setting	Type	Brief description
LDFS [101]	Arxiv	DG	T	A plug-and-play feature synthesis method to synthesize new domain features.
DSPL [102]	CVPR'24	DG	V,T	Disentangles text prompts using LLM then learn domain-specific and invariant prompts.
ODG-CLIP [103]	CVPR'24	DG	T	Generates OOD samples with diffusion models for training prompts.
Xiao <i>et al.</i> [104]	CVPR'24	DG	V,T	A probabilistic inference framework that considers both training and test distributions.
Hao <i>et al.</i> [105]	ECCV'24	DG	C	Utilizes quantization error as a kind of noise to explore quantization for regularization.
SPG [37]	ECCV'24	DG	T	Trains instance-specific soft prompts for unseen target domains.
CoOPood [106]	ICML'24	DG	T	A fine-grained prompt tuning method that mitigates spurious correlation.
OGEN [107]	ICLR'24	DG	T	Synthesizes OOD features to regularize decision boundaries between ID and OOD data.
STYLIP [99]	WACV'24	DG	T	Learns domain-agnostic prompts with a set of style projectors.
MetaPrompt [108]	TMLR'25	DG	T	Learns prompts to detect unknown class using meta-learning with momentum updates.
DAPrompt [34]	TNNLS'23	UDA	T	Learns domain information with domain-specific and domain-agnostic prompts.
AD-CLIP [98]	ICCVW'23	UDA	T	Learns domain-invariant prompts by conditioning on image style and content features.
DAMP [36]	CVPR'24	UDA	C,T	Learns domain-invariant semantics by mutually aligning visual and textual embeddings.
PDA [109]	AAAI'24	UDA	V,T	Learns a base and alignment branch to integrate class-related and cross-domain features.
Shi <i>et al.</i> [110]	IJCNN'24	UDA	T	Trains CLIP's prompts and an image encoder with data augmentation.
FUZZLE [111]	TFS'24	UDA	T	Integrates fuzzy C-means clustering and a fuzzy vector during prompt learning.
MPA [112]	NeurIPS'23	MSDA	T	Minimizes the domain gap between each source-target domain pair by prompt learning.
PGA [100]	NeurIPS'24	MSDA	T	Trains multi-domain prompts by solving a multi-objective optimization problem.
UPT [84]	Arxiv	FSL	V,T	Learns an additional network to simultaneously optimize vision and text prompts.
CoOp [31]	IJCV'22	FSL	T	Learns a class-agnostic text prompt to adapt VLMs to downstream tasks.
CoCoOp [32]	CVPR'22	FSL	T	Proposes to learn class-conditioned prompts with a meta-network.
ProDA [113]	CVPR'22	FSL	T	Learns from diverse prompts to handle the varying visual representations.
VPT [82]	ECCV'22	FSL	V	Learns visual prompts for the input of vision transformer.
MaPLe [83]	CVPR'23	FSL	C	Learns context prompts in both vision and text encoder.
PromptSRC [95]	CVPR'23	FSL	V,T	Promotes consistency between prompted and pretrained features to prevent forgetting.
KgCoOp [93]	CVPR'23	FSL	T	Regulates the learned prompts to be similar with the hand-crafted prompts.
ProGrad [94]	ICCV'23	FSL	T	Only updates prompts whose gradient align with the general knowledge.
DAPT [86]	ICCV'23	FSL	V,T	Learns multimodal prompts by finding the appropriate distribution in each modality.
DPT [85]	TMM'23	FSL	V,T	Simultaneously learns the visual and text prompts from the ends of both encoders.
DePT [89]	CVPR'24	FSL	T	Decouples and preserves base-specific knowledge in prompts.
DeCoOp [91]	ICML'24	FSL	T	Tunes prompts on base classes and evaluates on a combination of base and new classes.
ProText [97]	AAAI'25	FSL	C,T	Learns prompts from the rich contextual knowledge in LLM data.
ZIP [114]	ICLR'25	FSL	C	Reduces the problem dimensionality and the variance of zeroth-order gradient estimates.
DPC [90]	CVPR'25	FSL	T	Decouples the learning of base and new tasks at prompt level.

TABLE 2
Test-time prompt tuning methods for VLMs.

Method	Training loss
TPT [40]	\mathcal{L}_{tpt} : Test-time prompting loss in Eq. (9).
PromptAlign [122]	$\mathcal{L}_{\text{tpt}} + \ \varepsilon(\varphi(x; t)), \hat{\varepsilon}\ _1 + \ \sigma(\varphi(x; t)), \hat{\sigma}\ _1$.
DiffTPT [42]	\mathcal{L}_{tpt} with augmentations from diffusion model.
Swap-Prompt [123]	$\mathcal{L} = \ell_{\text{ce}}(\varphi(x; t'), \hat{y}) + \ell_{\text{ce}}(\varphi(x'; t'), \hat{y}) + \ell_{\text{ce}}(\varphi(x; t'), \varphi(x'; t)) + \ell_{\text{ce}}(\varphi(x'; t'), \varphi(x; t))$.
C-TPT [124]	$\mathcal{L}_{\text{tpt}} + \frac{1}{C} \sum_{c=1}^C \ \mathbf{t}_{\text{centroid}} - t_c\ _2$.
DynaPrompt [125]	\mathcal{L}_{tpt} on instance-relevant prompts.
R-TPT [126]	$\mathcal{L} = \frac{1}{B} \sum_{i=1}^B \text{Ent}(\varphi(x_i; t))$.
O-TPT [127]	$\mathcal{L}_{\text{tpt}} + \ \mathbf{E}\mathbf{E}^\top - \mathbf{I}\ _2^2$, \mathbf{I} is the identity matrix.

with further regularization or advanced augmentation techniques. Table 2 summarizes these methods and provides illustrative training losses. Note the losses are refined for better readability. Refer to the original papers for details.

The regularization-based TTA methods in Table 2 include: PromptAlign [122] additionally aligns the means ε and variances σ of the learned prompts with the source dataset statistics $\hat{\varepsilon}, \hat{\sigma}$ using L_1 norm. C-TPT [124] learns prompts similar to the text feature centroid $\mathbf{t}_{\text{centroid}} = \frac{1}{C} \sum_{c=1}^C t_c$. R-TPT [126] modifies the TPT loss in Eq. (9) by only minimizing point-wise entropy. O-TPT [127] applies orthogonal regularization on text feature matrix \mathbf{E} , where

the i_{th} row E_i represents prompt embedding of the i_{th} class. SwapPrompt [123] introduces target prompts t and online prompts t' that updates the target prompts in an Exponential Moving Average (EMA) manner, and encourages prediction consistencies among different prompts and augmented inputs x' . Other methods include DiffTPT [42] that replaces the standard augmentations, e.g., random cropping and flipping in [40], with images generated by diffusion model. DynaPrompt [125] only selects and optimizes relevant prompt for each test sample.

3.3 Discussion

Prompt tuning has become the mainstream approach to generalize VLMs in various cases. Categorized by prompting positions, prompts can be learned on text, visual embeddings and intermediate features. One can also tailor the prompts into domain-invariant and domain-specific parts to solve cross-domain generalization tasks. The computation efficiency of prompt tuning makes it suitable for test-time adaptation with proper calibration. Apart from generalization tasks, prompting can be applied as a task-agnostic tuning technique for VLMs [128], [129], [130]. Despite their effectiveness, prompt tuning may fall behind full fine-tuning in certain scenarios like large distribution gaps [131].

4 PARAMETER-BASED METHODS

Prompt-based methods introduced in Section 3 learn additional prompts with original parameters frozen, which may

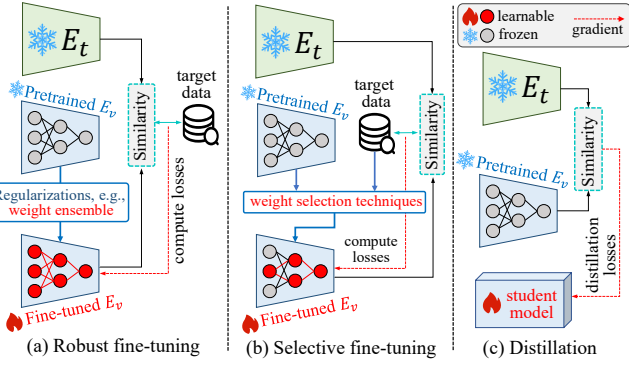


Fig. 6. Illustrations of parameter-based methods. For clarity, only the vision encoder is transferred. (a) Robust fine-tuning methods adopt regularizations like weight ensemble to prevent overfitting. (b) Selective fine-tuning methods choose task-relevant weights to update. (c) Distillation methods learn a new student model from pretrained VLMs' outputs.

not be sufficient for tasks with significant domain gap [131]. Fine-tuning on pretrained models is an effective approach for transfer learning [43], [51]. However, the delicate pretrained multimodal representations in VLMs require more strategical fine-tuning processes. Inspired by knowledge distillation (KD) [132], a more practical transfer approach distills the knowledge in VLMs to a smaller student for efficient deployment or data-sensitive cases. The two lines of study either modify the original parameters, or introduce new parameters, extending the ability of VLMs. Fig. 6 illustrates typical fine-tuning and distillation procedures.

4.1 Fine-Tuning Methods

Table 3 includes representative fine-tuning-based knowledge transfer methods for VLMs, mainly covering the setting of DG and TTA. Depending on the tuning process, these methods can be categorized as robust fine-tuning (FT-R) and selective fine-tuning (FT-S), as illustrated in Fig. 6(a),(b).

4.1.1 Robust Fine-Tuning

Traditional fine-tuning focuses more on acquiring new target knowledge, while the tuning process of VLMs highlights the challenge to *preserve* pretrained knowledge. Extensive researches have revealed that naive fine-tuning leads to severe forgetting and damages of pretrained knowledge [25], [26], [27], [131], which inspire the explorations on *robust* fine-tuning, aiming to adapt to target tasks without compromising the models' original abilities.

As categorized in Table 3, robust fine-tuning (FT-R) methods are mainly designed for DG tasks for better generalization ability. First revealed in [133], finding *flat minima* in losses has become a mainstream DG framework. Izmailov *et al.* [134] proposes a simple baseline stochastic weight averaging (SWA) that finds flat minima by averaging trained model weight snapshots every K epochs. Cha *et al.* [135] improves SWA by introducing iteration-granularity ensemble when the validation loss is low, namely Stochastic Weight Averaging Densely (SWAD). SWA and SWAD has been successfully applied to VLMs like CLIP and inspire further researches on weight-ensemble-based robust fine-tuning of VLMs. WiSE-FT [92] proposes to combine specifically the pretrained weights and fine-tuned weights. Instead

of preserving zero-shot weights via explicit mixing, follow-up works [46], [136] implicitly enforces trained weights to be similar with zero-shot ones. CLIPood [47] performs adaptive model ensemble with ensemble weights sampled from Beta distribution. DART [137] assembles model weights trained via different augmentations of input images, while WATT [138] obtain diverse model weights for ensemble via different prompt templates.

In addition to weight-level generalization, more diverse strategies have been explored to maintain zero-shot capabilities during fine-tuning. LP-FT [25] proposes to fine-tune with prediction heads initialized with frozen zero-shot weights. FLYP [139] preserves the pretraining *strategy* of zero-shot weights for fine-tuning. PADCLIP [131] prevents forgetting of zero-shot knowledge by adopting low and finely controlled learning rates. LipsumFT [140] connects zero-shot and fine-tuned vision encoder weights by the frozen text encoder. CaRot [48] adopts a new multimodal contrastive loss to promote larger smallest singular values then performs weight ensemble. UEO [141] extends the information maximization loss in [142] by minimizing in-distribution samples' information entropy while maximizing the OOD samples' entropy.

4.1.2 Selective Fine-Tuning

While robust full fine-tuning methods introduced in Section 4.1.1 are effective for generalization, they are usually computationally intensive thus cannot satisfy the fast adaptation and inference needs in TTA. Therefore, selective fine-tuning (FT-S) is proposed to adapt VLMs in limited response time. These methods either adjust certain statistics in model to align with the distribution of input data, or find the most task-relevant parameters to fine-tune.

Batch normalization (BN) layers [168] play an important part in ensuring fast convergence and reducing covariate shifts. Prior works [169], [170], [171] have discovered the effectiveness of simply adapting the statistics in BN layers for model generalization. Such finding further benefits the test-time adjustments of VLMs by replacing full parameter tuning with statistics tuning [138], [155], [158]. Rather than BN layers, Imam *et al.* [159] choose to update attention layers following current parameter-efficient fine-tuning advances [172], and Niu *et al.* [156] select batch-agnostic norm layers like group norm and layer norm.

The above introduced methods fine-tune a fix subset of model parameters, neglecting task specifics. The principles of selective fine-tuning [173], [174] are then utilized for VLMs, where model weights contributing more to the target task are updated. Furthermore, recent studies [175], [176] have shown that reduced fine-tuned parameters contribute to better generalization abilities, confirming the effectiveness of selective fine-tuning. DPLOT [157] explores such parameters by dividing the whole model into blocks then performs selection. SAFT [143] selects important parameters by evaluating the gradients of cross-entropy losses with respect to the parameters. Rather than selecting parameters, SAR [156] selects more reliable test samples for adaptation.

4.2 Generalizable Distillation

Fine-tuning methods in Section 4.1 need to access the model parameters for computing gradient, which is not viable for

TABLE 3

Parameter-based transfer methods for VLMs. FT-S, -R represent selective and robust fine-tuning methods. KD indicates knowledge distillation.

Method	Venue	Setting	Type	Brief description
SWAD [135]	NeurIPS'21	DG	FT-R	Finds flatter minima with a dense and overfit-aware stochastic weight sampling strategy.
WiSE-FT [92]	CVPR'22	DG	FT-R	Assembles the weights of pre-trained and fine-tuned models.
MIRO [46]	ECCV'22	DG	FT-R	Maximizes the mutual information between fine-tuned and pre-trained features.
LP-FT [25]	ICLR'22	DG	FT-R	Proposes fine-tuning with classification head initialized by linear-probing.
CLIPood [47]	ICML'23	DG	FT-R	Assembles model weights by Beta distribution and adopts a margin softmax loss for training.
FLYP [139]	CVPR'23	DG	FT-R	Proposes to use the exact pretraining loss for downstream fine-tuning.
DART [137]	CVPR'23	DG	FT-R	Combines the weights of models trained with differently augmented inputs.
UEO [141]	ICML'24	DG	FT-R	Leverages instance-level confidence to optimize through entropy.
SAFT [143]	ECCV'24	DG	FT-S	Selectively updates a small subset of parameters whose gradient magnitude is large.
CaRot [48]	NeurIPS'24	DG	FT-R	Proposes to fine-tune VLMs by enforcing a larger smallest singular value.
LipsumFT [140]	ICLR'24	DG	FT-R	Regularizes the Energy Gap from language model outputs for random texts during fine-tuning.
CAR-FT [136]	IJCV'24	DG	FT-R	Minimizes KL divergence between zero-shot and learned prompt weights during fine-tuning.
CLIP-TD [144]	Arxiv	DG	KD	Distills knowledge from CLIP into existing architectures using a dynamically weighted objective.
APD [145]	Arxiv	DG	KD	Multimodal knowledge KD framework that trains student model with adversarial examples.
DFKD [146]	MM'23	DG	KD	Distills a student model for distribution-agnostic downstream tasks with given categories.
Li <i>et al.</i> [33]	ICCV'23	DG	KD	Imitates teacher's visual representations to promote coherence in vision-language alignment.
RISE [147]	ICCV'23	DG	KD	Regularizes the student's learned representations to be close to the teacher's.
DAD [148]	NeurIPS'23	DG	KD	Leverages VLM teacher to generate adversarial examples and a VQGAN to discretize them.
SCI-PD [39]	CVPR'24	DG	KD	Transfers knowledge from VLMs to lightweight vision models with improved robustness.
VL2V-ADiP [38]	CVPR'24	DG	KD	Aligns vision and text modalities between the teacher and student model then perform KD.
PromptKD [149]	CVPR'24	DG	KD	Domain-specific prompt-based knowledge distillation with unlabeled target data.
KDPL [150]	ECCV'24	DG	KD	A KD framework that is applicable to prompt tuning methods of VLMs.
PADCLIP [131]	ICCV'23	UDA	FT-R	Fine-tunes CLIP with adjusted learning rates to prevent forgetting, and debias pseudo labels.
CLIP-Div [151]	Arxiv	UDA	KD	Learns a feature extractor and classifier with language-guided pseudo labels.
CMKD [152]	TCSVT'24	UDA	KD	Leverages VLMs as teacher models to guide the learning process in the target domain.
CustomKD [153]	CVPR'25	UDA	KD	Customizes the teacher features to reduce discrepancy between teacher and student models.
RLCF [154]	ICLR'24	TTA	KD	Adopts reinforcement learning to improve TTA with CLIP feedback to rectify the model outputs.
CLIP-OT [155]	Arxiv	TTA	FT-S	Adopts optimal transfer to integrate and distill multiple template knowledge.
SAR [156]	ICLR'23	TTA	FT-S	Removes samples with large gradients and encourages model weights towards flat minimum.
WATT [138]	NeurIPS'24	TTA	FT-S,R	Assembles weights learned from different text prompt templates.
DPLOT [157]	CVPR'24	TTA	FT-S	Selects and trains domain-specific model blocks with paired-view pseudo labeling.
CLIPArTT [158]	WACV'25	TTA	FT-S	Adapts norm layers by combining multiple class prompts as pseudo labels.
TTL [159]	WACV'25	TTA	FT-S	A PEFT TTA method that updates the attention weights of the transformer encoder.
POUF [160]	ICML'23	SFDA	FT	Directly fine-tunes the model or prompts on the unlabeled target data.
RCL [161]	Arxiv	SFDA	KD	Distillation from multiple VLMs with reliability-driven learning.
ViLAaD [162]	Arxiv	SFDA	KD	A framework that integrates VLMs as a powerful initialization for target adaptation.
DALL-V [163]	ICCV'23	SFDA	KD	Distills the rich multimodal knowledge in VLMs to a student model tailored for the target.
DIFO [164]	CVPR'24	SFDA	KD	First customizes a VLM from the target model then distills its knowledge to the target model.
Zhan <i>et al.</i> [165]	IJCAI'24	SFDA	KD	Fosters collaboration between the source trained model and the VLM.
Colearn++ [166]	IJCV'25	SFDA	KD	A co-learning strategy that generates reliable pseudo labels by pretrained models for finetuning.
ProDe [167]	ICLR'25	SFDA	KD	A proxy denoising mechanism that corrects VLM's predictions for domain-invariant KD.

black-box settings [177], [178] where the pretrained models are only accessed through API calls. Furthermore, the increasing scale of pretrained foundation models hinder the deployments on edge devices, which motivates distilling smaller student models with similar abilities with the large teacher model [38]. As summarized in Table 3, knowledge distillation (KD) methods aim to purify domain generalizable student models, or utilize the teacher knowledge in VLMs to refine pretrained source models.

