Unsupervised Domain Adaptive Visual Question Answering in the era of Multi-modal Large Language Models

Weixi Weng[1,*], Rui Zhang[2], Xiaojun Meng[3], Jieming Zhu[3], Qun Liu[3], Chun Yuan[1,†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University,
[2]School of Computer Science & Tech, Huazhong University of Science and Technology (ruizhang.info),
[3]Huawei Noah Ark's Lab
[*]wengwx22@mails.tsinghua.edu.cn, [†]yuanc@sz.tsinghua.edu.cn

Abstract

*Unsupervised domain adaptation (UDA) for visual question answering (VQA) has attracted research interest. However, with Multi-modal Large Language Models (MLLMs) showing great performance on VQA datasets, UDA for VQA based on MLLMs remains unexplored. To fill this gap, we propose the first systematic approach to **U**nsupervised **D**omain **A**daptation VQA based on **M**LLMs (UDAM). First, we introduce semantic context feature alignment and domain query feature alignment, which utilize a single token embedding for each modality to capture contextual domain information from unimodal inputs and conduct coarse-grained feature alignment on it, thus alleviating domain shifts in the unimodal feature space. Second, we propose the novel semantics-guided query feature alignment, which differentiates important domain-specific queries from learnable query outputs and conducts fine-grained feature alignment controlled by a semantics-guided weight map to reduce domain shifts in the cross-modal feature space. Third, we devise a pair-wise domain-aware prompt strategy, which aids UDA by prompting MLLMs to discern the commonality of tasks and the distinctiveness of domains in multi-modal inputs. Extensive experiments demonstrate UDAM's effectiveness in adapting MLLMs to unlabeled new domains.*

1. Introduction

Unsupervised domain adaptation (UDA) for visual question answering (VQA) has garnered much research attention [6, 51, 58, 59], focusing on specialized VQA models such as LXMERT [43], MCAN [56] *etc,*. The key of UDA for VQA is reducing domain shifts between the source and target domains in both image and text modalities by facilitating the acquisition of domain-invariant features, enabling the models to achieve optimal performance on the unlabeled target domain with labeled source domain data.

However, with Multi-modal Large Language Models (MLLMs) [1,9,29] showing excellent performance on VQA datasets, to the best of our knowledge, UDA research for VQA based on MLLMs remains unexplored and we aim to fill this gap. Investigating UDA within the context of MLLMs is not only a theoretical endeavor but also carries significant practical implications. Adapting MLLMs to specific domains often requires fine-tuning them on corresponding labeled data, while labeling multi-modal data is resource-intensive. UDA fine-tuning proposes a resource-friendly way to adapt MLLMs to specific, unlabeled, and large-sized target domains by leveraging labeled data from a generic and small-sized source domain, as annotations for these source domains are more accessible.

Take the widely studied BLIP2 [29] as an example; UDA based on such kind of MLLMs faces the following three challenges. **C1:** Given the domain shifts existing in the unimodal feature space of both two modalities among various VQA datasets and the different ways that BLIP2 handles text and image encoding, the first challenge lies in designing tailored unimodal feature alignment methods for both image and text modalities. **C2:** A set of pretrained learnable queries interact with multi-modal inputs in the Q-Former to generate cross-modal queries, resulting in a considerable domain shift in the cross-modal feature space. It's intuitive that each learnable query has a different perceptual capability to multi-modal inputs from various domains. The learnable queries that are more sensitive to multi-modal inputs from a specific domain and extract more important domain information are experimentally proven to significantly contribute to the domain shift in the cross-modal feature space. It is important to differentiate these queries from learnable query outputs and utilize them for effective cross-modal feature alignment. **C3:** Since hard prompt strategies greatly impact MLLMs' zero-shot as well as fine-tuning performance, it's necessary to design a specialized hard prompt strategy to aid in UDA fine-tuning.

