DeFT-LoRA: Decoupled and Fused Tuning with LoRA Experts for Universal Cross-Domain Retrieval

Ke Xu[1, 2], **Xiaozheng Shen**[1, 3], **Shanshan Wang**[1, 4*], **Mengzhu Wang**[5], **Xun Yang**[6]

[1]State Key Laboratory of Opto-Electronic Information Acquisition and Protection Technology, Anhui University, Hefei, China
[2]School of Artificial Intelligence, Anhui University, Hefei, China
[3]School of Computer Science and Technology, Anhui University, Hefei, China
[4]Institutes of Physical Science and Information Technology, Anhui University, Hefei, China
[5]Hebei University of Technology, Tianjin, China
[6]University of Science and Technology of China, Hefei, China
{wang.shanshan, xuke}@ahu.edu.cn, q23201182@stu.ahu.edu.cn, dreamkily@gmail.com, xyang21@ustc.edu.cn

Abstract

Universal Cross-Domain Retrieval (UCDR) aims to retrieve images across unseen domains and categories, a critical capability for real-world applications. While large-scale Vision-Language Models (VLMs) like CLIP offer strong zero-shot category generalization, they struggle with domain shifts. Existing methods often improve domain robustness at the cost of high computational overhead or by compromising the VLM's inherent knowledge. To address this, we propose Decoupled and Fused Tuning with LoRA (DeFT-LoRA), a novel and parameter-efficient framework that integrates Low-Rank Adaptation (LoRA) with a Mixture-of-Experts (MoE) mechanism. This approach resolves the intrinsic conflict between domain-invariant and domain-specific knowledge in a single adapter, enabling our model to construct a domain adapters for each input image. We propose a three-stage training strategy, which first learns a shared Base LoRA for domain-invariant features, then derives Domain-Specific Experts to capture specific styles, and finally fuses them dynamically with a lightweight gating network. Extensive experiments on three UCDR benchmarks demonstrate that DeFT-LoRA achieves comparable or superior performance to state-of-the-art methods while requiring only 1.46 percent of CLIP's image-encoder parameters and reducing computational overhead, thereby establishing an exceptional balance between accuracy and efficiency.

code — https://github.com/wildboarman/DeFT-LoRA

Introduction

Universal Cross-Domain Retrieval (UCDR) (Paul, Dutta, and Biswas 2021) is a challenging task that aims to achieve robust performance in open-world scenarios, where query images may originate from domains and categories entirely unseen during training. The core challenge of UCDR lies in learning transferable image representations from seen data while preserving the generalization ability of pre-trained models (Tian et al. 2022; Mondal and Biswas 2022), a principle that has proven critical even in more complex tasks like video retrieval (Yang et al. 2022, 2021, 2024a,b; Pan et al. 2024).

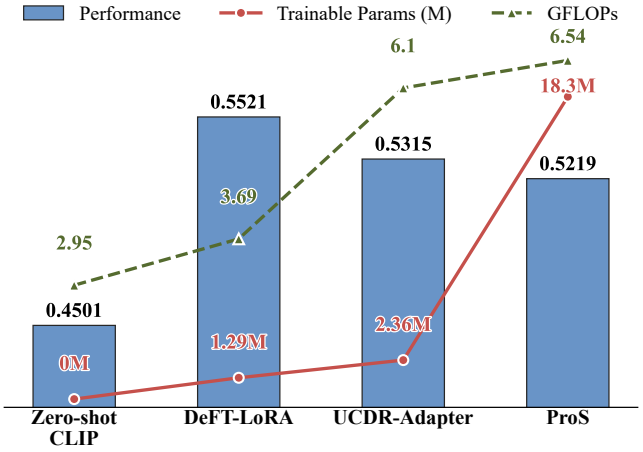

Figure 1: Performance and efficiency comparison on the DomainNet benchmark under UCDR setting with Infograph as the unseen domain. Our method not only achieves new state-of-the-art performance but does so with a substantial reduction in trainable parameters and FLOPs.

To address these challenges, large-scale VLMs such as CLIP (Radford et al. 2021), pre-trained on extensive image-text pairs, deliver robust zero-shot generalization to unseen categories. However, they remain prone to performance degradation when faced with significant domain shifts caused by variations in style or imaging conditions. Therefore, this work focuses on a critical objective: to *enhance CLIP's robustness to domain shifts while maximally preserving its innate ability to generalize to novel categories*.

Existing studies indicate that, compared to full fine-tuning, prompt-based methods (Jia et al. 2022; Zhou et al. 2022b; Fang et al. 2024; Jiang et al. 2025) are more effective at preserving CLIP's generalization capabilities while demonstrating superior parameter efficiency (Kumar et al. 2022]). Building on this, ProS (Fang et al. 2024) introduced the innovative "Prompting-to-Simulate" paradigm, which significantly boosts CLIP's UCDR performance. UCDR-Adapter (Jiang et al. 2025) further refined this paradigm, achieving state-of-

*Corresponding Authors.

the-art results. However, these methods rely on numerous prompts and attention-based modules, which substantially increase the input token count. Given that Transformer's self-attention mechanism has $O(n^2)$ computational complexity relative to token count, this results in significantly higher computational overhead during inference as the number of tokens grows quadratically. We argue that UCDR necessitates a *more efficient paradigm that retains strong generalization*.

Low-Rank Adaptation (LoRA) (Hu et al. 2022) preserves computational efficiency by avoiding the introduction of additional computational modules during fine-tuning. Moreover, its principle of minor intervention not only preserves this efficiency but also minimizes catastrophic forgetting of the pre-trained knowledge (Kumar et al. 2022]). Nevertheless, a naive application of LoRA struggles to enhance CLIP's UCDR performance. We argue that this difficulty stems from the forced coupling of general and domain-specific knowledge within a single LoRA module (Yang et al. 2024c; Wang et al. 2023). This intrinsic conflict hinders the single adapter from learning transferable representations. A logical extension is to employ multiple low-rank experts, but a naive Mixture-of-Experts (MoE) (Dou et al. 2023) approach still fails to resolve the underlying issue. By training independent experts in parallel, the network results in multiple, poorly decoupled experts, where each module still learns a redundant and uncoordinated mixture of shared and specialized knowledge (Xu et al. 2025; Komatsuzaki et al. 2022).