Domain generalizable distillation. In the context of vision language distillation depicted by Eq. (6), the teacher model is the pretrained VLMs and the student model is generally a smaller vision model. The goal is to distill the knowledge in VLMs to the student model while also enhancing its OOD abilities. Similar to robust fine-tuning, generalizable distillation follows the principle of *preserving pretrained knowledge* of VLMs [33], [38], [147]. To achieve this, VL2V-ADiP [38] first aligns the representation space between the student model and the VLM's image encoder before distillation. RISE [147] constrains the distilled representations to be close to the pretrained ones. CustomKD [153] aligns features to overcome the model discrepancies brought by different structures, scales, etc. PromptKD [149] distills knowledge to a smaller VLM by

preserving the teacher's pretrained text features. Adversarial distillation methods [145], [148] introduce adversarial examples to enhance the robustness of student models. In addition to typical DG, SCI-PD [39] considers a more challenging open-set DG setting where target data might include novel classes. There are also researches on distilling for UDA [151], [152], [153]. RLCF [154] proposes to adopt pretrained VLMs as reward models for test-time adaptation with reinforcement learning.

Distillation-guided source model refinement. The distillation methods for DG introduced above generally distill knowledge into an independently initialized task-irrelevant student model. The source-free domain adaptation (SFDA) setting [76] provides a source-trained student model and unlabeled target data. The goal is to adapt the student model to accurately predict target data. Single-modality methods refine the source model with unsupervised measurements like pseudo labels [142], [179], but cannot achieve satisfactory results due to their low accuracies. VLMs provide a novel approach by distilling the pretrained knowledge on target data to the source model to solve SFDA, as shown in Table 3. Pseudo-label-based methods [162], [166], [167] aim to generate more reliable pseudo labels for target data utilizing clustering [162], [166] or proxy denois-

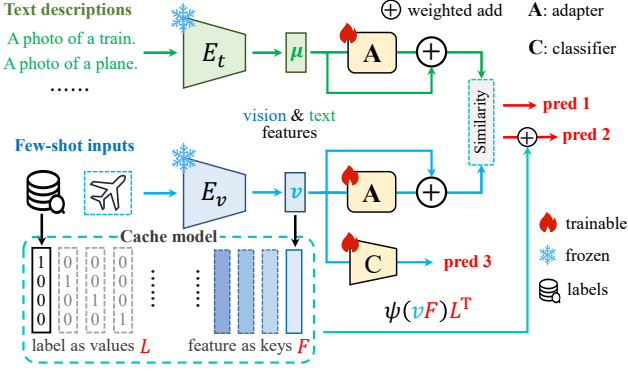


Fig. 7. Illustrations of feature-based methods, where **pred 1** is derived from features refined by adapters, **pred 2** includes cache model knowledge, and **pred 3** is from linear probing.

ing [167]. RCL [161] exploits multiple LLMs for more accurate pseudo labels and then transfer knowledge in a curriculum learning framework. Other methods explicitly distill the teacher predictions from target-customized VLMs [164] via adapters [163] or prompts [165]. Yu *et al.* [180] further explore the distillation methods that prevent forgetting in continual learning [181], [182].

While the methods introduced above mainly focus on *distribution-level* generalization for image classification tasks, KD is also applicable for *task-level* knowledge transfer, e.g., distilling VLMs for object detection [6] or semantic segmentation [183] tasks. The goal is to exploit and transfer the vast general knowledge in VLMs for downstream data-scarce or open-vocabulary scenarios [184], [185], [186] by bridging the granularity gap between image-level pretraining of VLMs and pixel-level target tasks [186], [187], [188], [189].

4.3 Discussion

Compared with prompt-based methods, parameter-based methods provide more flexibility for generalizing VLMs according to various problem settings, computation constraints, data availability, etc. Introduced in Section 4.1, fine-tuning methods extend the traditional pretraining-finetuning framework to VLMs by balancing the learning of target knowledge and preservation of zero-shot knowledge. Two major research lines include robust fine-tuning that regulates the updated parameters, and selective fine-tuning that finds crucial parameters to learn. Distillation methods in Section 4.2 introduce more flexible transfer methods that suit various practical scenarios and tasks.

5 FEATURE-BASED METHODS

The prompt-based methods in Section 3 and parameter-based methods in Section 4 require forwards and backwards of the whole VLM to compute gradients, bringing high computation expenses and training time. Observing that features extracted by pretrained VLMs are discriminative and informative, feature-based methods aim to perform adaptation and refinements on the feature level for better efficiencies. In this paper, feature-based methods are categorized as *cache-model-based* methods and *refinement-based* methods. As illustrated in Fig. 7, feature refinement is achieved by adapters [29] or linear-probing [11]. Cache models store representative features of input samples and

refine the final outputs. Recent advances in both fields are summarized in Table 4.

5.1 Feature Refinement

Rather than considering the whole VLM, feature refinement methods build upon the VLM-extracted vision and text features for further adaptation. These methods freeze the whole VLM and introduces external modules for various purposes, providing more dynamic adaptations.

5.1.1 Refine with Adapters

A typical refinement framework is Adapters proposed in [29], where a lightweight linear adapter module Φ generates feature-dependent residuals for complement, as shown in Eq. (7). Gao *et al.* [26] first adapts this scheme to VLMs by introducing dual adapters for each modality branch. Follow-up researches improve this paradigm by modifying the adapters' structures, positions, purposes and training objectives according to different adaptation settings.

Observing that multi-head attention [7] fosters cross-modal information exchanges, several methods [190], [192], [193], [197], [202] introduce attention layers as adapters for the purposes of multimodal information fusion [192], [193], [197] and interactions [190], [202]. In contrast to attention modules, adapters can be simplified into learnable parameters and variables. For example, TaskRes [199] learns task residuals that directly add to extracted text features. MTA [191] learns inlierness variables to assess reliability of augmented views for TTA. In addition to feature refinement, adapters are also used for disentangling and encompassing the domain and modality information in the multimodal features. VDPG [192] encompasses domain information, PromptTA [198] learns style features, and UniMoS [35] disentangles modality information by learning adapters. CLIP-CEIL [196] refines visual channels for domain-invariance and class-relevance. ReCLIP [207] and DAC [208] project vision and text features for distribution alignment. Adapters can also reconstruct or augment the features for various style information [200], [206].

5.1.2 Linear Probing

As discussed in [11], linear probing (LP) is a simple yet effective approach that directly tunes a linear classifier on extracted features for downstream predictions. It is pointed out that LP even surpass full fine-tuning on OOD data [25]. To inherent such merits, several researches have integrated LP into their method designs.

One direction is to linear probe with augmented style features. PromptStyler [194] learns diverse style word vectors to generate style-content features for training a domain generalizable linear classifier. Similar approach is adopted to generate scene styles (e.g., at night, under rain, in snow) for zero-shot adaptation of semantic segmentation [204], [205], where segmenter models are further trained. On the other hand, CLAP [201] retains the strong zero-shot abilities in VLMs and avoids prototype degradation during linear probing with Augmented Lagrangian Multiplier methods. Based on modality-disentangled representations, UniMoS [35] directly adopts LP for generating vision-modality outputs, and PromptTA [198] adopts a linear classifier and text adapter trained with style features for inference.

TABLE 4
Feature-based transfer methods for VLMs. Ca, A, TF represent cache-based, adapter-based and train-free methods.

Method	Venue	Setting	Type	Brief description
CALIP [190]	AAAI'23	TTA	A	Promotes interactions between visual and textual features using parameter-free attention layers.
MTA [191]	CVPR'24	TTA	A	Learns quality assessment variables to assess the quality of each augmented view.
VDPG [192]	ICLR'24	TTA	A	Trains a domain prompt generator to condense knowledge bank into domain-specific prompts.
L2C [193]	ICLR'25	TTA	A	Learns directly on the input space with an independent side branch via invert attention.
PromptStyler [194]	ICCV'23	DG	A	Learns diverse prompt vectors of the same class for training linear classifier.
GESTUR [195]	ICCVW'23	DG	A	Estimates unobservable gradients that reduces potential risks with parameterized experts.
CLIPCEIL [196]	NeurIPS'24	DG	A	Refines the visual feature channels for domain-invariance and class-relevance with an adapter.
Li <i>et al.</i> [197]	CVPR'24	DG	A	Learns a gating module to refine vision features with masks for both classification and detection.
PromptTA [198]	ICASSP'25	DG	A	Trains a text adapter to store domain information and a classifier for prediction.
TaskRes [199]	CVPR'23	FSL	A	Tunes a set of prior-independent parameters as a residual for encoding task-specific knowledge.
SHIP [200]	ICCV'23	FSL	A	Reconstructs the visual features with synthesized prompts by variational autoencoders.
CLAP [201]	CVPR'24	FSL	A	A hyperparameter-tuning-free method with linear probing constrained by Lagrangian method.
CLIPAdapter [26]	IJCV'24	FSL	A	Proposes to apply adapters to vision or language branch for feature refinement.
Candle [202]	KDD'24	FSL	A	Proposes compensating logit-adjusted loss for long-tail generalization with cross-modal attention.
UniMoS [35]	CVPR'24	UDA	A	Disentangles VLMs' feature into vision and language-associated components for finer adaptation.
Lai <i>et al.</i> [203]	WACV'24	UDA	A	Refines vision and text features with independent parameters for pseudo-labeling-based UDA
PODA [204]	ICCV'23	ZSL	A	Obtains stylistic features with different prompts then learns domain-wise segmenter models.
ULDA [205]	CVPR'24	ZSL	A	Aligns pixel-level, regional-level, and scene-level features with text embeddings.
LanDA [206]	Arxiv	MSDA	A	Learns domain-specific augmenters for image-free MSDA with only target textual descriptions.
ReCLIP [207]	WACV'24	SFDA	A	Learns a projection space to align visual and text features then update layer norms of encoders.
DAC [208]	WACV'24	FSL	A,Ca	Introduces text, visual caches and a visual adapter for inter- and intra-modal alignment.
CPL [209]	AAAI'24	FSL	A,Ca	Creates a visual concept cache and a projector for vision-to-language prompting.
DPE [210]	NeurIPS'24	TTA	Ca	Accumulates task-specific knowledge with evolving textual and visual prototypes.
TDA [41]	CVPR'24	TTA	Ca	Builds key-value cache that stores few-shot pseudo labels and corresponding features.
DMN [49]	CVPR'24	TTA	Ca	Builds dynamic and static memory for train-free zero- and few-shot adaptation.
BCA [211]	CVPR'25	TTA	Ca	Updates class embeddings to adapt likelihood while updating the prior of class embeddings.
TipAdapter [27]	ECCV'22	FSL	Ca	Builds non-parametric key-value cache from few-shot training sets.
SuS-X [212]	ICCV'23	FSL	Ca	Builds cache models with support set generated by Stable Diffusion or retrieved from datasets.
APE [213]	ICCV'23	FSL	Ca	Refines CLIP-extracted feature channels when building cache models.
Wang <i>et al.</i> [214]	ICLR'24	FSL	Ca	Builds a Gaussian Discriminate Analysis classifier to complement VLMs' zero-shot classifier.
CuPL [96]	ICCV'23	FSL	TF	Utilizes LLMs to construct more diverse category descriptions with more details.
ZERO [215]	NeurIPS'24	TTA	TF	Augments inputs, predicts then marginalizes over the most confident predictions.
BendVLM [216]	NeurIPS'24	TTA	TF	A nonlinear, train-free approach for TTA by VLM embedding debiasing.

5.2 Cache Model

Adapter-based methods still requires training the parameterized modules. Inspired by cache models [217], TipAdapter [27] proposes building cache models with VLM-extracted features. Denote the matrix of text features from all C classes as $\mathbf{W} = (\mu_1, \dots, \mu_C)$, $\mathbf{W} \in \mathbb{R}^{d \times C}$, where d is feature dimension. Given an input image with vision feature $v \in \mathbb{R}^d$, the zero-shot prediction logit before normalization can be computed with dot product: $\hat{y} = v\mathbf{W}$. The cache model is built by recording all vision features of few-shot training samples as keys $\mathbf{F} = (v_1, \dots, v_N)$, and their corresponding one-hot labels as values $\mathbf{L} = (y_1, \dots, y_N)$, where $\mathbf{F} \in \mathbb{R}^{d \times N}$, $\mathbf{L} \in \mathbb{R}^{C \times N}$. Given vision feature v of a tested sample, the cache model complements the zero-shot prediction results in the residual manner:

$$\hat{y} = \alpha\psi(v\mathbf{F})\mathbf{L}^\top + v\mathbf{W}, \quad (10)$$

where $\psi(x) = \exp(-\beta(1-x))$, and α, β are hyperparameters. The first term in Eq. (10) retrieves classification information of trained samples weighted by feature similarities, storing the target data information in a train-free manner.

Relevant works aim to improve this framework. SuS-X [212] enhances the vision-modality cache model in Eq. (10) by incorporating both vision and text features. SuS-X first computes cross-modal similarities between cache values and text features: $\mathbf{S} = \text{Softmax}(\mathbf{F}\mathbf{W}^\top)$, as well as normalized zero-shot probabilities of test samples: $s = \text{Softmax}(v\mathbf{W})$. An affinity matrix \mathbf{M} can be computed by

computing KL divergence: $M_{i,j} = \text{KL}(s_i, \mathbf{S}_j)$. Finally, prediction result in Eq. (10) is modified as:

$$\hat{y} = \alpha\psi(v\mathbf{F})\mathbf{L}^\top + v\mathbf{W} + \gamma\omega(-\mathbf{M})\mathbf{L}, \quad (11)$$

where γ is hyperparameter and $\omega(\cdot)$ rescales the values in \mathbf{M} . The third term in Eq. (11) introduces vision-language similarities when retrieving from the cache. APE [213] first refines the features in $\mathbf{F}, v, \mathbf{W}$ by selecting d' most informative channels from all d feature dimensions, obtaining $\mathbf{F}', v', \mathbf{W}'$. Then, the prediction in Eq. (10) is modified as:

$$\hat{y} = \alpha\psi(v'\mathbf{F}')(\text{diag}(\mathbf{R}')\mathbf{L}) + v\mathbf{W}, \quad (12)$$

where $\mathbf{R}' = \exp(\gamma\text{KL}(\mathbf{L}, \mathbf{F}'\mathbf{W}'^\top))$. Compared with Eq. (10), the first term in Eq. (12) introduces feature informativeness measured by the refined feature channels. Wang *et al.* [214] further improves the cache model by estimating weight and bias of the classifier using Gaussian Discriminate Analysis.