To address the challenges above, we propose the first systematic approach to **U**nsupervised **D**omain **A**daptation VQA based on **M**LLMs (UDAM). For **C1**, we introduce semantic context feature alignment (SCFA) and domain query feature alignment (DQFA) tailored for text and image modality, which utilize a single token embedding for each modality to capture global-level contextual domain information of unimodal input and conduct coarse-grained feature alignment on it, thus alleviating domain shifts in the unimodal feature space of each modality. For **C2**, we propose the novel semantics-guided query feature alignment (SQFA), which differentiates important domain-specific queries from learnable query outputs and conducts fine-grained feature alignment controlled by a semantics-guided weight map, thus reducing the domain shift in the cross-modal feature space. For **C3**, we devise a pair-wise domain-aware prompt strategy, which aids UDA for VQA by prompting MLLMs to discern the commonality of tasks and the distinctiveness of domains among diverse multi-modal inputs. Our contributions are summarized as follows:

- We introduce two coarse-grained unimodal feature alignment methods, including semantic context feature alignment (SCFA) and domain query feature alignment (DQFA) to efficiently alleviate the domain shift in the unimodal feature space of each single modality.
- We further propose a novel fine-grained cross-modal feature alignment method named semantics-guided query feature alignment (SQFA) to reduce the domain shift in the cross-modal feature space.
- We devise a simple yet practical pair-wise domain-aware prompt strategy, which aids UDA for VQA on MLLMs and demonstrates great research potential.
- Extensive experiments demonstrate the effectiveness of UDAM. UDAM provides up to a 1.98% accuracy improvement compared to conventional fine-tuning and a 14.66% improvement compared to BLIP2's original zero-shot accuracy on the CLEVR dataset.

2. Related Work

2.1. Multi-modal Large Language Models.

Research on MLLMs for various applications [32,41,63] has been flourishing recently. MLLMs [1, 9, 29, 55] usually develop various types of multi-modal perceivers to transform multi-modal inputs into soft prompts, which are fed into LLMs along with the text input for answer generation.

BLIP2 [29] introduces Q-Former as its multi-modal perceiver. Many subsequent works [15,21,30,62] have demonstrated the powerful capabilities of Q-Former. Q-Former takes the image, text, and a set of learnable queries as input. The learnable queries capture text-relevant features of the image during the forward process of Q-Former.

Some other work [1,4,53] employs simpler multi-modal perceivers, even integrates visual feature into different layers [52], and train the LLM to enhance its visual capabilities. Compared to these efforts, BLIP2 effectively models the image-text alignment with the Qformer module during pretraining and requires training only the Qformer during fine-tuning, which is more cost-effective. This makes it a more suitable starting point for conducting our research.

2.2. Unsupervised Domain Adaptation.

UDA has been investigated across various tasks, and the two most mainstream UDA methods are self-training methods and feature alignment methods. The former focuses on how to generate high-quality pseudo labels for unlabeled target data [5, 10, 14]. The latter can be further classified into two categories based on their underlying principles: one category aims to reduce domain shift by reducing various measures of distributional disparities such as the Maximum Mean Discrepancy (MMD) [46], CoRAL [42], Wasserstein Distance [40] *etc.*, while the other involves domain-adversarial training of neural networks (DANN) [7,45,49,50], where models learn domain-invariant features itself by competing against domain discriminators during training. We refer to methods based on the above two principles as discrepancy-based and adversarial-based methods.

Unimodal UDA on large-scale pretrained models is still in its early stage, revolving around the above two kinds of methods. A line of self-training methods [14, 26, 27] mainly focus on image classification, leveraging the zero-shot ability of CLIP [36] to generate pseudo labels for unlabeled target data; Saad-Falcon et al. [38] utilizes LLMs to create synthetic queries for target domain in passage retrieval. The major obstacle to applying self-training methods to VQA lies in the difficulty of reasonably evaluating and filtering the model's answers. In terms of feature alignment methods, Malik et al. [34] integrate Adapters [20] into LLMs and achieve UDA by reducing the domain discrepancy measured by MK-MMD [33]. Meanwhile, Yang et.al [54] apply adversarial-based methods to visual pretrained models. The above work provides insights for us to explore the application of feature alignment methods on MLLMs.