To address the above challenges, we propose DeFT-LoRA, a novel framework that strategically decouples domain-invariant and domain-specific knowledge before dynamically fusing them for domain adaptation. DeFT-LoRA implements a carefully designed three-stage training strategy. In the **Base Expert Learning** stage, we train a base lora module on all trainable source data to capture domain-invariant representations. In the subsequent **Domain Expert Learning** stage, we build upon this shared foundation of the Base Expert by learning incremental domain experts. Our key insight is to decouple the adaptation: for each source domain, its dedicated expert inherits the general knowledge structure from the base, while being trained exclusively on that domain's data to learn a unique component modeling specific styles. This strategy forces each expert to capture domain-specific knowledge as an incremental addition built upon the common foundation. Finally, in the **Joint Optimization of Gating and Experts** stage, with knowledge now cleanly decoupled into a reusable base and specialized components, we introduce a lightweight gating network to achieve dynamic Fusing. This gate composes a tailored adaptation for each image by blending the foundational knowledge from the base expert with the different stylistic knowledge from the domain experts. Furthermore, DeFT-LoRA demonstrates exceptional parameter efficiency and computational efficiency while maintaining superior performance compared to existing approaches, as shown in Figure 1. Unlike prompt-based methods, we do not introduce any large additional modules into the image encoder. This paper has the following key contributions:

- We first analyze the limitations of existing prompt-based methods for UCDR, particularly their high computational cost, and identify LoRA as a more efficient paradigm.
- We reveal the intrinsic conflict of naive LoRA-based tuning, showing that neither a single adapter nor a naive LoRA-MoE approach can effectively separate domain-invariant from domain-specific knowledge. The latter leads to poorly decoupled experts, each still suffering from the same underlying knowledge conflict.
- We propose DeFT-LoRA, a novel framework featuring an innovative three-stage training strategy. This strategy first decouples knowledge into a shared Base Expert and multiple Domain-Specific Experts, and then fuses them dynamically for each input.
- Extensive experiments show that DeFT-LoRA achieves state-of-the-art performance on major UCDR benchmarks. Compared to strong baselines like UCDR-Adapter, it drastically reduces the number of trainable parameters (from 2.36M to 1.29M) and computational overhead (from 6.1G to 3.69G FLOPs), demonstrating superior efficiency.

Related Works

Universal Cross-Domain Retrieval. Universal Cross-Domain Retrieval (UCDR) (Paul, Dutta, and Biswas 2021) is a challenging task requiring generalization across unseen domains and categories, combining principles from Domain Generalization (DG) (Wang et al. 2025b,a, 2024) and Zero-Shot Learning (ZSL) (Pourpanah et al. 2022). Early approaches (Paul, Dutta, and Biswas 2021; Tian et al. 2022) tackled this by designing specialized architectures that integrated data augmentation strategies or learned categorical prototypes to bridge the domain and semantic gaps.

CLIP (Radford et al. 2021) has established a new paradigm for UCDR with their strong zero-shot capabilities. Current state-of-the-art methods (Fang et al. 2024; Jiang et al. 2025) are built upon a "prompting-to-simulate" paradigm. However, these prompt-based approaches introduce substantial computational overhead and parameter count via costly additional modules. In contrast to prior works, our DeFT-LoRA is the first to apply LoRA to UCDR, featuring a novel architecture specifically designed for its distinct properties.

CLIP. The Contrastive Language-Image Pre-training (CLIP) (Radford et al. 2021) model has significantly advanced the field of Vision-Language Pre-training (VLP). It learns robust semantic representations by training separate visual and text encoders on a large-scale dataset of image-text pairs using a contrastive objective. A key characteristic of CLIP is its remarkable zero-shot generalization capability, allowing it to adapt to unseen classes without direct training. To further enhance its performance on specific downstream tasks, the mainstream approaches (Zhou et al. 2022b,a; Bose et al. 2024; Khattak et al. 2023; Abdul Samadh et al. 2023; Cheng et al. 2024; Cai et al. 2024) are to introduce a small number of learnable parameters, while keeping the pre-trained backbone frozen. However, standard adapters often fall short in bridging both category and domain gaps in UCDR, motivating the development of more targeted solutions.

Low-Rank Adaptation (LoRA). LoRA (Hu et al. 2022) is a parameter-efficient fine-tuning (PEFT) method that adapts

a pre-trained weight matrix $W_0 \in \mathbb{R}^{m\times n}$ by learning a low-rank update. Specifically, it learns the ΔW as a product of two much smaller low-rank matrices, $A \in \mathbb{R}^{r\times n}$ and $B \in \mathbb{R}^{m\times r}$, where the rank $r \ll \min(m, n)$. The effectiveness and simplicity have led to a wide range of applications and many variant extensions (Wang et al. 2025c; Liu et al. 2024; Meng, Wang, and Zhang 2024; Dettmers et al. 2024; Hayou, Ghosh, and Yu 2024; Si et al. 2025; Dong et al. 2024; He et al. 2023; Han et al. 2025) for vision and language tasks. Building on this foundation, recent researches have explored multi-LoRA architectures, often using a Mixture-of-Experts (MoE) framework for enhanced adaptability (Dou et al. 2023; Tian et al. 2024; Wang et al. 2023; Yang et al. 2024c; Chen et al. 2025; Wu, Huang, and Wei 2024; Wu et al. 2024). However, both single LoRA and naive LoRA-MoE struggle to resolve the conflict between learning domain-invariant and specific knowledge. To this end, we propose DeFT-LoRA to leverages a novel multi-stage training process to decouple and fuse these knowledge components.

Method

In this section, we introduce our novel parameter-efficient framework, Decoupled and Fused Tuning with LoRA (DeFT-LoRA), designed for Universal Cross-Domain Retrieval (UCDR). We begin by formally defining the UCDR problem. Subsequently, we will detail our core contribution: the DeFT-LoRA and a three-stage training strategy, which decouple and fuse general and domain-specific knowledge. The overall architecture of DEFT-LoRA is illustrated in Figure 2.