While the cache-model-based methods introduced above are designed for few-shot learning (FSL) with labeled samples, their high efficiency brought by the train-free design makes them extremely suitable for TTA. TDA [41] modifies Eq. (10) by constructing the key matrix \mathbf{F} with high quality test samples, whose highly-confident pseudo labels are used for the value matrix \mathbf{L} . A negative cache is also built with low-quality samples to further enhance TTA. To simultaneously handle test-time and few-shot adaptation settings, DMN [49] introduces a dynamic cache to store test-time samples and a static cache for optional few-shot training samples. Rather than building cache with extracted features, these representative prototypes can also be learned

TABLE 5
Widely-used datasets for evaluating VLMs.

Task	Dataset	Setting	# Class	# Sample	Brief description
Image classification	ImageNet [5]	FSL,DG,TTA	1000	1,281,167	A foundation image classification dataset.
	ImageNet-A [218]	FSL,DG,TTA	200	7,500	A variant of ImageNet with natural adversarial examples.
	ImageNet-R [219]	FSL,DG,TTA	200	30,000	A variant of ImageNet with renditions like art and cartoons.
	ImageNet-S [220]	FSL,DG,TTA	1000	50,889	A variant of ImageNet with black-and-white sketch images.
	ImageNet-V2 [221]	FSL,DG,TTA	1000	10,000	A non-overlapping variant with similar distribution of ImageNet
	Caltech101 [222]	FSL,TTA	101	9,146	Contains regular objects for recognition.
	OxfordPets [223]	FSL,TTA	37	7,349	Contains fine-grained breed labels of cats and dogs.
	StanfordCars [224]	FSL,TTA	196	16,185	Contains fine-grained vehicle models.
	Flowers102 [225]	FSL,TTA	102	8,189	Contains fine-grained flower species.
	Food101 [226]	FSL,TTA	101	101,000	Contains images of different foods with noises.
	FGVC-Aircraft [227]	FSL,TTA	100	10,000	Includes fine-grained annotation of aircraft models.
	SUN397 [228]	FSL,TTA	397	39,700	Vision dataset for scene recognition and understanding.
	DTD [229]	FSL,TTA	47	5,760	Includes images of regular texture for recognition.
	EuroSAT [230]	FSL,TTA	10	2,700	Includes satellite sensing images of land cover situations.
	UCF101 [231]	FSL,TTA	101	13,320	Action recognition dataset collected from video clips.
	PACS [232]	DG	7	9,991	Includes images from domains <i>photo</i> , <i>art</i> , <i>cartoon</i> and <i>sketch</i> .
	VLCS [233]	DG	5	10,729	Includes images from Caltech101, LabelMe, SUN09 and VOC2007.
	TerraIncognita [234]	DG	10	24,330	Includes animal images taken at different locations.
	OfficeHome [235]	DG,UDA	65	15,500	Includes office items from domains <i>Art</i> , <i>Clipart</i> , <i>Product</i> , <i>RealWorld</i> .
	DomainNet [119]	DG,UDA	345	586,575	Includes pictures from 6 domains for multi-source DA.
	VisDA [236]	UDA	12	207,785	Include synthetic and real images for knowledge transfer.
Image-text	VQAv2 [237]	QA	-	265,016	Includes images with at least 3 open-ended questions to answer.
	COCO caption [238]	Retrieval	-	82,783	Includes images with human-generated captions.
	Flickr30k [239]	Retrieval	-	31,783	Includes images with sentence-based image description.
Video-text	TVQA [240]	QA	-	21,793	Includes QA pairs from 6 TV shows and 21793 video clips.
	How2QA [241]	QA	-	9,035	Includes QA pairs with one correct answer and three distractors.
	TVC [242]	Caption	-	108,965	Includes short video moments and corresponding captions.
Navigation	R2R [243]	VLN	-	21,567	Includes building scenes with language instructions for navigation.
	REVERIE [244]	VLN	-	21,702	An indoor grounding dataset with target objects and instructions.
Object detection	COCO 2014 [245]	OD	80	83,000	Large-scale datasets for object detection, segmentation, keypoint detection, and captioning.
	COCO 2017 [245]	OD	80	118,000	
Semantic segmentation	Cityscapes [246]	SS	19	2,975	Focuses on semantic understanding of urban street scenes.
	GTA5 [247]	SS	19	24,966	Pixel-level semantic annotations from the game Grand Theft Auto 5.

as the test-time adaptation proceeds [210], [211]. Cache models can be integrated with adapters that further refine the cache contents [208], [209]. In addition to cache models, there are other train-free (TF) methods aided by external knowledge [96], utilizing various augmentations [215] or embedding debiasing [216], as summarized in Table 4.

5.3 Discussion

Feature-based methods exploit the strong representation ability of pretrained VLMs by performing adaptation mainly on extracted features. Refinement-based methods in Section 5.1 introduces external adaptation modules like Adapters [29] to modify the features, or replace the cross-modal prediction head (Eq. (4)) by directly training a linear classifier on vision features. Some researches have pointed out that VLMs' pretrained features sometimes surpasses the fine-tuned features [25], [201], which motivates train-free cache-model-based methods. These methods record the extracted features of train samples as class prototypes to benefit the zero-shot prediction. To conclude, feature-based methods provide more computation efficiency, can be suitable for agile adaptation and deployment of VLMs in resource-constrained scenarios like edge devices.

6 DATASETS AND EXPERIMENTAL RESULTS

This section reviews popular datasets used for evaluating the generalization of VLMs and comprehensively compares

TABLE 6
Performance comparison on domain-level DG tasks.

Method	Type	Backbone	PACS	VLCS	OH	DN	TI	Avg.
DFKD-VLFM [146]	KD	ResNet50	78.3	76.0	-	-	-	-
RISE [147]	KD	ResNet50	90.2	82.4	72.6	-	54.0	-
DART [137]	FT-R	ResNet50	88.9	80.3	71.9	47.1	51.3	67.9
CLIP-zeroshot [11]	-	ViT-B/16	96.2	81.7	82.0	57.5	33.4	70.2
VL2V-ADiP [38]	KD	ViT-B/16	96.7	83.3	87.4	62.8	58.5	77.7
STYLIP [99]	T	ViT-B/16	97.0	82.9	83.9	68.6	-	-
SPG [37]	T	ViT-B/16	97.0	82.4	83.6	60.1	50.2	74.7
DSPL [102]	V,T	ViT-B/16	97.5	86.4	86.1	62.1	57.1	77.8
Xiao <i>et al.</i> [104]	V,T	ViT-B/16	98.5	87.0	86.0	61.8	-	-
SWAD [135]	FT-R	ViT-B/16	91.3	79.4	76.9	51.7	45.4	68.9
MIRO [46]	FT-R	ViT-B/16	95.6	82.2	82.5	54.0	54.3	73.7
CLIPood [47]	FT-R	ViT-B/16	97.3	85.0	87.0	63.5	60.4	78.6
UEO [141]	FT-R	ViT-B/16	96.9	81.3	86.0	60.8	51.5	75.3
CAR-FT [136]	FT-R	ViT-B/16	96.8	85.5	85.7	62.5	61.9	78.5
DPL [248]	A	ViT-B/16	96.4	80.9	83.0	59.5	46.6	73.3
GESTUR [195]	A	ViT-B/16	96.0	82.8	84.2	58.9	55.7	75.5
CLIPCEIL [196]	A	ViT-B/16	97.6	88.4	85.4	62.0	53.0	77.3
PromptStyler [194]	A	ViT-B/16	97.2	82.9	83.6	59.4	-	-
PromptTA [198]	A	ViT-B/16	97.3	83.6	82.9	59.4	-	-

the performances of existing knowledge transfer methods.

6.1 Datasets

Table 5 summarizes representative datasets on image, text, video modalities for tasks including classification, retrieval, captioning, question answering (QA), vision-language navigation (VLN), object detection (OD) and semantic segmentation (SS). The **Setting** column indicates the problems that the dataset is designed to solve.

TABLE 7

Performance comparison over various methods, settings and backbones on dataset-level generalization. Method types correspond with the types introduced in Table 1, Table 3 and Table 4. All methods are based on CLIP [11] with different vision backbones.

Method	Type	Backbone	ImageNet family						Few-shot datasets										
			-I	-V2	-S	-A	-R	Avg.	Calt	Pets	Cars	Flower	Food	FGVC	SUN	DTD	Euro	UCF	Avg.
CLIP-zeroshot [11]		ResNet50	58.2	51.4	33.3	21.7	56.0	44.1	85.1	83.1	55.7	65.4	74.2	17.1	58.6	42.2	37.6	61.2	58.0
CLIP-zeroshot [11]		ViT-B/16	66.7	60.8	46.2	47.8	74.0	59.1	93.6	86.9	66.1	67.0	82.9	23.2	65.6	45.0	50.4	65.2	64.6
CLIP-zeroshot [11]		ViT-B/16 [†]	-	-	-	-	-	-	95.4	94.1	68.7	74.8	90.7	31.1	72.2	56.4	60.0	73.9	71.7
Few-shot learning (FSL) methods (16-shot results are provided)																			
VPT [82]	V	ViT-B/16	-	-	-	-	-	-	96.4	96.8	73.1	81.1	91.6	34.7	78.5	67.3	77.7	79.0	77.6
DPT [85]	V,T	ViT-B/16	-	-	-	-	-	-	95.6	91.2	82.6	96.6	79.3	48.4	71.0	70.2	91.2	81.4	80.7
UPT [84]	V,T	ViT-B/16	72.6	64.4	48.7	50.7	76.2	62.5	95.9	93.0	84.3	97.1	85.0	46.8	75.9	70.7	90.5	84.0	82.3
Wang et al. [214]	Ca	ViT-B/16	-	-	-	-	-	-	92.6	88.8	75.1	95.7	79.1	40.6	70.7	66.5	86.1	77.5	77.3
APE [213]	Ca	ViT-B/16	-	-	-	-	-	-	92.3	88.0	70.5	92.0	78.4	31.2	69.6	67.4	78.4	74.5	74.2
CLAP [201]	A	ViT-B/16	71.8	64.1	47.7	48.4	76.7	61.7	-	-	-	-	-	-	-	-	-	-	-
Candle [202]	A	ViT-B/16	71.6	62.8	48.3	49.1	75.0	61.4	91.3	88.9	64.6	68.3	85.5	24.2	66.1	44.6	48.4	67.2	64.9
TaskRes [199]	A	ViT-B/16	73.1	65.3	49.1	50.4	77.7	63.1	93.4	87.8	76.8	96.0	77.6	36.3	70.7	67.1	84.0	78.0	76.8
CoCoOp [32]	T	ViT-B/16 [†]	71.0	64.1	48.8	50.6	76.2	62.1	95.8	96.4	72.0	81.7	91.0	27.7	78.3	64.9	71.2	77.6	75.7
CoOp [31]	T	ViT-B/16 [†]	71.5	64.2	48.0	49.7	75.2	61.7	93.7	94.5	68.1	74.1	85.2	28.8	72.5	54.2	68.7	67.5	70.7
KgCoOp [93]	T	ViT-B/16 [†]	71.2	64.1	49.0	50.7	76.7	62.3	96.0	96.2	73.4	83.7	91.1	34.8	78.4	64.4	73.5	79.7	77.1
ProGrad [94]	T	ViT-B/16 [†]	70.5	63.4	48.2	49.5	75.2	61.3	95.9	96.3	72.9	82.0	90.0	32.8	77.6	62.5	72.7	79.4	76.2
DePT [89]	T	ViT-B/16 [†]	-	-	-	-	-	-	96.3	96.4	77.8	86.5	91.2	40.7	81.1	71.1	84.9	82.3	80.8
DeCoOp [91]	T	ViT-B/16 [†]	-	-	-	-	-	-	96.5	95.3	73.2	84.2	90.7	31.4	78.1	62.7	74.6	77.7	76.4
DPC [90]	T	ViT-B/16 [†]	-	-	-	-	-	-	96.5	96.7	78.5	83.6	91.5	40.2	80.8	66.8	84.1	83.2	80.2
MaPLE [83]	C	ViT-B/16 [†]	70.7	64.1	49.2	50.9	77.0	62.4	96.0	96.6	73.5	82.6	91.4	36.5	79.8	68.2	82.4	80.8	78.8
ZIP [114]	C	ViT-B/16 [†]	66.2	59.7	45.5	47.1	75.2	58.7	95.6	95.9	-	72.7	89.9	31.2	70.9	55.8	72.9	72.2	-
PromptSRC [95]	V,T	ViT-B/16 [†]	71.3	64.4	49.6	50.9	77.8	62.8	96.0	96.3	76.6	86.0	91.1	40.2	80.5	71.8	82.3	82.7	80.3
DAPT [86]	V,T	ViT-B/16 [†]	72.2	64.9	48.3	48.7	75.8	62.0	92.7	81.8	66.8	75.5	89.7	27.5	78.2	64.7	59.2	77.7	71.4
ProText [97]	C,T	ViT-B/16 [†]	70.2	63.5	49.5	51.5	77.4	62.4	96.8	96.5	69.8	76.4	91.1	32.4	77.6	62.3	68.7	77.5	74.9
CuPL [96]	TF	ViT-B/16 [†]	69.6	63.3	49.0	50.7	77.1	61.9	95.6	95.8	68.8	76.1	90.6	33.2	75.7	61.3	61.8	76.4	73.5
CPL [209]	A,Ca	ViT-B/16 [†]	-	-	-	-	-	-	96.7	97.0	78.0	88.4	92.9	40.5	80.8	70.4	87.1	83.3	81.5
SHIP [200]	A	ViT-B/16 [†]	-	-	-	-	-	-	96.3	96.0	77.2	85.7	91.1	39.0	79.4	70.2	87.1	81.9	80.4
Domain generalization (DG) methods																			
KDPL [150]	KD	ViT-B/32	66.3	58.6	43.1	31.7	68.8	53.7	93.6	88.2	60.4	65.1	80.9	18.2	65.7	43.3	44.3	63.4	62.3
OGEN [107]	T	ViT-B/16	-	-	-	-	-	-	95.1	91.4	66.0	72.9	86.5	23.0	68.4	46.4	45.8	69.7	66.5
STYLIP [99]	T	ViT-B/16	-	-	-	-	-	-	95.5	91.6	67.1	72.4	88.6	25.2	68.1	47.9	48.2	69.3	67.4
Hao et al. [105]	C	ViT-B/16	70.7	63.9	48.9	51.1	76.9	62.3	94.1	90.5	66.0	71.3	86.2	22.7	66.8	44.2	48.2	69.2	65.9
CLIPood [106]	FT-R	ViT-B/16	71.6	64.9	49.3	50.4	77.2	62.7	-	-	-	-	-	-	-	-	-	-	-
UEO [141]	FT-R	ViT-B/16	71.9	65.1	48.6	48.6	75.9	62.0	-	-	-	-	-	-	-	-	-	-	-
CaRot [48]	FT-R	ViT-B/16	83.1	74.1	52.7	51.6	77.7	67.9	-	-	-	-	-	-	-	-	-	-	-
LipsumFT [140]	FT-R	ViT-B/16	83.3	73.6	51.4	49.9	75.9	66.8	-	-	-	-	-	-	-	-	-	-	-
CAR-FT [136]	FT-R	ViT-B/16	83.8	74.0	53.0	49.5	75.4	67.1	-	-	-	-	-	-	-	-	-	-	-
SAFT [143]	FT-S	ViT-B/16	72.7	65.8	49.7	51.6	78.1	63.6	94.2	90.9	65.3	70.8	86.3	25.2	68.2	46.2	49.6	70.2	66.7
PromptKD [149]	KD	ViT-B/16	-	-	-	-	-	-	93.6	91.6	73.9	75.3	88.8	26.2	68.6	55.1	63.7	76.4	71.3
APD [145]	KD	ViT-B/16	-	-	-	-	-	-	93.9	88.5	81.1	94.4	71.5	47.8	71.4	65.5	76.9	77.1	76.8
Test-time adaptation (TTA) methods																			
Ma et al. [123]	T	ResNet50	61.8	53.9	38.2	24.5	60.9	47.9	89.9	89.1	59.6	70.2	75.1	-	63.9	47.3	46.6	65.7	-
CALIP [190]	A	ResNet50	-	-	-	-	-	-	87.7	86.2	56.3	66.4	77.4	17.8	58.6	42.4	38.9	61.7	59.3
R-TPT [126]	T	ResNet50	-	-	-	-	-	-	86.7	84.6	58.1	60.6	-	17.5	-	41.3	21.2	59.7	-
DiffTPT [42]	T	ViT-B/16	70.3	65.1	46.8	55.7	75.0	62.6	92.5	88.2	67.0	70.1	87.2	25.6	65.7	47.0	43.1	68.2	65.5
C-TPT [124]	T	ViT-B/16	69.3	63.4	48.5	52.9	78.0	62.4	94.1	87.4	66.7	69.9	84.5	23.9	66.0	46.8	48.7	66.7	65.5
TPT [40]	T	ViT-B/16	73.6	66.8	49.3	58.0	77.3	65.0	94.2	87.8	66.9	69.0	84.7	24.8	65.5	47.8	42.4	68.0	65.1
O-TPT [127]	T	ViT-B/16	67.3	61.7	47.1	49.9	72.6	59.7	94.0	88.0	64.5	70.1	84.1	23.6	64.2	45.7	42.8	64.2	64.1
Abdul et al. [122]	T	ViT-B/16	-	65.3	50.2	59.4	79.3	63.6	94.0	90.8	68.5	72.4	86.7	24.8	67.5	47.2	47.9	69.5	66.9
Xiao et al. [125]	T	ViT-B/16	69.6	64.7	48.2	56.2	78.2	63.4	94.3	88.3	67.7	70.0	85.4	24.3	66.3	48.0	42.3	68.7	65.5
WATT [138]	FT-S	ViT-B/16	64.1	62.4	49.0	51.0	75.4	60.4	93.6	88.1	66.8	68.5	84.8	24.2	65.8	46.9	52.0	65.7	65.6
TTL [159]	FT-S	ViT-B/16	70.2	64.6	48.6	60.5	77.5	64.3	93.6	88.7	68.0	70.5	85.1	23.8	66.3	46.7	42.0	69.2	65.4
DPE [210]	Ca	ViT-B/16	71.9	65.4	52.3	59.6	80.4	65.9	94.8	91.1	67.3	75.1	86.2	29.0	70.1	54.2	55.8	70.4	69.4
TDA [41]	Ca	ViT-B/16	69.5	64.7	50.5	60.1	80.2	65.0	94.2	88.6	67.3	71.4	86.1	23.9	67.6	47.4	58.0	70.7	67.5
DMN [49]	Ca	ViT-B/16	72.3	65.2	53.2	58.3	78.6	65.5	95.4	92.0	68.0	74.5	85.1	30.0	70.2	55.9	59.4	72.5	70.3
BCA [211]	Ca	ViT-B/16	70.2	64.9	50.9	61.1	80.7	65.6	94.7	90.4	66.9	73.1	86.0	28.6	68.4	53.5	56.6	67.6	68.6
MTA [191]	A	ViT-B/16	70.1	64.2	49.6	58.1	78.3	64.1	94.2	88.2	68.5	68.1	85.0	25.2	66.7	45.9	45.4	68.7	65.6
HisTPT [249]	T,Ca	ViT-B/16	-	-	-	-	-	-	94.5	89.1	69.2	71.2	89.3	26.9	67.2	48.9	49.7	70.1	67.6
RLCF [154]	KD	ViT-B/16	75.5	70.4	57.7	75.2	87.2	73.2	-	-	-	-	-	-	-	-	-	-	-
ZERO [215]	TF	ViT-B/16	70.9	65.2	50.3	64.1	80.8	66.2	94.1	87.2	68.5	66.8	84.6	24.4	66.9	45.9	-	68.6	-