Multi-modal UDA for VQA has received extensive attention. Chao et al. [6] demonstrates that there are considerable domain shifts in both text and image modalities across various VQA datasets and use DANN to reduce domain shifts between domains to achieve UDA. Xu et al. [51] propose a multi-modal UDA framework for open-ended VQA, which minimizes MMD of unimodal features between domains and utilizes DANN [12] to conduct feature alignment on fused features. The effectiveness of this approach is validated only in the scenarios between VQA 2.0 [16] and

Table 1. The names and formats of the hard prompt strategies we investigate on VQA.

Strategy Name	Strategy Format
qa [29]	Question: ⟨question⟩ Answer:
Short-qa [3]	Question: ⟨question⟩ Short answer:
following-qa [3]	Answer the following question. ⟨question⟩
task-qa [29]	Visual Question Answering. Question: ⟨question⟩ Answer:
task-domain-qa	Visual Question Answering: ⟨domain-specific style description⟩. Question: ⟨question⟩ Answer:

Vizwiz [19], which also poses certain limitations. Zhang et al. [59] employs DANN on the visual encoder of VQA models and does not investigate the domain shift in the text modality. Zhang et al. [58] explore the impact of MMD, self-supervised reconstruction, and supervised auxiliary co-training based on two 2019 VQA models [43, 57].

The aforementioned works on UDA for VQA validate their methods in different cross-domain scenarios, with both the source and target domains [2, 16, 24, 37, 60] utilizing similar real-world image sources *i.e.* the COCO dataset, which results in little domain shift in the image modality within their cross-domain scenarios and makes their methods less convincing. To address this issue, we carefully devise four cross-domain scenarios with more pronounced domain shifts in the image modality, which is further illustrated in Section 5.2. More importantly, previous studies have not explored MLLMs, and we also aim to fill this gap.

2.3. Prompt Engineering of Multi-modal Large Language Models

The effectiveness of prompt engineering [31] has been demonstrated in enabling pretrained models to achieve better performance on new tasks or in new data domains. [17] conducts a systematic survey of prompt engineering on MLLMs. This section focuses on illustrating the use of hard prompts. Different MLLMs [1, 29, 48] meticulously design various hard prompts for multi-modal tasks to achieve optimal zero-shot inference performance. Huang et al. [23] improves zero-shot image classification by employing hard prompts along with specific class descriptions. Awal et al. [3] systematically investigate the efficacy of various hard prompt techniques for zero-shot VQA inference on BLIP2, which includes task-specific instruction, in-context exemplars, and various informative captions [11, 18, 22, 44].

3. Problem Definition.

Firstly, a regular VQA dataset is defined as $D = \{(I_i, Q_i), A_i\}_{i=1}^{n}$ where i-th sample contains an image I_i and a corresponding question Q_i. A_i is the desired answer given a combination of an image and a question. Then, in the setting of UDA, we further use $D_S = \{(I_i^s, Q_i^s), A_i^s\}_{i=1}^{n_s}$ to denote the labeled source dataset with n_s labeled samples. The unlabeled target dataset is denoted as $D_T = \{(I_i^t, Q_i^t)\}_{i=1}^{n_t}$. We follow the open-ended setting of VQA, where no answer candidates are provided, and the model can respond with any free-form answers. We aim to utilize D_S and D_T to fine-tune MLLMs to enable them to achieve optimal performance on the target domain.

4. Methods

In Section 4.1, we introduce two coarse-grained unimodal feature alignment methods to alleviate the domain shifts in the unimodal feature space of each modality. In Section 4.2, we investigate how learnable queries interact with multi-modal inputs and propose a novel fine-grained cross-modal feature alignment method to reduce the domain shift in the cross-modal feature space. In Section 4.3, we illustrate our devised pair-wise domain-aware prompt strategy.

4.1. Coarse-grained Unimodal Feature Alignment

Given that domain shift exists in the unimodal feature space of both modalities among different VQA datasets [6, 14], we first propose two coarse-grained unimodal feature alignment methods for text and image modality, respectively. We use the unimodal masking strategy in Figure 1 to control the self-attention mechanism of the Q-former, preventing text input and learnable queries from interacting with each other to obtain unimodal feature outputs.