UCDR Problem Formulation

The task of Universal Cross-Domain Retrieval (UCDR) (Paul, Dutta, and Biswas 2021) aims to learn a model that can retrieve relevant images from a gallery, where both query and gallery images may come from domains and categories unseen during training. Formally, let $\mathcal{D}_{train} = \bigcup_{i=1}^{N_{dom}} \mathcal{D}_i$ be the training dataset, comprising images from N_{dom} distinct source domains. Each image $I \in \mathcal{D}_{train}$ is associated with a label y from the source category set, $\mathcal{C}_{train}$. The evaluation is performed on a query domain dataset, $\mathcal{D}_{query}$, which contains images from a disjoint set of categories $\mathcal{C}_{test}$. In the testing phase, the gallery set and query set are needed to enable image retrieval. The core assumptions of UCDR are:

- **Unseen Domain**: The target domain is not available during training, i.e., $query \notin \{1, ..., N_{dom}\}$.
- **Unseen Categories**: The categories in the query domain are disjoint from those in the source domains, i.e., $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \emptyset$.

There are two settings in gallery set: 1) all samples belong to the unseen test class, termed Unseen Gallery; (2) samples belong to both seen class and unseen class, termed Mixed Gallery. The UCDR protocol encompasses two primary evaluation settings:

- **Unseen-class Cross-Domain Retrieval (U$^{\mathbf{c}}$CDR):** In this setting, the model is evaluated on a domain seen during training, but with entirely new object categories. Formally, $query \in \{1, ..., N_{dom}\}$, but the category sets are disjoint: $\mathcal{C}_{train} \cap \mathcal{C}_{test} = \emptyset$.
- **Unseen-domain Cross-Domain Retrieval (U$^{\mathbf{d}}$CDR):** Here, the model is tested on a new domain, but the object categories are the same as those seen during training. Formally, $query \notin \{1, ..., N_{dom}\}$, but $\mathcal{C}_{train} = \mathcal{C}_{test}$.

DeFT-LoRA: A Three-Stage Framework

To efficiently adapt the image encoder ϕ_I for UCDR, we propose DeFT-LoRA, a framework built on a novel three-stage strategy. Our method uses LoRA "experts" and a gating mechanism to decouple and fuse domain-specific and invariant knowledge through a three-stage strategy: Base Expert Learning, Domain Expert Learning, and Joint Optimization.

Base Expert Learning for Domain-Invariant Knowledge. To capture the domain-invariant knowledge shared across all source domains, we introduce a **Base Expert**, which is a LoRA-based module designed to learn a generalizable representation update. Specifically, we augment the pre-trained image encoder ϕ_I, which consists of L transformer layers. For each attention block within every layer $l \in \{1, \dots, L\}$, according to the settings of the LoRA paper, we introduce LoRA matrices to adapt the query (W_q) and value (W_v) projection matrices. The set of trainable parameters for our Base Expert is therefore a collection of LoRA matrix pairs:

$$\Theta_B = \bigcup_{l=1}^{L} \{(A_{B,q}^{(l)}, B_{B,q}^{(l)}), (A_{B,v}^{(l)}, B_{B,v}^{(l)})\}, \tag{1}$$

where $(A_{B,q}^{(l)}, B_{B,q}^{(l)})$ are the LoRA matrices for the query projection in layer l, and $(A_{B,v}^{(l)}, B_{B,v}^{(l)})$ are for the value projection. Each pair has a rank r such that $A \in \mathbb{R}^{r\times n}$ and $B \in \mathbb{R}^{m\times r}$. During this stage, the original parameters of ϕ_I and the text encoder ϕ_T are kept frozen.

The forward pass through an adapted projection matrix (e.g., $W_{0,q}^{(l)}$) in layer l is modified as:

$$h_q^{(l)} = W_{0,q}^{(l)} x + B_{B,q}^{(l)}(A_{B,q}^{(l)} x). \tag{2}$$

The Base Expert is trained on the aggregated source dataset $\mathcal{D}_{train}$, ensuring it is exposed to all available domain characteristics simultaneously. For an input image I_i from this mixed dataset, the adapted image encoder produces the feature $z_{I,i} = \phi_I(I_i; \Theta_B)$. Concurrently, we utilize the learnable text prompts $\{v_{dom}\}$ to generate a corresponding text feature $z_{T,i} = \phi_T(t_i)$ (Zhou et al. 2022b). The training objective is to align these image and text features within the shared embedding space. This is achieved by minimizing the contrastive loss over each batch of M samples: The image-to-text loss for the i-th image is:

$$\mathcal{L} = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{N_t} y_{i,j} \log p(t_j | I_i), \tag{3}$$

where $y_{i,j}$ is the ground-truth label. The only trainable parameters in this stage are the Base Expert weights Θ_B and the prompt vectors $\{v_{domain}\}$. Upon completion, the optimized Base Expert parameters Θ_B^* are frozen and serve as the foundation for subsequent stages.