† applies harmonic mean (HM) on few-shot datasets, as defined in [250].

We can observe that the majority of works focus on image classification, where the VLM needs to generalize to shifted data distributions [119], [218], [219], [220], [235], fine-grained categories [223], [224], [225], [227] or specialized data not included in VLMs’ pretraining datasets [227], [229], [230]. With slight modifications, their vision perception abilities can be applied to more visual tasks like object detection [245] and semantic segmentation [246], [247]. In addition, the vision-language processing ability of VLMs

enables them to solve diverse multimodal tasks including visual question answering [237], [240], [241], image-text retrieval [238], [239], and vision-language navigation [243] where an agent is expected to complete navigation tasks as instructed by natural languages [243], [243].

6.2 Experimental Results

Aligning with recent research focus, this survey mainly provides performance comparisons for *image classification*

TABLE 8

Performance comparison of domain-wise transfer tasks. On OfficeHome, all 12 sets of cross-domain results are provided, e.g., task AC means adaptation from source domain A to target domain C. On other datasets the averaged results are provided.

Method	Type	OfficeHome														VisDA		DomainNet	
		Backbone	AC	AP	AR	CA	CP	CR	PA	PC	PR	RA	RC	RP	Avg.	Backbone	Avg.	Backbone	Avg.
CLIP-zeroshot [11]		ResNet50	51.7	81.5	82.3	71.7	81.5	82.3	71.7	51.7	82.3	71.7	51.7	81.5	71.8	ResNet101	84.4	ViT-B/16	56.2
CLIP-zeroshot [11]		ViT-B/16	67.8	89.0	89.8	82.9	89.0	89.8	82.9	67.8	89.8	82.9	67.8	89.0	82.4	ViT-B/16	88.9	-	-
Unsupervised domain adaptation (UDA) methods (sorted by mean accuracy on OfficeHome)																			
Xiao <i>et al.</i> [104]	V,T	ResNet50	51.6	81.9	82.6	71.9	81.9	82.6	71.9	51.6	82.6	71.9	51.6	81.9	72.0	ResNet101	84.4	ViT-B/16	56.2
CLIP-Div [151]	KD	ResNet50	57.2	80.4	82.9	73.9	80.7	81.1	72.8	58.6	83.5	73.3	59.9	81.9	73.9	ResNet101	80.8	ViT-B/16	54.6
DAPrompt [34]	T	ResNet50	54.1	84.3	84.8	74.4	83.7	85.0	74.5	54.6	84.8	75.2	54.7	83.8	74.5	ResNet101	86.9	ViT-B/16	59.8
Shi <i>et al.</i> [110]	T	ResNet50	54.6	84.6	85.1	75.8	84.2	85.1	74.5	54.1	85.2	75.2	54.9	84.0	74.8	ResNet101	86.8	ViT-B/16	59.5
FUZZLE [111]	T	ResNet50	55.9	84.6	85.5	75.0	84.8	84.3	74.0	56.2	85.1	75.1	56.2	86.4	75.3	ResNet101	86.5	ViT-B/16	59.5
PDA [109]	V,T	ResNet50	55.4	85.1	85.8	75.2	85.2	85.2	74.2	55.2	85.8	74.7	55.8	86.3	75.3	ResNet101	86.4	-	-
AD-CLIP [98]	T	ResNet50	55.4	85.2	85.6	76.1	85.8	86.2	76.7	56.1	85.4	76.8	56.1	85.5	75.9	ResNet101	87.7	-	-
Lai <i>et al.</i> [203]	A	ResNet50	58.1	85.0	84.5	77.4	85.0	84.7	76.5	58.8	85.7	75.9	60.4	86.4	76.5	ResNet101	89.2	ViT-B/16	62.3
PADCLIP [131]	FT-R	ResNet50	57.5	84.0	83.8	77.8	85.5	84.7	76.3	59.2	85.4	78.1	60.2	86.7	76.6	ResNet101	88.5	ViT-B/16	63.7
UniMoS [35]	A	ResNet50	59.5	89.4	86.9	75.2	89.6	86.8	75.4	58.4	87.2	76.9	59.5	89.7	77.9	ResNet101	88.1	ViT-B/16	63.6
DAMP [36]	C,T	ResNet50	59.7	88.5	86.8	76.6	88.9	87.0	76.3	59.6	87.1	77.0	61.0	89.9	78.2	ResNet101	88.4	ViT-B/16	63.2
CMKD [152]	KD	ResNet50	65.9	86.6	87.3	74.4	87.7	85.8	75.9	64.4	87.9	79.1	67.2	90.0	79.3	ResNet101	87.0	ViT-B/16	53.9
DAPrompt [34]	T	ViT-B/16	70.6	90.2	91.0	84.9	89.2	90.9	84.8	70.5	90.6	84.8	70.1	90.8	84.0	ViT-B/16	89.5	-	-
PDA [109]	V,T	ViT-B/16	73.5	91.4	91.3	86.0	91.6	91.5	86.0	73.5	91.7	86.4	73.0	92.4	85.7	ViT-B/16	89.7	-	-
ADCLIP [98]	T	ViT-B/16	70.9	92.5	92.1	85.4	92.4	92.5	86.7	74.3	93.0	86.9	72.6	93.8	86.1	ViT-B/16	90.7	-	-
PADCLIP [131]	FT-R	ViT-B/16	76.4	90.6	90.8	86.7	92.3	92.0	86.0	74.5	91.5	86.9	79.1	93.1	86.7	ViT-B/16	90.9	-	-
UniMoS [35]	A	ViT-B/16	74.9	94.0	92.5	86.4	94.3	92.5	86.0	73.9	93.0	86.4	74.2	94.5	86.9	ViT-B/16	90.1	-	-
DAMP [36]	C,T	ViT-B/16	75.7	94.2	92.0	86.3	94.2	91.9	86.2	76.3	92.4	86.1	75.6	94.0	87.1	ViT-B/16	90.9	-	-
Lai <i>et al.</i> [203]	A	ViT-B/16	78.2	90.4	91.0	87.5	91.9	92.3	86.7	79.7	90.9	86.4	79.4	93.5	87.3	ViT-B/16	91.7	-	-
CMKD [152]	KD	ViT-B/16	79.4	94.2	92.7	86.3	93.4	92.2	86.7	79.5	92.1	88.2	81.2	94.5	88.4	ViT-B/16	90.3	-	-
Source-free domain adaptation (SFDA) methods																			
DIFO [164]	KD	ResNet50	62.6	87.5	87.1	79.5	87.9	87.4	78.3	63.4	88.1	80.0	63.3	87.7	79.4	ResNet101	88.8	-	-
ProDe [167]	KD	ResNet50	64.0	90.0	88.3	81.1	90.1	88.6	79.8	65.4	89.0	80.9	65.5	90.2	81.1	ResNet101	88.7	-	-
ViLaAD [162]	KD	ViT-B/32	70.1	91.6	89.9	83.2	92.0	90.0	81.0	71.7	89.9	83.3	71.3	92.3	83.9	ViT-B/32	90.5	-	-
RCL [161]	KD	ViT-B/32	83.1	95.7	93.1	89.2	95.3	92.6	89.2	82.3	92.9	90.0	83.2	95.5	90.2	ViT-B/32	93.2	-	-
Zhan <i>et al.</i> [165]	KD	ViT-B/16	71.1	87.1	91.3	86.3	90.9	91.6	86.6	74.1	91.8	87.6	75.0	91.8	85.4	-	-	-	-
Co-learn++ [166]	KD	ModelZoo [†]	80.0	91.2	91.8	83.4	92.7	91.3	83.4	78.9	92.0	85.5	80.6	94.7	87.1	ModelZoo [†]	91.1	-	-

[†] ModelZoo means the results are from a mixture of different pretrained models detailed in [166].

tasks evaluated by classification accuracies. The comparison covers various generalization settings introduced in Section 2.2. The compared methods adopt CLIP [11] with different backbones in the vision encoder, which are indicated in the **Backbone** column of each result table. The method types summarized in Table 1, Table 4, Table 3 are also included.

6.2.1 Dataset-Level Generalization

In dataset-level generalization, VLMs need to adapt to various tasks with different scenarios, categories and data distributions. Table 7 presents comprehensive results and comparisons over various methods introduced in this survey. Following current research lines, we present dataset-level VLM generalization tasks evaluated on the ImageNet family and other 10 specialized or fine-grained image datasets introduced in Table 5. There are mainly three settings that tackle dataset-level generalization. **(1) Few-shot learning (FSL)** provides few-shot labeled training data for each dataset and each category. Note that FSL on the ImageNet family only trains on ImageNet-1k (-I) and evaluates on all other variants. Some FSL methods split dataset categories into base and novel parts, where the model is only adapted with data from base classes. Harmonic mean [250] is adopted to assess the general model performance across both base and novel classes. **(2) Domain generalization (DG)** assumes no available data on the evaluated datasets, where the model is only trained on ImageNet and expected to generalize to unseen tasks. **(3) Test-time adaptation (TTA)** directly generalizes and evaluates pretrained VLMs on target datasets on-the-fly without the training process.

From the results we can conclude the following factors that affect the classification performance. **(1) Backbones.**

The majority of methods base on the vision transformer (ViT) family [3] and ResNet family [2], where ViT-based method exhibit a significant accuracy gain of $\sim 15\%$. Such phenomenon indicates that the generalization ability of VLMs is largely related to that of vision encoders. **(2) Transfer settings.** With labeled samples from the target dataset, FSL methods exhibit the best overall performance on few-shot datasets. The overall performances of DG and TTA methods are comparable. Specifically, on hard OOD datasets with large distribution gaps, e.g., ImageNet-A, FGVC, EuroSAT, TTA methods exhibit superior performances. The reason is that TTA methods can access these hard samples and make quick adjustments to mitigate the gap. On more general tasks with in-distribution data, DG methods perform better due to sufficient tuning on source data with similar distributions. **(3) Method designs.** Under the same setting and backbone, different method types contribute to the subtle performance fluctuations. For example, on FSL tasks, prompt-tuning methods perform generally better at the expenses of higher computation overheads. On the contrary, on TTA tasks, prompt-tuning methods deteriorate due to insufficient training and the absence of ground truths. Cache models are better at capturing the representative features of streamlined data flows, making them suitable for TTA tasks. Distillation-based methods (KD methods) surpass other competitors by introducing external knowledge from larger teacher models (ViT-L/14 in Table 7).

6.2.2 Domain-Level Generalization

On domain-level DG tasks, each dataset is composed of several distinct *domains* with the same label space but different data distribution. Table 6 presents evaluation results

based on the five domain-level DG datasets (OH is short for OfficeHome, DN is short for DomainNet, TI is short for TerraIncognita) established in DomainBed [251]. The methods follow leave-one-out protocol that generalizes to each target domain by tuning on all other domains. The reported accuracies are averaged over all tested target domains. It can be observed that fine-tuning-based (FT) methods [47], [136] perform generally better than other methods, especially on harder datasets like TerraIncognita. The reason is that pretrained VLMs cannot overcome significant domain gaps without full fine-tuning [131]. The adapter-based methods achieve slightly better results on easier tasks where CLIP’s zero-shot accuracies are high, but fall far behind FT methods on challenging tasks, negatively affecting their overall performances. However, the absolute accuracies of FT methods on large distribution shifts (around 60%) still leave significant room for improvements, highlighting the challenges of generalizing pretrained VLMs to unseen domains that deviate far from the training data.

6.2.3 Domain-Wise Transfer

While domain-level generalization trains on the mixture of source domains, domain-wise transfer performs finer adaptation from exactly one source domain to one target domain. As shown in Table 8, for dataset OfficeHome with domains *Art (A)*, *Clipart (C)*, *Product (P)* and *RealWorld (R)*, knowledge transfer is performed on each source-target pair, leading to 12 subtasks in total. DomainNet includes 6 domains and 30 cross-domain tasks, and their means are reported due to space limits. Please refer to the origin paper for full domain-wise results. On VisDA there is only one synthetic-to-real adaptation task and the mean-class accuracies are reported following previous works. Note that the CLIP’s zero-shot results on tasks with the *same target domain* are *identical* since the model is not trained on the source domain.

From Table 8 we can make a similar observation that the backbone of CLIP’s vision encoder is the main factor that affects the performance. However, the accuracy gap brought by backbone is narrowing with the advances in research works, especially on easier tasks (e.g., AP, CP, RP on OfficeHome and VisDA). Unlike dataset-level generalizations, the differences in method design do not bring significant changes in performance. It is worth noting that CLIP-based UDA methods significantly surpass single-modality baselines, mainly due to the strong zero-shot ability of pretrained CLIP - even CLIP’s zero-shot accuracies easily surpass that of various single-modality methods. However, current CLIP-based UDA methods overly rely on pseudo labels generated by CLIP, leading to *similar* accuracies among tasks with the same target domain. This phenomenon indicates that the source knowledge are still under-explored, thus hardly affecting the transfer performance. Such observation also holds true for SFDA tasks.