For the text modality, we introduce **semantic context feature alignment (SCFA)**. The first token of the textual input to the Q-Former is always designated as the [CLS] token in BLIP2. Thanks to the pretraining process on extensive data, the [CLS] embedding output of the Q-Former (denoted by *t*) represents the contextual semantics of the given text input and should contain valuable domain information of the text modality. We input *t* into a single fully connected layer f_t before feeding it into a domain discriminator D_{uni} to distinguish its domain. We adopt a binary cross-entropy (BCE) loss as the domain classification loss of SCFA, which can formulated as follows:

$$\mathcal{L}_{SCFA} = -d log D_{uni}(f_t(t)) - (1-d) log(1 - D_{uni}(f_t(t))) \quad (1)$$

where the domain label d is 0 for the source domain and 1 for the target domain.

For the image modality, we introduce **domain query feature alignment (DQFA)** inspired by [28, 49]. In the Q-Former, a set of learnable queries is used to cross-attend

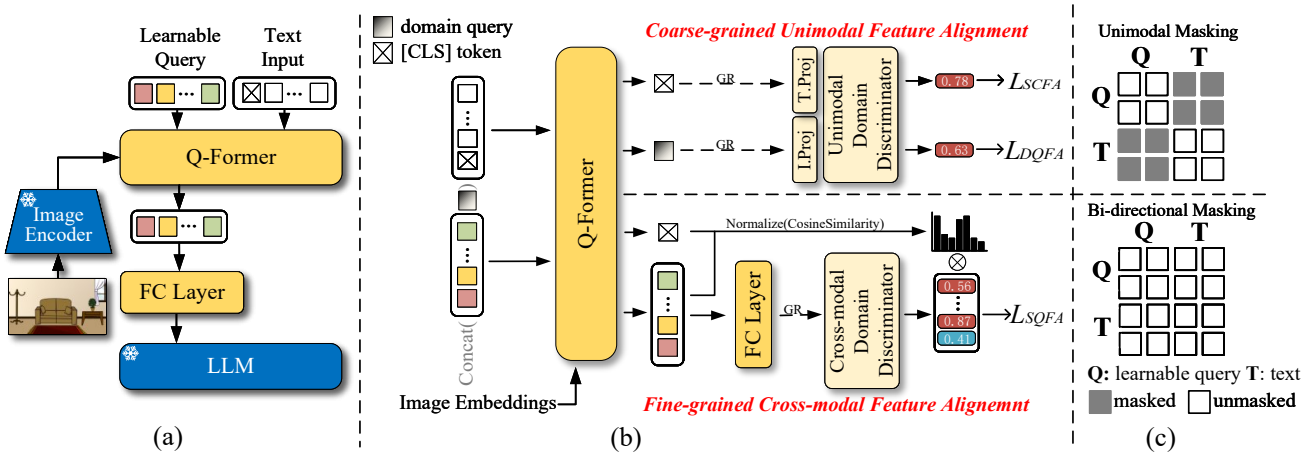

Figure 1. Subfigure (a) depicts the regular finetuning paradigm of BLIP2. UDAM only requires simple modifications to the Q-Former, as illustrated in Subfigure (b). GR means reversing the sign of gradients during backpropagation. FC Layer refers to the fully connected layer. Both coarse-grained unimodal feature alignment and fine-grained cross-modal feature alignment share the same input format and model parameters, employing different self-attention masking strategies between learnable queries and text to control their interaction, which is shown in Subfigure (c).

to the image embeddings generated by a frozen image encoder, generating unimodal queries representing image information. We first define an additional learnable domain query token dq tasked with capturing domain information from image embeddings. The weights of the domain query dq are initialized by averaging the weights of all learnable queries, ensuring that the domain query possesses the capability to capture valuable information in the early stages. dq is appended to the learnable queries to form the input for Q-Former, only attending to other learnable queries in the forward process. We denote the output of Q-Former as $[\{\tilde{q}_i\}^n : \tilde{dq}]$. In this way, the domain query gathers global domain information of the image, while the learnable queries extract discriminative features. The weights of the learnable queries are frozen during this process to preserve their discriminability. $\tilde{dq}$ is then fed to a fully connected layer f_i as well as the domain discriminator D_{uni}. The domain classification loss of DQFA is as follows.

$$\mathcal{L}_{DQFA} = -d log D_{uni}(f_i(\tilde{dq})) - (1-d) log(1 - D_{uni}(f_i(\tilde{dq}))) \quad (2)$$

In summary, we aim to perform coarse-grained feature alignment to alleviate the domain shift in the unimodal feature space of each modality. We introduce two efficient feature alignment methods for text and image modalities, which both utilize a single token embedding to capture global-level contextual domain information and conduct feature alignment on it by adversarial training, *i.e.* the Q-Former competing against D_{uni} during training.