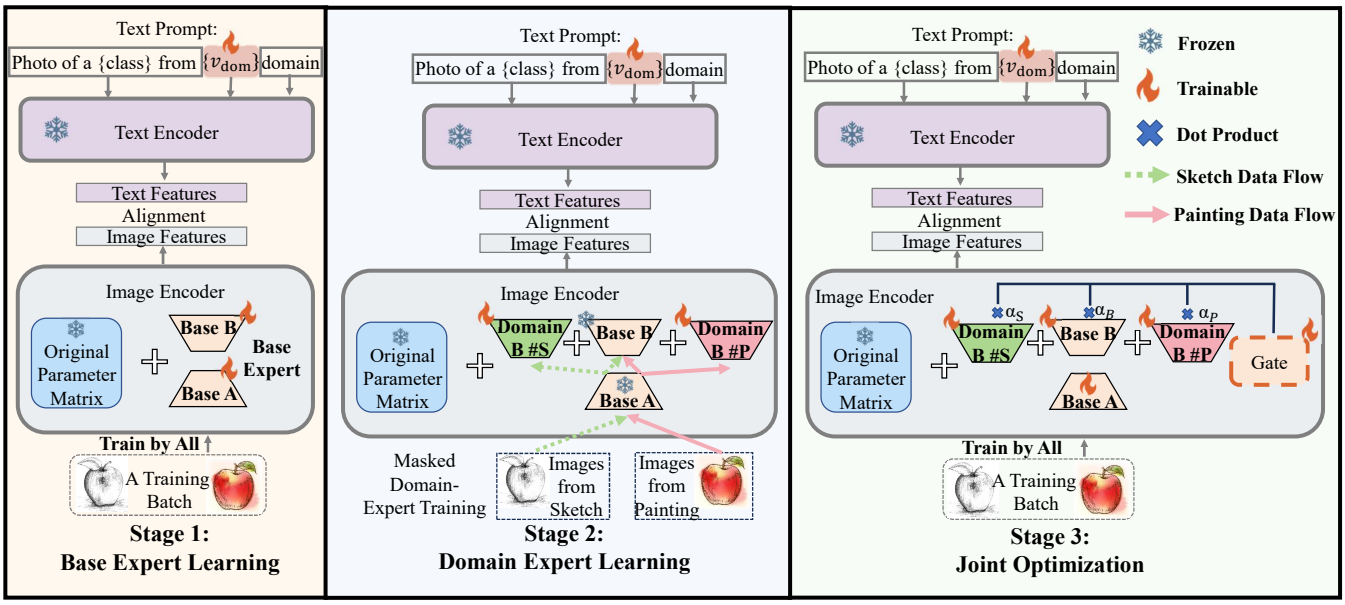

Figure 2: The three-stage training strategy of DEFT-LoRA. (1) A shared Base Expert learns a domain-invariant foundation. (2) Specialized Domain Experts are then trained to capture unique styles upon a frozen base projection. (3) Finally, a gating network is jointly optimized to dynamically fuse all experts.

Domain Expert Learning for Domain-Specific Knowledge. Building upon the frozen Base Expert Θ_B^*, this stage introduces a set of parameter-efficient Domain-Specific Experts to capture specialized knowledge. Our design is motivated by recent findings (Tian et al. 2024) that when LoRA modules are trained on different but related tasks or domains, their learned A_B matrices tend to be similar, while the B_B matrices capture more task-specific or domain-specific variations.

Therefore, for each domain $d \in \{1, \ldots, N_{dom}\}$, we employ a decoupled strategy: we introduce only a new set of B matrices $\Theta_{D,d} = \bigcup_{l=1}^{L} \{(B_{D,d,q}^{(l)}), (B_{D,d,v}^{(l)})\}$ while reusing the frozen A_B matrices from the Base Expert. This architecture forces all experts to share a common low-rank projection space, where the shared A_B matrices provide a basis of fundamental visual edits, and each domain-specific $B_{D,d}$ matrix learns a unique style for combining them. This approach not only facilitates knowledge transfer but also drastically reduces the number of trainable parameters.

The training process is crucial for isolating this domain-specific learning. We employ a masked optimization scheme (Xu et al. 2025): for a batch of images exclusively from a single source domain d, only its corresponding Domain Expert $\Theta_{D,d}$ is activated for the gradient update, while all other domain experts remain frozen. During the forward pass, the input benefits from the combined output of the Base Expert Θ_B^* and this active Domain Expert $\Theta_{D,d}$. However, during the backward pass, we also freeze the Base Expert's parameters Θ_B^*. This dual-freezing strategy ensures that gradients only update the active Domain Expert, compelling it to learn the domain-specific residual on top of the stable, domain-invariant foundation. The update to the original weight matrix's output for an input x from domain d is:

$$\Delta W x = \underbrace{B_{B,q}^{(l)}(A_{B,q}^{(l)}x)}_{\text{Base Expert Contribution}} + \underbrace{B_{D,d,q}^{(l)}(A_{B,q}^{(l)}x)}_{\text{Domain-Specific Contribution}}. \quad (4)$$

This can be simplified by factoring out the shared projection:

$$\Delta W x = (B_{B,q}^{(l)} + B_{D,d,q}^{(l)})(A_{B,q}^{(l)}x). \quad (5)$$

The total adapted image feature is $z_I = \phi_I(I; \Theta_B^*, \Theta_{D,d})$. The training objective remains the loss in Equation (3).

Joint Optimization of Gating and Experts. While the previous stages provide robust initializations for the experts, they are trained in isolation. It allows the expert modules to slightly refine their parameters to be more effectively fused, while simultaneously training the gating network to learn content-aware routing strategies. So during this stage, all components—the Gating Network $G(\cdot)$, the Base Expert Θ_B, and all Domain-Specific Experts $\{\Theta_{D,d}\}_{d=1}^{N_{dom}}$—are trainable. At each transformer layer l, the gating module $G^{(l)}$ takes the output feature from the previous layer, $x^{(l-1)}$, as input to compute the fusion weights. The gating module $G^{(l)}$ first generates $N_{dom}+1$ scores, which are normalized via Softmax to produce the layer-specific weights:

$$[\alpha_0^{(l)}, \alpha_1^{(l)}, \ldots, \alpha_{N_{dom}}^{(l)}] = \text{Softmax}(G^{(l)}(x^{(l-1)})). \quad (6)$$

Here, $\alpha_0^{(l)}$ is the weight for the Base Expert at layer l, and $\{\alpha_d^{(l)}\}_{d=1}^{N_{dom}}$ are for the Domain-Specific Experts. The total update, $\Delta W^{(l)} x^{(l-1)}$, applied to the output of the original weight matrix (e.g., for query or value) in layer l is a dynamically weighted sum of all expert contributions. This

entire process can be summarized in a single, comprehensive equation. For clarity, we have omitted the q subscripts:

$$\Delta W^{(l)} x^{(l-1)} = \left(\underbrace{\alpha_0^{(l)} B_B^{(l)}}_{\text{Base Expert}} + \underbrace{\sum_{d=1}^{N_{dom}} \alpha_d^{(l)} B_{D,d}^{(l)}}_{\text{Domain Experts}} \right) \underbrace{A_B^{(l)} x^{(l-1)}}_{\text{Shared Projection}} . \tag{7}$$

The entire model is trained end-to-end on the aggregated source dataset $\mathcal{D}_{train}$. The training objective is the same image-text contrastive loss (Zhou et al. 2022b) as used in the preceding stages. This forces the gating modules to learn meaningful, depth-aware routing strategies, while simultaneously refining the expert parameters to maximize their collective effectiveness for the contrastive task. This ensures the emergence of a highly adaptive and cohesive system.