7 MULTIMODAL LARGE MODELS

Recent years have witnessed the revolutionary breakthroughs and astonishing capabilities of large language models (LLMs), exemplified by the GPT family [10], [55]. As the language branch plays an important part in VLMs for encoding and aligning textual information, one intuitive

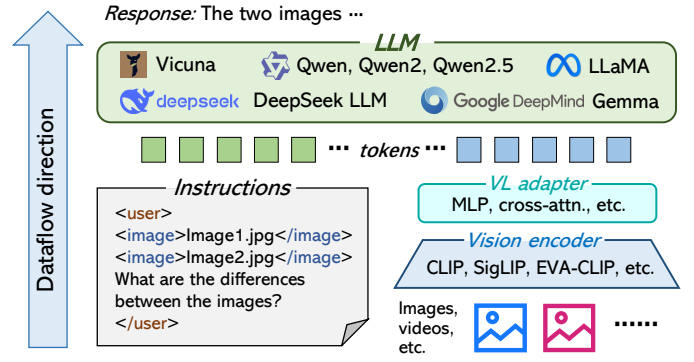


Fig. 8. General structure of MLLMs, composing of a pretrained vision encoder, an LLM and a VL-adapter to connect the two. The inputs include vision-modality materials (images, videos) and textual instructions.

question is that, *is it possible to enhance VLMs by integrating the general-purpose LLMs?* Research efforts have proven the viability of such approach. Furthermore, the resultant LLMs enhanced by pretrained VLMs show strong generalization abilities to a wide range of vision-language tasks, e.g., visual question answering (QA), segmentation, captioning, etc. These large models are summarized as multimodal large language models (MLLMs), due to their capabilities in handling vision and text in various forms. This section first introduces typical structures and training strategies for general MLLMs, then conduct a more detailed review on representative MLLMs.

7.1 MLLM Structure

Current MLLMs are mostly obtained by integrating a vision encoder with an LLM, as shown in Fig. 8. The vision encoder perceives various visual information, including videos, images of different resolutions, and transforms them into high-dimensional features. A lightweight vision-language adapter (VL adapter) aligns the dimension and distribution between the vision and text representation space. The resultant vision tokens are concatenated with text tokens of corresponding instruction data, which are then sent to an LLM to generate the desired outputs. In practice, the tokens may be concatenation of multilingual texts, images of different scales, videos, and potentially more modalities encoded by other encoders [302], [303].

To mitigate modality gap [304] between vision and text tokens, current MLLMs [18], [57], [58], [274] typically adopt the vision branch of pretrained VLMs [11], [252], [269], [275] as the vision encoder. The choice of LLM varies for different research teams, where individual researchers usually adopt public LLMs [277], [278], while AI institutions prefer their own LLMs trained with in-house data [256], [292], [302]. The adapters are usually one- or two-layer MLP.

Training data and strategy is the key factor that decides model performances. Introduced in [18], instruction tuning [305] has been widely adopted for tuning MLLMs to follow human instructions for various tasks. Multimodal datasets, e.g., captioning [264], [266], visual QA [237], [272], [306], image-text pairs [260], [264], [273], are then transformed to instruction-following data using automatic pipelines like GPT [55]. Every of the three components in the

TABLE 9
Summarization of representative multimodal large language models (MLLM). The **F** column represents the MLLM family.

F	Model	Release date	Training datasets	Vision encoder	VL Adapter	LLM	Brief description
DeepSeek	DeepSeek-VL [56]	2024.3	Image-text, image caption, table & charts, web code, scene text OCR, document OCR, text corpus, in-house data, shared gpt4v datasets.	SigLIP [252], SAM-B [253], ViT-Det [254]	two-layer MLP	DeepSeek LLM [52]	Introduces hybrid vision encoder to efficiently handle high resolution inputs. Proposes to preserve the language ability of VLMs by dynamically adjust the ration between vision and text modality data.
	DeepSeek-VL2 [255]	2024.12	Image-text, image captioning, OCR, visual QA, visual grounding, conversations, reasoning, logic, mathematics, textbook, academic questions, code generation, etc.	SigLIP-SO400M [252]	two-layer MLP	DeepSeek MoE [256]	Introduces tiling strategy for processing high-resolution images with different aspect ratios. Replaces the LLM with DeepSeek-MoE with the Multi-head Latent Attention mechanism.
	DeepSeek Janus [257]	2024.10	Image-text paired captions [258], images, text, table & charts, visual generation data.	SigLIP-Large [252]	task-specific adapters	DeepSeek LLM [52]	Decouples the visual encoding process into separate encoders to handle multimodal understanding and generation, respectively.
	DeepSeek Janus-Pro [259]	2025.1	Image caption, table, chart, document understanding data, aesthetic data, and datasets from DeepSeek-VL2.	SigLIP-Large [252]	task-specific adapters	DeepSeek LLM [52]	Improves DeepSeek Janus with extended training data and model scale. Improves the train strategy by focusing on the training of VL-adapter and image head, as well as image-text data for pretraining.
Qwen	Qwen-VL [57]	2023.10	LAION-5B [260], LAION-COCO [261], DataComp [262], Coyo [263], CC12M [264], CC3M [265], SBU [266], COCO Caption [238] and VL annotation data.	OpenCLIP ViT-B and ViT-L	single-layer cross-attn.	Qwen 7B [54]	Qwen-VL introduces a vision encoder, VL adapter and a 3-stage training pipeline to endow Qwen-LM with various multimodal abilities. An addition set of dialogue data incorporates localization and multi-image comprehension abilities into the model.
	Qwen2-VL [268]	2024.10	Image-text pairs, OCR data, interleaved image-text articles, visual QA datasets, video dialogues, image knowledge datasets, etc.	Modified DFN's ViT [269]	single-layer cross-attn.	Qwen2 [53]	Proposes Naive Dynamic Resolution mechanism to process dynamic image resolutions, and Multimodal Rotary Position Embedding (MRPE) to fuse positional information across modalities.
	Qwen2.5-VL [270]	2025.2	Image caption, OCR, pure text, interleaved data, vision QA, video grounding, agent, long video, document, etc.	Modified ViT	two-layer MLP	Qwen2.5 [271]	Building on Qwen-VL2, proposes MRPE aligned to absolute time, native dynamic resolution and redesigned ViT to achieve strong generalization ability across domains without specific fine-tuning.
BLIP	BLIP [16]	2022.2	129M images from COCO, Visual Genome [272], Conceptual Captions [264], CC 12M [264], SBU captions [266], LAION400M [273].	ImageNet ViT-B and ViT-L	None	BERT-base	A multimodal mixture of encoder-decoder structure, composed of aligned unimodal encoders, image-grounded text encoder, and image-grounded text decoder to generate answers.
	BLIP2 [274]	2023.1	The same as BLIP.	CLIP ViT-L, EVA-CLIP ViT-g [275]	Q-Former	OPT [276], FlanT5 [277]	Integrates a lightweight Query Transformer (Q-Transformer) to extract vision features that mitigate the modality gap between frozen image encoder and LLM.
	Instruct-BLIP [17]	2023.6	Datasets including image captioning, visual reasoning, visual QA, image question generation, video QA, image classification, etc.	EVA-CLIP ViT-g/14 [275]	Q-Former with instructions	FlanT5 [277], Vicuna [278]	Introduces instructions to Q-Former of BLIP2 to extract instruction-aware visual features. The training datasets are extended to a wider range of multimodal tasks in instruction tuning format.
	xGen-MM [279]	2024.8	Interleaved dataset mixture, including MINT-1T [280], OBELICS [281], BLIP3-KALE, BLIP3-OCR-200M, BLIP3-GROUNDING-50M, and other public datasets.	CLIP-DFN [269], SigLIP [252]	perceiver resampler [282]	Phi3-mini [283]	Scales up BLIP2 with an ensemble of multimodal interleaved datasets, and replaces Q-Former with a scalable vision token sampler to down sample the encoded any-resolution vision tokens.
LLaVA	LLaVA [18]	2023.4	COCO images [245], filtered CC3M [265], transformed to instruction data using GPT-4.	CLIP ViT-L/14	single linear layer	Vicuna [278]	Generates instruction-following data from multimodal data using GPT-4, then trains a MLLM that connects an image encoder and LLM for multimodal chat, science QA, etc.
	LLaVA-1.5 [284]	2023.10	The same as LLaVA, plus open-knowledge VQA and OCR data.	CLIP ViT-L 336px	two-layer MLP	Vicuna [278]	Improves LLaVA by replacing the VL adapter with a two-layer MLP, improved QA data format, and grid-split the images to up-scale to high resolution visual perception.
	LLaVA-plus [285]	2023.11	LLaVA data and curated tool-use instruction data.	Based on pretrained MLLM weights.			A generalizable agent that uses a large and diverse set of external tools according to users' needs, trained in an end-to-end fashion.
	LLaVA-OneVision [286]	2024.10	Re-captioned description data, document / OCR data, Chinese and language data, visual instruction data, etc.	SigLIP [252]	two-layer MLP	Qwen2 [53]	A family of MLLM to handle single-, multi-image and videos with transfer learning ability across modalities and domains, supported by an AnyRes strategy to handle different resolutions.
	Dynamic-LLaVA [287]	2024.12	The same as LLaVA-1.5.	Based on pretrained MLLM weights.			A dynamic vision-language context sparsification framework to save memory consumption, GPU memory, and inference time, by reducing redundancy in prefill and decoding stages.
PaliGemma	Pali-Gemma [288]	2024.7	Image captioning, visual QA, image segmentation, video, etc. Please refer to the paper for full details.	SigLIP-SO400M [252]	single linear layer	Gemma-2B [289]	A family of MLLM following the PaLI family [20], [290] to handle images of various resolutions, and can be transferred to domain-specific tasks including remote-sensing and segmentation.
	Pali-Gemma 2 [291]	2024.12	Data in PaliGemma, plus text detection, table recognition, molecular structure, optical music score, long caption generation, spatial reasoning, radiography report generation.	SigLIP-SO400M [252]	single linear layer	Gemma 2 [289]	Enhances PaliGemma with improved LLM and a wider range of transfer tasks including OCR data, fine-grained captioning and radiography report generation.
Apple MM	MM1 [292]	2024.4	Mixture of 45% captioned images, 45% interleaved image-text document, 10% text.	ViT trained on DFN [269]	C-Abstractor [293]	Private LLM	Investigates several factors when building a MLLM, e.g., data composition, image encoder, VL adapter, pretrain loss. Builds a family of MLLM with enhanced in-context, multi-image, few-shot abilities.
	MM1.5 [294]	2024.9	Data in MM1, plus OCR data, and newly introduced supervised fine tuning data with optimal ratio.	ViT-H trained on DFN [269]	C-Abstractor [293]	Private LLM	Improves MM1 from the perspective of data composition by exploring optimal data mixture ratios and curation. Two specialized variants, MM1.5-Video and MM1.5-UI are tailored for video and mobile UI understanding.
MiniGPT	MiniGPT-4 [58]	2023.10	Conceptual Captions [264], CC 12M [264], SBU [266], LAION400M [273], and a newly curated dataset.	EVA-CLIP ViT-g [275] as in [274]	single linear layer	Vicuna [278]	Demonstrates that by connecting a pretrained vision encoder and LLM with one projection layer, the resultant MLLM can achieve similar multimodal abilities as in GPT4.
	MiniGPT-v2 [295]	2023.11	CC3M [265], SBU [266], LAION400M [273], vision QA datasets, LLaVA instruction data [18], Flickr30k entities [296], KOSMOS-2 [297]	EVA-CLIP ViT-g [275]	single linear layer	LLaMA2-chat [298]	Builds a unified interface for various vision-language tasks, i.e., a MLLM, on pretrained LLM and vision encoder using instruction-format data.
ByteDance	Seed1.5-VL [299]	2025.6	Image-text, OCR, visual grounding, STEM, video, GUI, 3D Spatial, long Chain-of-Thought data, and RLHF.	private Seed-ViT	two-layer MLP	private Seed1.5-LLM	Latest 20B MLLM obtained with high-quality synthetic vision-language data and hybrid training infrastructure to achieve leading multimodal abilities with reduced expenses.
	BAGEL [300]	2025.6	Pure text, image-text, vision-text interleaved data, video, reasoning-augmented data (e.g., image manipulation, conceptual edits).	SigLIP2-so400m/14 [301]	two-layer MLP	Qwen2.5 LLM [271]	An unified 7B MLLM supporting general understanding and generation tasks across modalities (e.g., free-form visual manipulation) with mixture-of-Transformer structure.

MLLM is trainable. Considering their parameter count and functionalities are fundamentally different, current methods adopt multi-stage training paradigm to adjust different components with different data and tuning strategy. Finer discussions on the training data, structure, training strategy of different MLLMs are provided in the following section.

7.2 General-Purpose MLLMs

Table 9 introduces recent advances (2022-2025.6) of MLLMs. Note that the **Training datasets** column are high-level summarization of the vast training data. For more details please refer to the original papers.

7.2.1 Research-Oriented MLLMs

BLIP family are one of the earliest efforts in building MLLMs. Similar to its ancestor CLIP [11], BLIP builds on web-crawled image-text data but with denoising. Unlike the two-tower structure of CLIP, BLIP introduces mixture of encoder-decoder trained with multiple cross-modality alignment losses. BLIP is based on single-modality vision encoders (ImageNet-pretrained ViT) and small-scale language decoders, which is more like a VLM rather than a MLLM. One year later, the same research team proposes BLIP2 [274], a MLLM following the structure in Fig. 8 using EVA-CLIP [275] vision encoder and OPT [276], FlanT5 [277] as LLM. The VL-adapter is a newly-proposed Query Transformer (Q-Former) to mitigate modality gap. BLIP2 adopts a two-stage training paradigm to endow Q-Former with multimodal representation abilities, while keeping the pre-trained vision encoder and LLM frozen. Based on BLIP2, InstructBLIP [17] further introduces instruction-tuning for the training of Q-Former to obtain a general vision-language task solver. Recently, a family of MLLM termed xGen-MM [279], also known as BLIP-3, have been released. With high-quality interleaved datasets and improved VL-adapter [282], BLIP-3 achieves competitive performances on various vision-language tasks.

LLaVA family are another group of instruction-tuned MLLMs built upon public VLMs and LLMs. Aiming to conduct visual instruction tuning, Liu *et al.* [18] propose to transform available image-text data into instruction-following data by prompting language-only LLMs, e.g., GPT4 [55]. Based on instruction data, LLaVA [18] is obtained by tuning an LLM Vicuna [278] and a single projection layer in a two-step paradigm. The vision encoder from pretrained CLIP remains frozen all the time. With systematic exploration on the training principles of LLaVA, including data composition, data scale, choices in MLLM components, a more powerful variant LLaVA-1.5 [284] is obtained. Building on the success of LLaVA family, follow-up researches have explored multimodal agents [285] that operate on various tools to complete users' complex requirements, or try to reduce computation resource consumption when building MLLMs [287]. LLaVA-OneVision [286] summarizes the extensions and advances of the LLaVA-NeXT series achieved by handling high-resolution images, improving data quality and LLM capabilities.

MiniGPT family are early efforts to build MLLMs that reach the capabilities of private GPT4, using public base models and datasets. By aligning the vision encoder and

Q-Former in BLIP2 [274] with the Vicuna [278] LLM using instruction-data-tuned VL adapter, MiniGPT-4 [58] achieves comparable abilities with GPT4. Similar observations are obtained in MiniGPT-v2 [295] with different base models. Differently, MiniGPT-v2 also updates parameters in the LLM with a coarse-to-fine three-step training scheme.

7.2.2 Commercial MLLMs

Although commercial and research-oriented MLLMs share similar structures, the abundant high-quality in-house training data and computation resources in large AI institutions allow more in-depth tuning of MLLM parameters, resulting in more powerful and practical commercial MLLMs.

Qwen family are based on Alibaba's Qwen-LLM series [53], [54], [271]. Qwen-VL [57] is trained with three stages: general pretraining of *vision encoder and VL adapter* with weakly-supervised image-text pairs, multi-task pre-training of *the whole model* with high-quality interleaved multi-task vision-language data, and supervised fine-tuning of *LLM and VL adapter* with instruction-following data. Instead of tuning a small part of MLLM as in research-oriented MLLMs, we can observe that commercial MLLMs generally tune the whole model progressively for better task generalization. One critical challenge in building MLLMs is how to handle various input image resolutions. Qwen2-VL [268] tackles this by introducing dynamic resolution support [307] and Multimodal Rotary Position Embedding (M-RoPE) to encode position information of multimodal inputs. Qwen2.5-VL [270] further advances the MLLM on various modalities, e.g., video, by considering the information in spatial and temporal dimensions.

DeepSeek family are famous for their superior performance and low training costs. DeepSeek-VL [56] focuses on the preservation of language abilities when building MLLMs. To achieve this, a local-to-global training strategy is applied, where the VL-adapter is first warmed-up, followed by joint pretraining of LLM and VL-adapter. Finally, the whole model is fine-tuned with instruction-following data. Specifically, the ratio of language and multimodal data in step 2 is 7:3 to alleviate the loss of language abilities. DeepSeek-VL2 [255] replaces the DeepSeek-LLM [52] with DeepSeek-MoE [256] for more efficient inference. A dynamic tiling vision encoding strategy is proposed to handle high-resolution image inputs, which divides large images into tiles. DeepSeek-Janus [257] is a unified autoregressive framework with multimodal understanding and generation ability. Slightly different from Fig. 8, Janus introduces separate understanding/generation encoders and decoders to handle both tasks in isolation. The training procedure is similar to DeepSeek-VL, except the generation decoder is also trainable in step 1, and only the understanding encoder is trainable in step 3. The improved version Janus-Pro [259] extends the data and model scale. The training strategy is improved by lengthening train step 1 on pure images, and drop these images in step 2 to focus on image-text data.