Experiment

Experimental Setup

DataSet. To comprehensively evaluate our method, we conduct experiments on three widely-used benchmark datasets: DomainNet (Peng et al. 2019), Sketchy (Sangkloy et al. 2016; Liu et al. 2017), and TU-Berlin (Eitz et al. 2010; Liu et al. 2017). Following established protocols (Fang et al. 2024), we assess DeFT-LoRA under three challenging cross-domain retrieval settings: UCDR, U$^{\text{d}}$CDR, and U$^{\text{c}}$CDR.

Evaluation Metrics. To ensure fair and consistent comparison with prior art, we employ standard retrieval metrics (Paul, Dutta, and Biswas 2021; Tian et al. 2022). For DomainNet and Sketchy, we report Precision@200 (Prec@200) and mean Average Precision@200 (mAP@200). For TU-Berlin, we report Prec@100 and mAP@all.

Implementation Details. Our framework is built upon the pre-trained CLIP model, which uses a ViT-B/32 (Dosovitskiy et al. 2021) backbone for its image encoders. The key hyperparameters for DeFT-LoRA are configured as follows: we set the LoRA rank to $r = 2$ when *Painting* or *Clipart* are the unseen domains and $r = 4$ for all other cases. LoRA is applied to the W_q and W_v matrices, except when *Quickdraw* is the unseen domain, in which case the W_k matrix is also adapted. $G^{(l)}$ is a single linear layer followed by a Softmax function. Following (Zhou et al. 2022b), our text prompts use a context length of 16 with a 512-dimensional feature vector. For the training schedule, Stage 1 and Stage 2 are each trained for a single epoch, while Stage 3 is trained for up to 10 epochs with an early stopping patience of 2, monitored on the validation mAP. We employ the Adam optimizer with a learning rate of 1e-4 and a cosine decay schedule.

Main results

We compare with two groups of methods. The first group comprises SnMpNet (Paul, Dutta, and Biswas 2021) and SASA (Tian et al. 2022), which respectively adopt ResNet and ViT as backbones. The second group includes CLIP-Full (Radford et al. 2021), ProS (Fang et al. 2024), and UCDR-Adapter (Jiang et al. 2025), all of which are built upon the CLIP backbone.

Results on UCDR. The detailed results of UCDR on the DomainNet dataset are presented in Table 1. Our DeFT-LoRA framework demonstrates superior performance among all CLIP-based competitors. On average, DeFT-LoRA achieves the highest scores across both two gallery settings, which represents a notable improvement over the strongest prompt-based baselines. Specifically, our method outperforms UCDR-Adapter by 1.41% in average mAP200, highlighting its enhanced generalization capability. What's more, the superiority of DeFT-LoRA is particularly pronounced on domains with significant distribution shifts. We highlight its performance on two notably challenging unseen domains: Infograph and Quickdraw. The Infograph domain characterized by complex scenes and object co-occurrences. Here, our method achieves an mAP@200 of 0.6093 on the Unseen Gallery, substantially outperforming the strongest competitor by a significant margin of 2.95%. Conversely, the Quickdraw domain presents an extreme case of information scarcity. Even under this challenging condition, DeFT-LoRA establishes a new state-of-the-art with an mAP@200 of 0.3012, surpassing all other methods.

Result on U$^{\text{d}}$CDR. We now evaluate our method under the U$^{\text{d}}$CDR setting, where the model is tested on unseen domains but with categories seen during training. As shown in Table 1, DeFT-LoRA achieves a new state-of-the-art average mAP@200 of 0.6547. While the ViT-based SASA method is highly competitive in this setting—as the need for zero-shot category generalization is removed, leveling the playing field for non-CLIP models—our DeFT-LoRA still demonstrates superior robustness. This is particularly evident in semantically complex domains like Infograph, where our method achieves an mAP@200 of 0.6454, significantly outperforming all rivals. This result underscores our model's strong capability in aligning high-level semantic features even when faced with significant stylistic shifts.

Result on U$^{\text{c}}$CDR. We now evaluate our method under the U$^{\text{c}}$CDR setting. The detailed results on the two datasets are presented in Table 2. DeFT-LoRA achieves state-of-the-art U$^{\text{c}}$CDR performance on both the Sketchy and TU-Berlin benchmarks. This leading performance demonstrates our method's superior ability to capture discriminative features from the complex, multi-source domain inputs present during training. More importantly, achieving these top results on an unseen-class task strongly confirms that our structured, parameter-efficient approach successfully preserves and leverages CLIP's foundational zero-shot generalization capability.

Discussion

Ablation Study

To validate the architectural choices behind DeFT-LoRA, we systematically ablate its key components. We conduct experiments on the DomainNet (Peng et al. 2019) benchmark under the UCDR setting with Infograph as the unseen domain. The results are summarized in Table 3. We observe the following from the ablations:

Query Domain	Method	UCDR				U^{d}CDR	
		Unseen Gallery		Mixed Gallery			
		mAP@200	Prec@200	mAP@200	Prec@200	mAP@200	Prec@200
Sketch	SnMpNet	0.3007	0.2432	0.2624	0.2134	0.3529	0.1657
	SASA	0.5262	0.4468	0.4732	0.4025	0.5733	**0.5290**
	CLIP-Full	0.5367	0.4666	0.4788	0.4136	0.6128	0.3806
	ProS	0.6457	0.6001	0.5843	0.5463	0.7385	0.4911
	UCDR-Adapter	0.6591	0.6142	**0.6073**	**0.5707**	0.7332	0.4893
	DeFT-LoRA(Ours)	**0.6663**	**0.6207**	0.6014	0.5643	**0.7496**	0.5021
Quickdraw	SnMpNet	0.1736	0.1284	0.1512	0.1111	0.1077	0.0509
	SASA	0.2564	0.1970	0.2116	0.1651	0.1805	**0.1549**
	CLIP-Full	0.2011	0.1522	0.1622	0.1196	0.1820	0.0723
	ProS	0.2842	0.2544	0.2318	0.2127	0.2889	0.1186
	UCDR-Adapter	0.2794	0.2534	0.2317	0.2154	0.2900	0.1181
	DeFT-LoRA(Ours)	**0.3012**	**0.2668**	**0.2522**	**0.2279**	**0.3010**	0.1194
Painting	SnMpNet	0.4031	0.3332	0.3635	0.3019	0.4808	0.4424
	SASA	0.5898	0.5188	0.5463	0.4804	0.5596	**0.5178**
	CLIP-Full	0.6558	0.5923	0.6083	0.5478	0.6189	0.3688
	ProS	0.7516	0.6955	0.7120	0.6612	0.7227	0.4615
	UCDR-Adapter	0.7538	0.6974	**0.7203**	**0.6693**	0.7306	0.4634
	DeFT-LoRA(Ours)	**0.7551**	**0.6994**	0.7135	0.6639	**0.7415**	0.4752
Infograph	SnMpNet	0.2079	0.1717	0.1800	0.1496	0.1957	0.1764
	SASA	0.2823	0.2425	0.2491	0.2113	0.2340	0.2093
	CLIP-Full	0.5332	0.4893	0.4718	0.4309	0.5311	0.3330
	ProS	0.5798	0.5442	0.5219	0.4956	0.6056	0.3962
	UCDR-Adapter	0.5714	0.5364	0.5315	0.5022	0.6064	0.3922
	DeFT-LoRA(Ours)	**0.6093**	**0.5735**	**0.5521**	**0.5242**	**0.6454**	**0.4171**
Clipart	SnMpNet	0.4198	0.3323	0.3765	0.2959	0.5520	0.5074
	SASA	0.5392	0.4300	0.4902	0.3886	0.6840	**0.6361**
	CLIP-Full	0.6880	0.6200	0.6423	0.5755	0.6922	0.4174
	ProS	0.7648	0.7186	0.7228	0.6815	0.8105	0.5298
	UCDR-Adapter	0.7718	0.7263	**0.7391**	**0.6979**	0.8251	0.5392
	DeFT-LoRA(Ours)	**0.7739**	**0.7276**	0.7259	0.6851	**0.8358**	0.5389
Average	SnMpNet	0.3010	0.2418	0.2667	0.2144	0.3378	0.2685
	SASA	0.4388	0.3670	0.3941	0.3296	0.4462	0.4094
	CLIP-Full	0.5230	0.4641	0.4727	0.4175	0.5274	0.3144
	ProS	0.6052	0.5626	0.5546	0.5195	0.6332	0.3994
	UCDR-Adapter	0.6071	0.5655	0.5660	0.5311	0.6370	0.4004
	DeFT-LoRA(Ours)	**0.6212**	**0.5776**	**0.5690**	**0.5331**	**0.6547**	**0.4105**

Table 1: UCDR and U^{d}CDR evaluation results on DomainNet. UCDR has two different gallery settings as we introduced in Preliminaries. The best performance is bolded, and the suboptimal performance is underlined.

Method	Sketchy		TU-Berlin	
	m@200	P@200	m@All	P@100
SnMpNet	0.5781	0.2155	0.3568	0.5226
SASA	0.6910	0.6090	0.4715	0.6682
ProS	0.6991	0.6545	0.6675	0.7442
UCDR-Adapter	0.7285	0.6888	0.6581	0.7317
DeFT-LoRA(Ours)	**0.7331**	**0.6914**	**0.6700**	**0.7476**

Table 2: U^{c}CDR results on Sketchy and TU-Berlin. The best performance is bolded, and the suboptimal is underlined.

Decoupling is Essential: Removing the decoupling mechanism entirely and reverting to a single LoRA adapter (Single LoRA) causes a severe performance drop to an mAP@200 of 0.5909 on Unseen Gallery, a decline of nearly 1.9 percentage points. This empirically confirms that a single adapter cannot resolve the intrinsic conflict of learning both general and specific knowledge, validating our core motivation.

Structured Decoupling is Superior: A Naive LoRA-MoE approach (LoRA-MoE), which lacks our structured Base and Domain Expert design, achieving an mAP@200 of only 0.5970. While better than a single LoRA, it still falls short of

Method	Unseen Gallery		Mixed Gallery	
	m@200	P@200	m@200	P@200
DeFT-LoRA	0.6093	0.5735	0.5521	0.5242
Single-LoRA	0.5909	0.5541	0.5363	0.5064
LoRA-MoE	0.5970	0.5614	0.5398	0.5115
w/o Base Expert	0.6014	0.5681	0.5426	0.5160
w/o Gate	0.6019	0.5655	0.5464	0.5170
Independent A	0.6070	0.5714	0.5477	0.5219

Table 3: Different components evaluated on DomainNet under UCDR setting with Infograph as the unseen domain.

Position	Unseen Gallery		Mixed Gallery	
	m@200	P@200	m@200	P@200
zero-shot CLIP	0.5007	0.4474	0.4501	0.3990
Shallow(1-6)	0.5558	0.5150	0.5022	0.4666
Deep(7-12)	0.6068	0.5706	0.5476	0.5212
ALL(DeFT-LoRA)	0.6093	0.5735	0.5521	0.5242

Table 4: Resuls of adaptation within different layers under UCDR setting with Infograph as the unseen domain.

our full model by 1.2 points, highlighting the importance of our three-stage strategy in preventing redundant learning and building upon a shared knowledge foundation.

The Base Expert is Irreplaceable: Removing the Base Expert also significantly degrades performance, with the mAP@200 dropping to 0.6014. The Base Expert acts as a crucial anchor for domain-invariant knowledge and ensures robust generalization, especially for inputs that do not strongly align with any specific domain. Without it, the model's adaptability and robustness are compromised.

Dynamic Gating is Critical: Replacing the dynamic gate with static averaging (w/o Gating) leads to a noticeable performance decline, with the mAP@200 falling to 0.6019. This highlights the necessity of our content-aware gating mechanism, which tailors the expert composition for each input to generate a customized adapter.