PaliGemma family are open-source variants of the Gemini series [308], [309], built on SigLIP [252] and public Gemma [289] LLM. PaliGemma follows a pretrain - transfer strategy. The whole model is first pretrained on extensive multimodal tasks and finer trained with increased image

resolutions. The obtained base model is then fine-tuned to transfer towards specific goals.

MM family are private MLLMs developed by Apple. MM focus on exploring data-centric training strategies and reveal several properties when tuning MLLM with different data composition in different stages. For example, MM1.5 [294] adopts a three-step training recipe: (1) Large-scale pretraining on low-resolution image-text pairs and text-only data. (2) Continual pretraining with high-resolution OCR data. (3) Supervised fine-tuning with mixture of data. In each stage, the impacts of data mixture ratios and each data type are explored in detail.

Seed family developed by ByteDance are among the latest advances in MLLM. SeedVL-1.5 [299] is built with pretraining and post-training steps. In pretraining, the VL adapter is first warmed-up, followed by full training of the whole model with carefully decided tasks and data mixture ratio. In post-training, the MLLM is equipped with instruction-following and understanding abilities by finely-constructed supervised fine-tuning data, e.g., chain-of-thought data, and boosted by reinforcement learning with human feedback [310] or verifiable rewards [311].

7.3 Discussions

The development of general MLLMs is a research focus and heated topic in the community, with new generations of MLLM being released monthly. As an extension of VLM, MLLM obtains strong generalization ability across various modalities and language-image tasks by connecting language-aligned image encoders with powerful LLMs. They generally follow the structure in Section 7.1 with progressively improved base models and training corpus. A part of popular or representative MLLMs are introduced in this survey, categorized into research (Section 7.2.1) or commercial Section 7.2.2 purpose MLLMs. It is expected that more advanced MLLMs with more general abilities, e.g., multimodal agents, be released in the future.

8 FUTURE DIRECTIONS

The research on the vision-language field, integrating with advances in large-scale training, is a promising direction in developing modern intelligent systems. The goal is to build general-purpose VLMs generalizable to any data distribution and multimodal tasks. While current models have shown astonishing capabilities in certain tasks, they are still limited by out-of-distribution data and tasks. For example, *none* of the cutting-edge models can solve a most recent coding benchmark [312]. Therefore, there is still plenty of room for improvements. Some promising directions are discussed as following.

(1) Generalize under black-box setting. Most of the current advances aiming to transfer a general VLM to downstream tasks are based on fully open-source frameworks like CLIP [11]. However, the most advanced achievements are generally private commercial services accessed by API, e.g., methods introduced in Section 7.2.2. The pretrained models in data-sensitive fields, e.g., medical, military, may also constrain the accessibility of their structures and weights. The goal is to build a target model by transferring the

knowledge behind the API. Current attempts [38], [114] rely on model distillation, but their assumptions on label distribution, accessibility of final-layer features, etc., limit their practicability. A general VLM transfer technique without accessing the base model's weights and structures is needed.

(2) Generalize efficiently. Although current VLM generalization methods adopt various parameter-efficient fine-tuning (PEFT) strategies, e.g., prompt tuning in Section 3, adapters in Section 5.1.1, their efficiencies are limited to the parameter dimension, but neglecting other factors like memory, GPU consumption. For example, the training resources required by typical prompt tuning in [31] scales with increasing category numbers, and requires computing gradient on all VLM parameters. More dimensions of transfer efficiency need to be considered. One recent attempt [105] adopts quantification to achieve generalization with memory efficiency. In order to deploy and adapt more powerful VLMs on resource-constrained scenarios like edge devices, researches on a larger spectrum of generalization efficiency are desirable.

(3) General multimodal agents. Agents are general helpers that aid the user to accomplish goals by operating on various tools, e.g., coding [313]. With general multimodal ability, agents are expected to handle a wider range of tasks. There have been pioneering studies on multimodal agents [314], but they often rely on pretrained LLMs, whose generalizability across modalities, novel tasks and domains remains underexplored. The daily tasks in human society span across various domains and involve substantial operations. Trained and evaluated with limited data, the agents' behavior in open real world cannot be well assessed, and their lack of true generalizability may lead to undesirable performances. Therefore, investigating and improving the generalizability of large-model-driven multimodal agents is a promising research direction. This can be further extended to applications on embodied intelligence.

(4) Native vision-language structure. Current MLLMs are built by connecting pretrained LLMs with language-aligned vision encoders. As the number and complexity of input modalities increase, the demands for the encoders - vision, audio ones, and probably more - are making the overall structure overly complicated and less scalable. The computation limitations posed by the attention mechanism also hinder the models' generalization to the infinite open world. Is it possible to construct an all-in-one multimodal model from scratch, which supports any modality natively?

(5) Generalizability of vision-language training data. When building general MLLMs, the quality and scale of multimodal training data play a dominant role. However, it is unclear that whether all the data contribute to the generalization of the target model. There have been researches [94], [156] showing that training with domain and modality-aligned data can improve adaptation performance. Extensive researches [259], [292], [294] have also investigated the optimal mixture proportion and composition of vision-language data to achieve a balance between model performance and data efficiency. In the era of Internet, the quality of web-crawled data is not guaranteed, and polluted data may even introduce negative effects during training. Therefore, the community is in need of an automatic framework that determine the quality and usefulness of raw

multimodal data given target tasks.

9 CONCLUSION

This paper systematically reviews recent advances in generalizing and adapting vision-language models (VLMs) to novel tasks and domains. Based on the mostly adopted VLM structures, this survey divides the generalization methods into prompt-based, parameter-based and feature-based methods, according to the adapted model component. The reviewed methods are introduced based on the traditional transfer learning settings, revealing how these transfer problems are being solved in the era of vision-language studies. Thorough performance comparisons among reviewed methods on standard generalization benchmarks are provided. Following advances of large-scale training, this survey also includes most up-to-date and representative multimodal large language models (MLLMs), which enhances the generalizability of vision-language systems with powerful large language models. Finally, current challenges and future directions of vision-language researches are discussed to inspire further advancements of the community.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
- [2] 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.
- [3] 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.
- [4] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.
- [5] 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.
- [6] 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.
- [7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li *et al.*, "Regionclip: Region-based language-image pretraining," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 793–16 803.
- [13] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.
- [14] 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.
- [15] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.
- [16] 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.
- [17] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," *arXiv preprint arXiv:2305.06500*, 2023.
- [18] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.
- [19] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.
- [20] X. Chen, J. Djolonga, P. Padlewski, B. Mustafa, S. Changpinyo, J. Wu, C. R. Ruiz, S. Goodman, X. Wang, Y. Tay *et al.*, "Pali-x: On scaling up a multilingual vision and language model," *arXiv preprint arXiv:2305.18565*, 2023.
- [21] T. Nguyen, M. Wallingford, S. Santy, W.-C. Ma, S. Oh, L. Schmidt, P. W. W. Koh, and R. Krishna, "Multilingual diversity improves vision-language representations," *Advances in Neural Information Processing Systems*, vol. 37, pp. 91 430–91 459, 2024.
- [22] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.
- [23] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008, pp. 722–729.
- [24] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.
- [25] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.
- [26] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.
- [27] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*. Springer, 2022, pp. 493–510.
- [28] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.
- [29] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Larousilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *International conference on machine learning*. PMLR, 2019, pp. 2790–2799.
- [30] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.
- [31] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
- [32] —, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.
- [33] X. Li, Y. Fang, M. Liu, Z. Ling, Z. Tu, and H. Su, "Distilling large vision-language model with out-of-distribution generalizability,"

- in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2492–2503.
- [34] C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang, “Domain adaptation via prompt learning,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [35] X. Li, Y. Li, Z. Du, F. Li, K. Lu, and J. Li, “Split to merge: Unifying separated modalities for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 364–23 374.
- [36] Z. Du, X. Li, F. Li, K. Lu, L. Zhu, and J. Li, “Domain-agnostic mutual prompting for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 375–23 384.
- [37] S. Bai, Y. Zhang, W. Zhou, Z. Luan, and B. Chen, “Soft prompt generation for domain generalization,” in *European Conference on Computer Vision*. Springer, 2024, pp. 434–450.
- [38] S. Addepalli, A. R. Asokan, L. Sharma, and R. V. Babu, “Leveraging vision-language models for improving domain generalization in image classification,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 922–23 932.
- [39] Z. Chen, W. Wang, Z. Zhao, F. Su, A. Men, and H. Meng, “Practicaldgl: Perturbation distillation on vision-language models for hybrid domain generalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 501–23 511.
- [40] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, “Test-time prompt tuning for zero-shot generalization in vision-language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 274–14 289, 2022.
- [41] A. Karmanov, D. Guan, S. Lu, A. El Saddik, and E. Xing, “Efficient test-time adaptation of vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 162–14 171.
- [42] C.-M. Feng, K. Yu, Y. Liu, S. Khan, and W. Zuo, “Diverse data augmentation with diffusions for effective test-time prompt tuning,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2704–2714.
- [43] 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.
- [44] Y. Du, Z. Liu, J. Li, and W. X. Zhao, “A survey of vision-language pre-trained models,” *arXiv preprint arXiv:2202.10936*, 2022.
- [45] A. Ghosh, A. Acharya, S. Saha, V. Jain, and A. Chadha, “Exploring the frontier of vision-language models: A survey of current methodologies and future directions,” *arXiv preprint arXiv:2404.07214*, 2024.
- [46] J. Cha, K. Lee, S. Park, and S. Chun, “Domain generalization by mutual-information regularization with pre-trained models,” in *European conference on computer vision*. Springer, 2022, pp. 440–457.
- [47] Y. Shu, X. Guo, J. Wu, X. Wang, J. Wang, and M. Long, “Clipood: Generalizing clip to out-of-distributions,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 31 716–31 731.
- [48] C. Oh, H. Lim, M. Kim, D. Han, S. Yun, J. Choo, A. Hauptmann, Z.-Q. Cheng, and K. Song, “Towards calibrated robust fine-tuning of vision-language models,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 12 677–12 707, 2024.
- [49] Y. Zhang, W. Zhu, H. Tang, Z. Ma, K. Zhou, and L. Zhang, “Dual memory networks: A versatile adaptation approach for vision-language models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 28 718–28 728.
- [50] K. Weiss, T. M. Khoshgoftaar, and D. Wang, “A survey of transfer learning,” *Journal of Big data*, vol. 3, pp. 1–40, 2016.
- [51] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning,” in *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4–7, 2018, Proceedings, Part III* 27. Springer, 2018, pp. 270–279.
- [52] X. Bi, D. Chen, G. Chen, S. Chen, D. Dai, C. Deng, H. Ding, K. Dong, Q. Du, Z. Fu *et al.*, “Deepseek llm: Scaling open-source language models with longtermism,” *arXiv preprint arXiv:2401.02954*, 2024.
- [53] Q. Team, “Qwen2 technical report,” *arXiv preprint arXiv:2407.10671*, 2024.
- [54] Qwen, “Introducing qwen-7b: Open foundation and human-aligned models (of the state-of-the-arts),” <https://github.com/zsc19/Qwen-7B/tree/main?tab=readme-ov-file>, 2023.
- [55] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, “Gpt-4 technical report,” *arXiv preprint arXiv:2303.08774*, 2023.
- [56] H. Lu, W. Liu, B. Zhang, B. Wang, K. Dong, B. Liu, J. Sun, T. Ren, Z. Li, H. Yang *et al.*, “Deepseek-vl: towards real-world vision-language understanding,” *arXiv preprint arXiv:2403.05525*, 2024.
- [57] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, “Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond,” 2023. [Online]. Available: <https://arxiv.org/abs/2308.12966>
- [58] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigt-4: Enhancing vision-language understanding with advanced large language models,” *arXiv preprint arXiv:2304.10592*, 2023.
- [59] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, “Contrastive learning of medical visual representations from paired images and text,” in *Machine learning for healthcare conference*. PMLR, 2022, pp. 2–25.
- [60] L. Torrey and J. Shavlik, “Transfer learning,” in *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.
- [61] M. Long, Z. Cao, J. Wang, and M. I. Jordan, “Conditional adversarial domain adaptation,” *Advances in neural information processing systems*, vol. 31, 2018.
- [62] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, “Transfer independently together: A generalized framework for domain adaptation,” *IEEE transactions on cybernetics*, vol. 49, no. 6, pp. 2144–2155, 2018.
- [63] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, “Maximum density divergence for domain adaptation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3918–3930, 2020.
- [64] J. Li, Z. Du, L. Zhu, Z. Ding, K. Lu, and H. T. Shen, “Divergence-agnostic unsupervised domain adaptation by adversarial attacks,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8196–8211, 2021.
- [65] J. Hu, J. Lu, and Y.-P. Tan, “Deep transfer metric learning,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 325–333.
- [66] H. Yan, Y. Ding, P. Li, Q. Wang, Y. Xu, and W. Zuo, “Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2272–2281.
- [67] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.
- [68] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, “Domain generalization: A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.
- [69] Z. Du, J. Li, K. Lu, L. Zhu, and Z. Huang, “Learning transferrable and interpretable representations for domain generalization,” in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3340–3349.
- [70] K. Muandet, D. Balduzzi, and B. Schölkopf, “Domain generalization via invariant feature representation,” in *International conference on machine learning*. PMLR, 2013, pp. 10–18.
- [71] Z. Du, J. Li, L. Zuo, L. Zhu, and K. Lu, “Energy-based domain generalization for face anti-spoofing,” in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1749–1757.
- [72] Y. Li, X. Tian, M. Gong, Y. Liu, T. Liu, K. Zhang, and D. Tao, “Deep domain generalization via conditional invariant adversarial networks,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 624–639.
- [73] S. Zhao, M. Gong, T. Liu, H. Fu, and D. Tao, “Domain generalization via entropy regularization,” *Advances in neural information processing systems*, vol. 33, pp. 16 096–16 107, 2020.
- [74] D. Li, Y. Yang, Y.-Z. Song, and T. Hospedales, “Learning to generalize: Meta-learning for domain generalization,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
- [75] Y. Balaji, S. Sankaranarayanan, and R. Chellappa, “Metareg: Towards domain generalization using meta-regularization,” *Advances in neural information processing systems*, vol. 31, 2018.