Shared-A is Efficient and Effective: Training independent LoRA experts for each domain without sharing the A-matrix (Independent A) slightly hurts performance, resulting in the mAP@200 falling to 0.6070, while substantially increasing the parameter count. This confirms that sharing the A-matrix is not only a parameter-efficient strategy but also encourages beneficial knowledge transfer by enforcing a common low-rank projection space (Tian et al. 2024).

In conclusion, these results collectively affirm that every component of DeFT-LoRA is integral to achieving its state-of-the-art performance and efficiency.

Analysis of Layer-wise Application

To investigate where domain adaptation is most effective within the ViT-B/32 architecture of CLIP, we applied DeFT-LoRA to different sets of encoder layers: all layers (1-12), only the shallow layers (1-6) and only the deep layers (7-12). The results are shown in Table 4. Notably, even when DeFT-LoRA is applied solely to the shallow layers, we observe a substantial improvement over the zero-shot CLIP baseline. However, adapting the deep layers (7-12) proves more beneficial, closely approaching the performance achieved by adapting all layers. This is consistent with the understanding that shallow layers primarily extract low-level, domain-agnostic features, while deeper layers encode high-level semantic representations that are more sensitive to domain shifts (Neyshabur, Sedghi, and Zhang 2020). By focusing adaptation on the deeper layers, DeFT-LoRA effectively aligns semantic information across domains without disrupting foundational features. While deep-layer adaptation accounts for most gains, a modest boost is achieved when all layers are adapted, indicating that slight adjustments throughout the network further enhance performance.

Analysis of Performance and Efficiency

A core motivation for DeFT-LoRA is to achieve state-of-the-art performance without the high computational overhead of previous methods. In Figure, we present a comparative analysis of performance (mAP@200) against both trainable parameters and computational cost (FLOPs) for our method and Prompt-based methods like ProS and UCDR-Adapter. As illustrated in Figure 1, our proposed DeFT-LoRA not only establishes a new SOTA in performance, but also achieves this performance with remarkable efficiency. DeFT-LoRA introduces only 1.29M trainable parameters. This is approximately 45% fewer parameters than UCDR-Adapter (2.36M) and a staggering 93% fewer than ProS (18.3M). This demonstrates the exceptional parameter efficiency of our structured decoupling and Shared-A design. While the MoE architecture requires dynamic, input-dependent routing and thus cannot be merged into the backbone weights prior to inference like a standard LoRA module, its computational overhead is minimal. The gating network is a lightweight MLP, and only a small subset of expert weights are activated for each token. As a result, the increase in FLOPs is negligible compared to the substantial gains in performance. The plot clearly shows that DeFT-LoRA's computational cost is significantly lower than that of prompt-based methods like ProS and UCDR-Adapter, which introduce costly additional modules.

Conclusion

We propose DeFT-LoRA, a novel parameter-efficient framework for universal cross-domain image retrieval that mitigates knowledge conflict. Its core innovation is a Mixture-of-Experts (MoE) architecture that decouples general and domain-specific knowledge via a shared base expert and multiple domain experts. Combined with a dynamic, content-aware gating network, this design adapts efficiently to diverse domains. DeFT-LoRA sets a new state-of-the-art on the DomainNet benchmark, achieving superior accuracy with remarkably fewer trainable parameters. While our MoE-based design introduces a minimal computational overhead during inference, as experts cannot be merged into the backbone like a standard LoRA module, we demonstrate this is a worthwhile trade-off for the substantial gains in adaptability and performance. Future work will involve extending this strategy to Multi-modal Large Language Models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62577001, No. 62576001, No. 62206003, No. U21A20512) and the National Key Research and Development Project (NO. 2018AAA0100105).

References

Abdul Samadh, J.; Gani, M. H.; Hussein, N.; Khattak, M. U.; Naseer, M. M.; Shahbaz Khan, F.; and Khan, S. H. 2023. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. *NeruIPS*, 36: 80396–80413.

Bose, S.; Jha, A.; Fini, E.; Singha, M.; Ricci, E.; and Banerjee, B. 2024. Stylip: Multi-scale style-conditioned prompt learning for clip-based domain generalization. In *WACV*, 5542–5552.

Cai, Y.; Liu, Y.; Zhang, Z.; and Shi, J. Q. 2024. Clap: Isolating content from style through contrastive learning with augmented prompts. In *ECCV*, 130–147. Springer.

Chen, Q.; Wang, C.; Wang, D.; Zhang, T.; Li, W.; and He, X. 2025. Lifelong knowledge editing for vision language models with low-rank mixture-of-experts. In *CVPR*, 9455–9466.

Cheng, D.; Xu, Z.; Jiang, X.; Wang, N.; Li, D.; and Gao, X. 2024. Disentangled prompt representation for domain generalization. In *CVPR*, 23595–23604.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2024. Qlora: Efficient finetuning of quantized llms. *NeurIPS*, 36.

Dong, W.; Zhang, X.; Chen, B.; Yan, D.; Lin, Z.; Yan, Q.; Wang, P.; and Yang, Y. 2024. Low-rank rescaled vision transformer fine-tuning: A residual design approach. In *CVPR*, 16101–16110.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Dou, S.; Zhou, E.; Liu, Y.; Gao, S.; Zhao, J.; Shen, W.; Zhou, Y.; Xi, Z.; Wang, X.; Fan, X.; et al. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*, 4(7).

Eitz, M.; Hildebrand, K.; Boubekeur, T.; and Alexa, M. 2010. An evaluation of descriptors for large-scale image retrieval from sketched feature lines. *Computers & Graphics*, 34(5): 482–498.

Fang, K.; Song, J.; Gao, L.; Zeng, P.; Cheng, Z.-Q.; Li, X.; and Shen, H. T. 2024. Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. In *CVPR*, 17292–17301.

Han, Z.; Zhu, B.; Xu, Y.; Song, P.; and Yang, X. 2025. Benchmarking and Bridging Emotion Conflicts for Multimodal Emotion Reasoning. In *ACM MM*, 5528–5537.

Hayou, S.; Ghosh, N.; and Yu, B. 2024. LoRA+: Efficient Low Rank Adaptation of Large Models. In *ICML*.

He, H.; Cai, J.; Zhang, J.; Tao, D.; and Zhuang, B. 2023. Sensitivity-Aware Visual Parameter-Efficient Fine-Tuning. In *ICCV*, 11791–11801.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *ECCV*, 709–727. Springer.

Jiang, H.; Cheng, Z.-Q.; Moreira, G.; Zhu, J.; Sun, J.; Ren, B.; He, J.-Y.; Dai, Q.; and Hua, X.-S. 2025. UCDR-Adapter: Exploring Adaptation of Pre-Trained Vision-Language Models for Universal Cross-Domain Retrieval. In *WACV*, 5429–5438. IEEE.

Khattak, M. U.; Rasheed, H.; Maaz, M.; Khan, S.; and Khan, F. S. 2023. Maple: Multi-modal prompt learning. In *CVPR*, 19113–19122.

Komatsuzaki, A.; Puigcerver, J.; Lee-Thorp, J.; Ruiz, C. R.; Mustafa, B.; Ainslie, J.; Tay, Y.; Dehghani, M.; and Houlsby, N. 2022. Sparse Upcycling: Training Mixture-of-Experts from Dense Checkpoints. In *ICLR*.

Kumar, A.; Raghunathan, A.; Jones, R. M.; Ma, T.; and Liang, P. 2022]. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *ICLR*.

Liu, L.; Shen, F.; Shen, Y.; Liu, X.; and Shao, L. 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *CVPR*, 2862–2871.

Liu, S.-y.; Wang, C.-Y.; Yin, H.; Molchanov, P.; Wang, Y.-C. F.; Cheng, K.-T.; and Chen, M.-H. 2024. DoRA: Weight-Decomposed Low-Rank Adaptation. In *ICML*.

Meng, F.; Wang, Z.; and Zhang, M. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *arXiv preprint arXiv:2404.02948*.

Mondal, B.; and Biswas, S. 2022. Seic: Semantic embedding with intermediate classes for zero-shot domain generalization. In *ACCV*, 789–806.

Neyshabur, B.; Sedghi, H.; and Zhang, C. 2020. What is being transferred in transfer learning? *NeurIPS*, 33: 512–523.

Pan, H.; Cao, Y.; Wang, X.; Yang, X.; and Wang, M. 2024. Finding and editing multi-modal neurons in pre-trained transformers. In *ACL*, 1012–1037.

Paul, S.; Dutta, T.; and Biswas, S. 2021. Universal cross-domain retrieval: Generalizing across classes and domains. In *ICCV*, 12056–12064.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment matching for multi-source domain adaptation. In *ICCV*, 1406–1415.

Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C. P.; Wang, X.-Z.; and Wu, Q. J. 2022. A review of generalized zero-shot learning methods. *TPAMI*, 45(4): 4051–4070.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PmLR.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 35(4): 1–12.

Si, C.; Wang, X.; Yang, X.; Xu, Z.; Li, Q.; Dai, J.; Qiao, Y.; Yang, X.; and Shen, W. 2025. Maintaining Structural Integrity in Parameter Spaces for Parameter Efficient Fine-tuning. In *ICLR*.

Tian, C.; Shi, Z.; Guo, Z.; Li, L.; and Xu, C.-Z. 2024. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *NeurIPS*, 37: 9565–9584.

Tian, J.; Xu, X.; Wang, K.; Cao, Z.; Cai, X.; and Shen, H. T. 2022. Structure-aware semantic-aligned network for universal cross-domain retrieval. In *SIGIR*, 278–289.

Wang, M.; Su, H.; Wang, S.; Wang, S.; Yin, N.; Shen, L.; Lan, L.; Yang, L.; and Cao, X. 2025a. Graph Convolutional Mixture-of-Experts Learner Network for Long-Tailed Domain Generalization. *TCSVT*.

Wang, S.; ALuSi; Yang, X.; Xu, K.; Tan, H.; and Zhang, X. 2024. Dual-stream Feature Augmentation for Domain Generalization. In *ACM MM*, 1111–1119.

Wang, S.; He, H.; Yang, X.; Liu, Z.; Zhong, Y.; Zhang, X.; and Wang, M. 2025b. Exploring Invariance Matters for Domain Generalization. *TIP*.

Wang, S.; Shen, X.; Yang, X.; Xu, K.; and Zhang, X. 2025c. Feature Responsive LoRA: Towards Parameter-Efficient Transfer Learning for Self-Supervised Visual Models. *IEEE TCSVT*.

Wang, Y.; Lin, Y.; Zeng, X.; and Zhang, G. 2023. Multilora: Democratizing lora for better multi-task learning. *arXiv preprint arXiv:2311.11501*.

Wu, J.; Hu, X.; Wang, Y.; Pang, B.; and Soricut, R. 2024. Omni-smola: Boosting generalist multimodal models with soft mixture of low-rank experts. In *CVPR*, 14205–14215.

Wu, X.; Huang, S.; and Wei, F. 2024. Mixture of LoRA Experts. In *ICLR*.

Xu, Z.; Qu, B.; Qi, Y.; Du, S.; Xu, C.; Yuan, C.; and Guo, J. 2025. ChartMoE: Mixture of Diversely Aligned Expert Connector for Chart Understanding. In *ICLR*.

Yang, X.; Chang, T.; Zhang, T.; Wang, S.; Hong, R.; and Wang, M. 2024a. Learning hierarchical visual transformation for domain generalizable visual matching and recognition. *IJCV*, 132(11): 4823–4849.

Yang, X.; Feng, F.; Ji, W.; Wang, M.; and Chua, T.-S. 2021. Deconfounded video moment retrieval with causal intervention. In *SIGIR*, 1–10.

Yang, X.; Wang, S.; Dong, J.; Dong, J.; Wang, M.; and Chua, T.-S. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE TIP*, 31: 1204–1216.

Yang, X.; Zeng, J.; Guo, D.; Wang, S.; Dong, J.; and Wang, M. 2024b. Robust video question answering via contrastive cross-modality representation learning. *Science China Information Sciences*, 67(10): 202104.

Yang, Y.; Jiang, P.-T.; Hou, Q.; Zhang, H.; Chen, J.; and Li, B. 2024c. Multi-task dense prediction via mixture of low-rank experts. In *CVPR*, 27927–27937.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *CVPR*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *IJCV*, 130(9): 2337–2348.