- [76] J. Liang, R. He, and T. Tan, "A comprehensive survey on test-time adaptation under distribution shifts," *International Journal of Computer Vision*, vol. 133, no. 1, pp. 31–64, 2025.
- [77] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," *arXiv preprint arXiv:2006.10726*, 2020.
- [78] J. Li, Z. Yu, Z. Du, L. Zhu, and H. T. Shen, "A comprehensive survey on source-free domain adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [79] X. Li, Z. Du, J. Li, L. Zhu, and K. Lu, "Source-free active domain adaptation via energy-based locality preserving transfer," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 5802–5810.
- [80] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM computing surveys (csur)*, vol. 53, no. 3, pp. 1–34, 2020.
- [81] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7402–7411.
- [82] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European conference on computer vision*. Springer, 2022, pp. 709–727.
- [83] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 113–19 122.
- [84] Y. Zang, W. Li, K. Zhou, C. Huang, and C. C. Loy, "Unified vision and language prompt learning," *arXiv preprint arXiv:2210.07225*, 2022.
- [85] Y. Xing, Q. Wu, D. Cheng, S. Zhang, G. Liang, P. Wang, and Y. Zhang, "Dual modality prompt tuning for vision-language pre-trained model," *IEEE Transactions on Multimedia*, vol. 26, pp. 2056–2068, 2023.
- [86] E. Cho, J. Kim, and H. J. Kim, "Distribution-aware prompt tuning for vision-language models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 22 004–22 013.
- [87] H. Bahng, A. Jahanian, S. Sankaranarayanan, and P. Isola, "Exploring visual prompts for adapting large-scale models," *arXiv preprint arXiv:2203.17274*, 2022.
- [88] H. Salman, A. Ilyas, L. Engstrom, S. Vemprala, A. Madry, and A. Kapoor, "Unadversarial examples: Designing objects for robust vision," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 270–15 284, 2021.
- [89] J. Zhang, S. Wu, L. Gao, H. T. Shen, and J. Song, "Dept: Decoupled prompt tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 12 924–12 933.
- [90] H. Li, L. Wang, C. Wang, J. Jiang, Y. Peng, and G. Long, "Dpc: Dual-prompt collaboration for tuning vision-language models," *arXiv preprint arXiv:2503.13443*, 2025.
- [91] Z. Zhou, M. Yang, J.-X. Shi, L.-Z. Guo, and Y.-F. Li, "Decoop: robust prompt tuning with out-of-distribution detection," *arXiv preprint arXiv:2406.00345*, 2024.
- [92] M. Wortsman, G. Ilharco, J. W. Kim, M. Li, S. Kornblith, R. Roelofs, R. G. Lopes, H. Hajishirzi, A. Farhadi, H. Namkoong et al., "Robust fine-tuning of zero-shot models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7959–7971.
- [93] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6757–6767.
- [94] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned gradient for prompt tuning," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 15 659–15 669.
- [95] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 15 190–15 200.
- [96] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? generating customized prompts for zero-shot image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 691–15 701.
- [97] M. U. Khattak, M. F. Naeem, M. Naseer, L. Van Gool, and F. Tombari, "Learning to prompt with text only supervision for vision-language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 4230–4238.
- [98] M. Singha, H. Pal, A. Jha, and B. Banerjee, "Ad-clip: Adapting domains in prompt space using clip," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4355–4364.
- [99] S. Bose, A. Jha, E. Fini, M. Singha, E. Ricci, and B. Banerjee, "Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5542–5552.
- [100] V. H. Phan, T. L. Tran, Q. Tran, and T. Le, "Enhancing domain adaptation through prompt gradient alignment," *Advances in Neural Information Processing Systems*, vol. 37, pp. 45 518–45 551, 2024.
- [101] S. Yan, C. Luo, Z. Yu, and Z. Ge, "Enhancing vision-language models generalization via diversity-driven novel feature synthesis," *arXiv preprint arXiv:2405.02586*, 2024.
- [102] D. Cheng, Z. Xu, X. Jiang, N. Wang, D. Li, and X. Gao, "Disentangled prompt representation for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 595–23 604.
- [103] M. Singha, A. Jha, S. Bose, A. Nair, M. Abdar, and B. Banerjee, "Unknown prompt the only lacuna: Unveiling clip's potential for open domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 309–13 319.
- [104] Z. Xiao, J. Shen, M. M. Derakhshani, S. Liao, and C. G. Snoek, "Any-shift prompting for generalization over distributions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 849–13 860.
- [105] T. Hao, X. Ding, J. Feng, Y. Yang, H. Chen, and G. Ding, "Quantized prompt for efficient generalization of vision-language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 54–73.
- [106] J. Zhang, X. Ma, S. Guo, P. Li, W. Xu, X. Tang, and Z. Hong, "Amend to alignment: decoupled prompt tuning for mitigating spurious correlation in vision-language models," in *Forty-first International Conference on Machine Learning*, 2024.
- [107] Y. Zang, H. Goh, J. Susskind, and C. Huang, "Overcoming the pitfalls of vision-language model finetuning for ood generalization," *arXiv preprint arXiv:2401.15914*, 2024.
- [108] S. Bose, M. Singha, A. Jha, S. Mukhopadhyay, and B. Banerjee, "Meta-learning to teach semantic prompts for open domain generalization in vision-language models," *Transactions on Machine Learning Research*, 2025.
- [109] S. Bai, M. Zhang, W. Zhou, S. Huang, Z. Luan, D. Wang, and B. Chen, "Prompt-based distribution alignment for unsupervised domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 2, 2024, pp. 729–737.
- [110] K. Shi, J. Lu, Z. Fang, and G. Zhang, "Clip-enhanced unsupervised domain adaptation with consistency regularization," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.
- [111] —, "Unsupervised domain adaptation enhanced by fuzzy prompt learning," *IEEE Transactions on Fuzzy Systems*, 2024.
- [112] H. Chen, X. Han, Z. Wu, and Y.-G. Jiang, "Multi-prompt alignment for multi-source unsupervised domain adaptation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 74 127–74 139, 2023.
- [113] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt distribution learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5206–5215.
- [114] S. Park, J. Jeong, Y. Kim, J. Lee, and N. Lee, "Zip: An efficient zeroth-order prompt tuning for black-box vision-language models," *arXiv preprint arXiv:2504.06838*, 2025.
- [115] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. S. Yu, "Generalizing to unseen domains: A survey on domain generalization," *IEEE transactions on knowledge and data engineering*, vol. 35, no. 8, pp. 8052–8072, 2022.
- [116] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 556–12 565.
- [117] K. Zhou, Y. Yang, T. Hospedales, and T. Xiang, "Learning to generate novel domains for domain generalization," in *Computer vision—ECCV 2020: 16th European conference, glasgow, UK, August 23–28, 2020, proceedings, part XVI 16*. Springer, 2020, pp. 561–578.
- [118] P. Li, D. Li, W. Li, S. Gong, Y. Fu, and T. M. Hospedales, "A simple feature augmentation for domain generalization," in *Proceedings*

- of the IEEE/CVF international conference on computer vision, 2021, pp. 8886–8895.
- [119] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang, “Moment matching for multi-source domain adaptation,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1406–1415.
- [120] X. Li, J. Li, F. Li, L. Zhu, and K. Lu, “Agile multi-source-free domain adaptation,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 12, 2024, pp. 13 673–13 681.
- [121] C. E. Shannon, “A mathematical theory of communication,” *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
- [122] J. Abdul Samadh, M. H. Gani, N. Hussein, M. U. Khattak, M. M. Naseer, F. Shahbaz Khan, and S. H. Khan, “Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 80 396–80 413, 2023.
- [123] X. Ma, J. Zhang, S. Guo, and W. Xu, “Swapprompt: Test-time prompt adaptation for vision-language models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 65 252–65 264, 2023.
- [124] H. S. Yoon, E. Yoon, J. T. J. Tee, M. Hasegawa-Johnson, Y. Li, and C. D. Yoo, “C-tp: Calibrated test-time prompt tuning for vision-language models via text feature dispersion,” *arXiv preprint arXiv:2403.14119*, 2024.
- [125] Z. Xiao, S. Yan, J. Hong, J. Cai, X. Jiang, Y. Hu, J. Shen, Q. Wang, and C. G. Snoek, “Dyaprompt: Dynamic test-time prompt tuning,” *arXiv preprint arXiv:2501.16404*, 2025.
- [126] L. Sheng, J. Liang, Z. Wang, and R. He, “R-tp: Improving adversarial robustness of vision-language models through test-time prompt tuning,” in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 29 958–29 967.
- [127] A. Sharifdeen, M. A. Munir, S. Baliah, S. Khan, and M. H. Khan, “O-tp: Orthogonality constraints for calibrating test-time prompt tuning in vision-language models,” *arXiv preprint arXiv:2503.12096*, 2025.
- [128] T. Lüddecke and A. Ecker, “Image segmentation using text and image prompts,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 7086–7096.
- [129] V. Vedit, M. Engilberge, and M. Salzmann, “Clip the gap: A single domain generalization approach for object detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3219–3229.
- [130] R. Jiang, L. Liu, and C. Chen, “Clip-count: Towards text-guided zero-shot object counting,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4535–4545.
- [131] Z. Lai, N. Vesdapunt, N. Zhou, J. Wu, C. P. Huynh, X. Li, K. K. Fu, and C.-N. Chuah, “Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 16 155–16 165.
- [132] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.
- [133] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, “On large-batch training for deep learning: Generalization gap and sharp minima,” *arXiv preprint arXiv:1609.04836*, 2016.
- [134] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson, “Averaging weights leads to wider optima and better generalization,” in *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*. Association For Uncertainty in Artificial Intelligence (AUAI), 2018, pp. 876–885.
- [135] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, “Swad: Domain generalization by seeking flat minima,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 405–22 418, 2021.
- [136] X. Mao, Y. Chen, X. Jia, R. Zhang, H. Xue, and Z. Li, “Context-aware robust fine-tuning,” *International Journal of Computer Vision*, vol. 132, no. 5, pp. 1685–1700, 2024.
- [137] S. Jain, S. Addepalli, P. K. Sahu, P. Dey, and R. V. Babu, “Dart: Diversify-aggregate-repeat training improves generalization of neural networks,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 048–16 059.
- [138] D. Osowiecki, M. Noori, G. A. V. Hakim, M. Yazdanpanah, A. Bahri, M. Cheraghalikhani, S. Dastani, F. Bezaee, I. B. Ayed, and C. Desrosiers, “Watt: Weight average test-time adaptation of clip,” *arXiv preprint arXiv:2406.13875*, 2024.
- [139] S. Goyal, A. Kumar, S. Garg, Z. Kolter, and A. Raghunathan, “Finetune like you pretrain: Improved finetuning of zero-shot vision models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 338–19 347.
- [140] G. Nam, B. Heo, and J. Lee, “Lipsum-ft: Robust fine-tuning of zero-shot models using random text guidance,” *arXiv preprint arXiv:2404.00860*, 2024.
- [141] J. Liang, L. Sheng, Z. Wang, R. He, and T. Tan, “Realistic unsupervised clip fine-tuning with universal entropy optimization,” in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 29 667–29 681.
- [142] J. Liang, D. Hu, and J. Feng, “Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation,” in *International conference on machine learning*. PMLR, 2020, pp. 6028–6039.
- [143] B. Nguyen, S. Uhlich, F. Cardinaux, L. Mauch, M. Edraki, and A. Courville, “Saft: Towards out-of-distribution generalization in fine-tuning,” in *European Conference on Computer Vision*. Springer, 2024, pp. 138–154.
- [144] Z. Wang, N. Codella, Y.-C. Chen, L. Zhou, J. Yang, X. Dai, B. Xiao, H. You, S.-F. Chang, and L. Yuan, “Clip-td: Clip targeted distillation for vision-language tasks,” *arXiv preprint arXiv:2201.05729*, 2022.
- [145] L. Luo, X. Wang, B. Zi, S. Zhao, X. Ma, and Y.-G. Jiang, “Adversarial prompt distillation for vision-language models,” *arXiv preprint arXiv:2411.15244*, 2024.
- [146] Y. Xuan, W. Chen, S. Yang, D. Xie, L. Lin, and Y. Zhuang, “Distilling vision-language foundation models: A data-free approach via prompt diversification,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 4928–4938.
- [147] Z. Huang, A. Zhou, Z. Ling, M. Cai, H. Wang, and Y. J. Lee, “A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 685–11 695.
- [148] A. Zhou, J. Wang, Y.-X. Wang, and H. Wang, “Distilling out-of-distribution robustness from vision-language foundation models,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 32 938–32 957, 2023.
- [149] Z. Li, X. Li, X. Fu, X. Zhang, W. Wang, S. Chen, and J. Yang, “Promptkd: Unsupervised prompt distillation for vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 617–26 626.
- [150] M. Mistretta, A. Baldrati, M. Bertini, and A. D. Bagdanov, “Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation,” in *European Conference on Computer Vision*. Springer, 2024, pp. 459–477.
- [151] J. Zhu, Y. Chen, and L. Wang, “Clip the divergence: Language-guided unsupervised domain adaptation,” *arXiv preprint arXiv:2407.01842*, 2024.
- [152] W. Zhou and Z. Zhou, “Unsupervised domain adaption harnessing vision-language pre-training,” *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- [153] J. Lee, D. Das, M. Hayat, S. Choi, K. Hwang, and F. Porikli, “Customkd: Customizing large vision foundation for edge model improvement via knowledge distillation,” *arXiv preprint arXiv:2503.18244*, 2025.
- [154] S. Zhao, X. Wang, L. Zhu, and Y. Yang, “Test-time adaptation with clip reward for zero-shot generalization in vision-language models,” in *12th International Conference on Learning Representations, ICLR 2024*, 2024.
- [155] S. Mishra, J. Silva-Rodriguez, I. B. Ayed, M. Pedersoli, and J. Dolz, “Words matter: Leveraging individual text embeddings for code generation in clip test-time adaptation,” *arXiv preprint arXiv:2411.17002*, 2024.
- [156] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan, “Towards stable test-time adaptation in dynamic wild world,” *arXiv preprint arXiv:2302.12400*, 2023.
- [157] Y. Yu, S. Shin, S. Back, M. Ko, S. Noh, and K. Lee, “Domain-specific block selection and paired-view pseudo-labeling for on-line test-time adaptation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 22 723–22 732.
- [158] G. A. V. Hakim, D. Osowiecki, M. Noori, M. Cheraghalikhani, A. Bahri, M. Yazdanpanah, I. B. Ayed, and C. Desrosiers, “Clipart: Adaptation of clip to new domains at test time,” in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 7092–7101.

- [159] R. Imam, H. Gani, M. Huzaifa, and K. Nandakumar, "Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 5449–5459.
- [160] K. Tanwisuth, S. Zhang, H. Zheng, P. He, and M. Zhou, "Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 33 816–33 832.
- [161] D. Chen, K. Patwari, Z. Lai, S.-c. Cheung, and C.-N. Chuah, "Empowering source-free domain adaptation with mllm-driven curriculum learning," *arXiv preprint arXiv:2405.18376*, 2024.
- [162] S. Tarashima, X. Shu, and N. Tagawa, "Vilaad: Enhancing" attracting and dispersing" source-free domain adaptation with vision-and-language model," *arXiv preprint arXiv:2503.23529*, 2025.
- [163] G. Zara, A. Conti, S. Roy, S. Lathuilière, P. Rota, and E. Ricci, "The unreasonable effectiveness of large language-vision models for source-free video domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 307–10 317.
- [164] S. Tang, W. Su, M. Ye, and X. Zhu, "Source-free domain adaptation with frozen multimodal foundation model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 711–23 720.
- [165] M. Zhan, Z. Wu, R. Hu, P. Hu, H. T. Shen, and X. Zhu, "Towards dynamic-prompting collaboration for source-free domain adaptation," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 1643–1651.
- [166] W. Zhang, L. Shen, and C.-S. Foo, "Source-free domain adaptation guided by vision and vision-language pre-training," *International Journal of Computer Vision*, vol. 133, no. 2, pp. 844–866, 2025.
- [167] S. Tang, W. Su, Y. Gan, M. Ye, J. Zhang, and X. Zhu, "Proxy denoising for source-free domain adaptation," *arXiv preprint arXiv:2406.01658*, 2024.
- [168] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.
- [169] Y. Li, N. Wang, J. Shi, J. Liu, and X. Hou, "Revisiting batch normalization for practical domain adaptation," *arXiv preprint arXiv:1603.04779*, 2016.
- [170] S. Schneider, E. Rusak, L. Eck, O. Bringmann, W. Brendel, and M. Bethge, "Improving robustness against common corruptions by covariate shift adaptation," *Advances in neural information processing systems*, vol. 33, pp. 11 539–11 551, 2020.
- [171] F. You, J. Li, and Z. Zhao, "Test-time batch statistics calibration for covariate shift," *arXiv preprint arXiv:2110.04065*, 2021.
- [172] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models," *ICLR*, vol. 1, no. 2, p. 3, 2022.
- [173] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Machine Intelligence*, vol. 5, no. 3, pp. 220–235, 2023.
- [174] X. Yang, D. Zhou, S. Liu, J. Ye, and X. Wang, "Deep model reassembly," *Advances in neural information processing systems*, vol. 35, pp. 25 739–25 753, 2022.
- [175] A. Aghajanyan, L. Zettlemoyer, and S. Gupta, "Intrinsic dimensionality explains the effectiveness of language model fine-tuning," *arXiv preprint arXiv:2012.13255*, 2020.
- [176] S. Arora, R. Ge, B. Neyshabur, and Y. Zhang, "Stronger generalization bounds for deep nets via a compression approach," in *International conference on machine learning*. PMLR, 2018, pp. 254–263.
- [177] Y. Ouali, A. Bulat, B. Martinez, and G. Tzimiropoulos, "Black box few-shot adaptation for vision-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 534–15 546.
- [178] J. Liang, D. Hu, J. Feng, and R. He, "Dine: Domain adaptation from single and multiple black-box predictors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8003–8013.
- [179] J. Dong, Z. Fang, A. Liu, G. Sun, and T. Liu, "Confident anchor-induced multi-source free domain adaptation," *Advances in neural information processing systems*, vol. 34, pp. 2848–2860, 2021.
- [180] Y.-C. Yu, C.-P. Huang, J.-J. Chen, K.-P. Chang, Y.-H. Lai, F.-E. Yang, and Y.-C. F. Wang, "Select and distill: Selective dual-teacher knowledge transfer for continual learning on vision-language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 219–236.
- [181] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 831–839.
- [182] M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesaro, "Learning to learn without forgetting by maximizing transfer and minimizing interference," in *International Conference on Learning Representations*, 2019.
- [183] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.
- [184] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.
- [185] Z. Ma, G. Luo, J. Gao, L. Li, Y. Chen, S. Wang, C. Zhang, and W. Hu, "Open-vocabulary one-stage detection with hierarchical visual-language knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 074–14 083.
- [186] H. Bangalath, M. Maaz, M. U. Khattak, S. H. Khan, and F. Shahbaz Khan, "Bridging the gap between object and image-level representations for open-vocabulary detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 781–33 794, 2022.
- [187] M. Xu, Z. Zhang, F. Wei, Y. Lin, Y. Cao, H. Hu, and X. Bai, "A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model," in *European Conference on Computer Vision*. Springer, 2022, pp. 736–753.
- [188] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin, "Scaling open-vocabulary image segmentation with image-level labels," in *European conference on computer vision*. Springer, 2022, pp. 540–557.
- [189] C. Zhou, C. C. Loy, and B. Dai, "Extract free dense labels from clip," in *European Conference on Computer Vision*. Springer, 2022, pp. 696–712.
- [190] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, and B. Cui, "Calip: Zero-shot enhancement of clip with parameter-free attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 746–754.
- [191] M. Zanella and I. Ben Ayed, "On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 783–23 793.
- [192] Z. Chi, L. Gu, T. Zhong, H. Liu, Y. Yu, K. N. Plataniotis, and Y. Wang, "Adapting to distribution shift by visual domain prompt generation," *arXiv preprint arXiv:2405.02797*, 2024.
- [193] Z. Chi, L. Gu, H. Liu, Z. Wang, Y. Wu, Y. Wang, and K. N. Plataniotis, "Learning to adapt frozen clip for few-shot test-time domain adaptation," in *The Thirteenth International Conference on Learning Representations*, 2025.
- [194] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak, "Promptstyler: Prompt-driven style generation for source-free domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 702–15 712.
- [195] B. Lew, D. Son, and B. Chang, "Gradient estimation for unseen domain risk minimization with pre-trained models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4436–4446.
- [196] X. Yu, S. Yoo, and Y. Lin, "Clipceil: Domain generalization through clip via channel refinement and image-text alignment," *Advances in Neural Information Processing Systems*, vol. 37, pp. 4267–4294, 2024.
- [197] D. Li, A. Wu, Y. Wang, and Y. Han, "Prompt-driven dynamic object-centric learning for single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 606–17 615.
- [198] H. Zhang, S. Bai, W. Zhou, J. Fu, and B. Chen, "Prompttta: Prompt-driven text adapter for source-free domain generalization," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
- [199] T. Yu, Z. Lu, X. Jin, Z. Chen, and X. Wang, "Task residual for tuning vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 899–10 909.

- [200] Z. Wang, J. Liang, R. He, N. Xu, Z. Wang, and T. Tan, "Improving zero-shot generalization for clip with synthesized prompts," *arXiv preprint arXiv:2307.07397*, 2023.
- [201] J. Silva-Rodriguez, S. Hajimiri, I. Ben Ayed, and J. Dolz, "A closer look at the few-shot adaptation of large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 681–23 690.
- [202] J.-X. Shi, C. Zhang, T. Wei, and Y.-F. Li, "Efficient and long-tailed generalization for pre-trained vision-language model," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 2663–2673.
- [203] Z. Lai, H. Bai, H. Zhang, X. Du, J. Shan, Y. Yang, C.-N. Chuah, and M. Cao, "Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 2691–2701.
- [204] M. Fahes, T.-H. Vu, A. Bursuc, P. Pérez, and R. De Charette, "Poda: Prompt-driven zero-shot domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 18 623–18 633.
- [205] S. Yang, Z. Tian, L. Jiang, and J. Jia, "Unified language-driven zero-shot domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 407–23 415.
- [206] Z. Wang, L. Zhang, L. Wang, and M. Zhu, "Landa: Language-guided multi-source domain adaptation," *arXiv preprint arXiv:2401.14148*, 2024.
- [207] X. Hu, K. Zhang, L. Xia, A. Chen, J. Luo, Y. Sun, K. Wang, N. Qiao, X. Zeng, M. Sun *et al.*, "Reclip: Refine contrastive language image pre-training with source free domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 2994–3003.
- [208] M. W. Gondal, J. Gast, I. A. Ruiz, R. Droste, T. Macri, S. Kumar, and L. Staudigl, "Domain aligned clip for few-shot classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5721–5730.
- [209] Y. Zhang, C. Zhang, K. Yu, Y. Tang, and Z. He, "Concept-guided prompt learning for generalization in vision-language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7377–7386.
- [210] C. Zhang, S. Stepputtis, K. Sycara, and Y. Xie, "Dual prototype evolving for test-time generalization of vision-language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 32 111–32 136, 2024.
- [211] L. Zhou, M. Ye, S. Li, N. Li, X. Zhu, L. Deng, H. Liu, and Z. Lei, "Bayesian test-time adaptation for vision-language models," *arXiv preprint arXiv:2503.09248*, 2025.
- [212] V. Udandarao, A. Gupta, and S. Albanie, "Sus-x: Training-free name-only transfer of vision-language models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2725–2736.
- [213] X. Zhu, R. Zhang, B. He, A. Zhou, D. Wang, B. Zhao, and P. Gao, "Not all features matter: Enhancing few-shot clip with adaptive prior refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2605–2615.
- [214] Z. Wang, J. Liang, L. Sheng, R. He, Z. Wang, and T. Tan, "A hard-to-beat baseline for training-free clip-based adaptation," *arXiv preprint arXiv:2402.04087*, 2024.
- [215] M. Farina, G. Franchi, G. Iacca, M. Mancini, and E. Ricci, "Frustratingly easy test-time adaptation of vision-language models," *Advances in Neural Information Processing Systems*, 2024.
- [216] W. Gerych, H. Zhang, K. Hamidieh, E. Pan, M. K. Sharma, T. Hartvigsen, and M. Ghassemi, "Bendvln: Test-time debiasing of vision-language embeddings," *Advances in Neural Information Processing Systems*, vol. 37, pp. 62 480–62 502, 2024.
- [217] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "Generalization through memorization: Nearest neighbor language models," *arXiv preprint arXiv:1911.00172*, 2019.
- [218] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 262–15 271.
- [219] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8340–8349.
- [220] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, "Learning robust global representations by penalizing local predictive power," *Advances in neural information processing systems*, vol. 32, 2019.
- [221] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.
- [222] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.
- [223] 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.
- [224] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.
- [225] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008, pp. 722–729.
- [226] 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.
- [227] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.
- [228] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.
- [229] 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.
- [230] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.
- [231] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.
- [232] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5542–5550.
- [233] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 1521–1528.
- [234] S. Beery, G. Van Horn, and P. Perona, "Recognition in terra incognita," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 456–473.
- [235] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan, "Deep hashing network for unsupervised domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5018–5027.
- [236] X. Peng, B. Usman, N. Kaushik, J. Hoffman, D. Wang, and K. Saenko, "Visda: The visual domain adaptation challenge," *arXiv preprint arXiv:1710.06924*, 2017.
- [237] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.
- [238] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.
- [239] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the association for computational linguistics*, vol. 2, pp. 67–78, 2014.
- [240] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," *arXiv preprint arXiv:1809.01696*, 2018.

- [241] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "Hero: Hierarchical encoder for video+ language omni-representation pre-training," *arXiv preprint arXiv:2005.00200*, 2020.
- [242] J. Lei, L. Yu, T. L. Berg, and M. Bansal, "Tvr: A large-scale dataset for video-subtitle moment retrieval," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*. Springer, 2020, pp. 447–463.
- [243] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Ständerhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3674–3683.
- [244] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. v. d. Hengel, "Reverie: Remote embodied visual referring expression in real indoor environments," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9982–9991.
- [245] 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.
- [246] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.
- [247] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 102–118.
- [248] X. Zhang, S. S. Gu, Y. Matsuo, and Y. Iwasawa, "Domain prompt learning for efficiently adapting clip to unseen domains," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 38, no. 6, pp. B–MC2_1, 2023.
- [249] J. Zhang, J. Huang, X. Zhang, L. Shao, and S. Lu, "Historical test-time prompt tuning for vision foundation models," *arXiv preprint arXiv:2410.20346*, 2024.
- [250] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning-the good, the bad and the ugly," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4582–4591.
- [251] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv preprint arXiv:2007.01434*, 2020.
- [252] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.
- [253] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.
- [254] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *European conference on computer vision*. Springer, 2022, pp. 280–296.
- [255] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang *et al.*, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," *arXiv preprint arXiv:2412.10302*, 2024.
- [256] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu *et al.*, "Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models," *arXiv preprint arXiv:2401.06066*, 2024.
- [257] C. Wu, X. Chen, Z. Wu, Y. Ma, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, C. Ruan *et al.*, "Janus: Decoupling visual encoding for unified multimodal understanding and generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 12 966–12 977.
- [258] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "Sharegpt4v: Improving large multi-modal models with better captions," in *European Conference on Computer Vision*. Springer, 2024, pp. 370–387.
- [259] X. Chen, Z. Wu, X. Liu, Z. Pan, W. Liu, Z. Xie, X. Yu, and C. Ruan, "Janus-pro: Unified multimodal understanding and generation with data and model scaling," *arXiv preprint arXiv:2501.17811*, 2025.
- [260] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in neural information processing systems*, vol. 35, pp. 25 278–25 294, 2022.
- [261] C. Schuhmann, A. Köpf, R. Vencu, T. Coombes, and R. Beaumont, "Laion coco: 600m synthetic captions from laion2b-en," *URL https://laion.ai/blog/laion-coco*, vol. 5, 2022.
- [262] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang *et al.*, "Datacomp: In search of the next generation of multimodal datasets," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 092–27 112, 2023.
- [263] M. Byeon, B. Park, H. Kim, S. Lee, W. Baek, and S. Kim, "Coyo-700m: Image-text pair dataset," *https://github.com/kakaobrain/coyo-dataset*, 2022.
- [264] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3558–3568.
- [265] 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.
- [266] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Advances in neural information processing systems*, vol. 24, 2011.
- [267] G. Ilharco, M. Wortsman, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," Jul. 2021. [Online]. Available: <https://doi.org/10.5281/zenodo.5143773>
- [268] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.
- [269] A. Fang, A. M. Jose, A. Jain, L. Schmidt, A. Toshev, and V. Shankar, "Data filtering networks," *arXiv preprint arXiv:2309.17425*, 2023.
- [270] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.
- [271] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu, "Qwen2.5 technical report," 2025. [Online]. Available: <https://arxiv.org/abs/2412.15115>
- [272] 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.
- [273] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv preprint arXiv:2111.02114*, 2021.
- [274] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.
- [275] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, and Y. Cao, "Eva: Exploring the limits of masked visual representation learning at scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 358–19 369.
- [276] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin *et al.*, "Opt: Open pre-trained transformer language models," *arXiv preprint arXiv:2205.01068*, 2022.
- [277] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
- [278] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-

- source chatbot impressing gpt-4 with 90%* chatgpt quality," *See <https://vicuna.lmsys.org> (accessed 14 April 2023)*, vol. 2, no. 3, p. 6, 2023.
- [279] L. Xue, M. Shu, A. Awadalla, J. Wang, A. Yan, S. Purushwalkam, H. Zhou, V. Prabhu, Y. Dai, M. S. Ryoo *et al.*, "xgen-mm (blip-3): A family of open large multimodal models," *arXiv preprint arXiv:2408.08872*, 2024.
- [280] A. Awadalla, L. Xue, O. Lo, M. Shu, H. Lee, E. Guha, S. Shen, M. Awadalla, S. Savarese, C. Xiong *et al.*, "Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens," *Advances in Neural Information Processing Systems*, vol. 37, pp. 36 805–36 828, 2024.
- [281] H. Laurençon, L. Tronchon, M. Cord, and V. Sanh, "What matters when building vision-language models?" *Advances in Neural Information Processing Systems*, vol. 37, pp. 87 874–87 907, 2024.
- [282] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.
- [283] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.
- [284] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 296–26 306.
- [285] S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu *et al.*, "Llava-plus: Learning to use tools for creating multimodal agents," in *European Conference on Computer Vision*. Springer, 2024, pp. 126–142.
- [286] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.
- [287] W. Huang, Z. Zhai, Y. Shen, S. Cao, F. Zhao, X. Xu, Z. Ye, Y. Hu, and S. Lin, "Dynamic-llava: Efficient multimodal large language models via dynamic vision-language context sparsification," *arXiv preprint arXiv:2412.00876*, 2024.
- [288] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello *et al.*, "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.
- [289] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Riviere, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.
- [290] X. Chen, X. Wang, L. Beyer, A. Kolesnikov, J. Wu, P. Voigtlaender, B. Mustafa, S. Goodman, I. Alabdulmohsin, P. Padlewski *et al.*, "Pali-3 vision language models: Smaller, faster, stronger," *arXiv preprint arXiv:2310.09199*, 2023.
- [291] A. Steiner, A. S. Pinto, M. Tschannen, D. Keysers, X. Wang, Y. Bitton, A. Gritsenko, M. Minderer, A. Sherbondy, S. Long *et al.*, "Paligemma 2: A family of versatile vlms for transfer," *arXiv preprint arXiv:2412.03555*, 2024.
- [292] B. McKinzie, Z. Gan, J.-P. Fauconnier, S. Dodge, B. Zhang, P. Dufter, D. Shah, X. Du, F. Peng, A. Belyi *et al.*, "Mm1: methods, analysis and insights from multimodal llm pre-training," in *European Conference on Computer Vision*. Springer, 2024, pp. 304–323.
- [293] J. Cha, W. Kang, J. Mun, and B. Roh, "Honeybee: Locality-enhanced projector for multimodal llm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 817–13 827.
- [294] H. Zhang, M. Gao, Z. Gan, P. Dufter, N. Wenzel, F. Huang, D. Shah, X. Du, B. Zhang, Y. Li *et al.*, "Mm1.5: Methods, analysis & insights from multimodal llm fine-tuning," *arXiv preprint arXiv:2409.20566*, 2024.
- [295] J. Chen, D. Zhu, X. Shen, X. Li, Z. Liu, P. Zhang, R. Krishnamoorthi, V. Chandra, Y. Xiong, and M. Elhoseiny, "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.
- [296] 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.
- [297] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei, "Kosmos-2: Grounding multimodal large language models to the world," *arXiv preprint arXiv:2306.14824*, 2023.
- [298] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.
- [299] D. Guo, F. Wu, F. Zhu, F. Leng, G. Shi, H. Chen, H. Fan, J. Wang, J. Jiang, J. Wang *et al.*, "Seed1.5-v1 technical report," *arXiv preprint arXiv:2505.07062*, 2025.
- [300] C. Deng, D. Zhu, K. Li, C. Gou, F. Li, Z. Wang, S. Zhong, W. Yu, X. Nie, Z. Song *et al.*, "Emerging properties in unified multimodal pretraining," *arXiv preprint arXiv:2505.14683*, 2025.
- [301] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa *et al.*, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.
- [302] Y. Li, H. Sun, M. Lin, T. Li, G. Dong, T. Zhang, B. Ding, W. Song, Z. Cheng, Y. Huo *et al.*, "Baichuan-omni technical report," *arXiv preprint arXiv:2410.08565*, vol. 3, no. 7, 2024.
- [303] Y. Li, J. Liu, T. Zhang, S. Chen, T. Li, Z. Li, L. Liu, L. Ming, G. Dong, D. Pan *et al.*, "Baichuan-omni-1.5 technical report," *arXiv preprint arXiv:2501.15368*, 2025.
- [304] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 612–17 625, 2022.
- [305] 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.
- [306] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.
- [307] M. Dehghani, B. Mustafa, J. Djolonga, J. Heek, M. Minderer, M. Caron, A. Steiner, J. Puigcerver, R. Geirhos, I. M. Alabdulmohsin *et al.*, "Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution," *Advances in Neural Information Processing Systems*, vol. 36, pp. 2252–2274, 2023.
- [308] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.
- [309] G. Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.
- [310] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.
- [311] N. Lambert, J. Morrison, V. Pyatkin, S. Huang, H. Ivison, F. Brahman, L. J. V. Miranda, A. Liu, N. Dziri, S. Lyu *et al.*, "T\`ulu 3: Pushing frontiers in open language model post-training," *arXiv preprint arXiv:2411.15124*, 2024.
- [312] Z. Zheng, Z. Cheng, Z. Shen, S. Zhou, K. Liu, H. He, D. Li, S. Wei, H. Hao, J. Yao *et al.*, "Livecodebench pro: How do olympiad medalists judge llms in competitive programming?" *arXiv preprint arXiv:2506.11928*, 2025.
- [313] K. Zhang, J. Li, G. Li, X. Shi, and Z. Jin, "Codeagent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges," *arXiv preprint arXiv:2401.07339*, 2024.
- [314] J. Xie, Z. Chen, R. Zhang, X. Wan, and G. Li, "Large multimodal agents: A survey," *arXiv preprint arXiv:2402.15116*, 2024.