4.2. Fine-grained Cross-modal Feature Alignment

When adopting the bi-directional masking strategy, a set of pretrained learnable queries interact with multi-modal inputs in Q-Former to generate cross-modal queries whose feature space exhibits non-negligible domain shift. In this section, we will demonstrate our proposed novel fine-grained cross-modal feature alignment method.

When dealing with multi-modal inputs from a specific dataset, it is intuitive that the learnable query at different positions inherently has different sensitivity to inputs; certain queries may be more sensitive to multi-modal inputs of a specific domain and capture more valuable domain information, indicating their greater domain specificity. We experiment on Vicuna7B-based BLIP2 in the AOKVQA→VQA-Abstract scenario to validate this intuitive idea. For each image-question pair, we feed it into Q-Former and calculate the cosine similarities between the 32 cross-modal queries and the [CLS] embedding output, *i.e.* the contextual semantics of the multi-modal input. We compute the average similarities of all 32 queries on the AOKVQA validation set and rank them in descending order. We find that approximately 12.5% of the cross-modal queries exhibited significant semantic cosine similarities. Contrastively, the similarities of the other cross-modal queries are very small and even less than 0. We further present t-SNE visualizations [47] of all cross-modal queries and the TOP K=4 domain-specific queries with the highest similarities between the two domains and find that the domain-specific queries contribute greatly to the domain shift in the cross-modal feature space. We believe that differentiating domain-specific queries from cross-modal queries is the key to fine-grained cross-modal feature alignment.

Given the analysis above, we propose a fine-grained cross-modal feature alignment method named **semantics-guided query feature alignment (SQFA)** to mitigate the

Table 2. Zero-shot inference VQA accuracy of different prompt strategies on four VQA datasets based on BLIP2 models with decoder-based Vicuna7B [61] and encoder-decoder-based FlanT5XL [8].

LLM	Prompt Strategy	ArtVQA	AOKVQA	CLEVR	VQA-Abstract
Vicuna7B	qa [29]	4.25	12.34	23.55	48.54
	short-qa [3]	2.45	7.04	26.68	51.16
	following-qa [3]	3.90	14.32	23.07	43.25
	task-qa [9]	5.03	11.49	29.44	51.29
	task-domain-qa	**5.80**	**16.36**	**31.60**	**53.93**
FlanT5XL	qa [29]	5.23	40.69	33.91	50.62
	short-qa [3]	5.46	42.24	34.09	50.02
	following-qa [3]	5.30	41.54	33.29	49.67
	task-short-qa [9]	5.87	41.55	34.14	50.63
	task-domain-short-qa	**6.26**	**42.30**	**34.29**	**50.68**

domain shift in the cross-modal feature space controlled by a semantics-guided weight map. In a forward pass of Q-Former with the bi-directional self-attention masking strategy shown in Figure 1, the learnable queries are able to interact with the textual inputs, thus generating cross-modal queries. Under this setup, the [CLS] token still represents the contextual semantics of the entire multi-modal inputs in the encoding process, and we compute the cosine similarity between each cross-modal query and the [CLS] embedding output. We then calculate a semantic-guided weight map W based on the computed similarities as follows: queries with a cosine similarity less than 0 are assigned a weight of 0, while the weights for the other queries are obtained by sum normalization based on their cosine similarities.

We then project all cross-modal queries by the pretrained fully connected (FC) layer after Q-Former, in which the cross-modal queries undergo dimensional expansion and further enrich their semantic information. Since the output of the FC layer, denoted as $\{\hat{q_i}\}^n$, is directly used as the soft prompts for the downstream LLM, we believe that performing fine-grained domain alignment on these embeddings would be beneficial. We feed $\{\hat{q_i}\}^n$ into a domain discriminator D_{cross} to get their classification losses. We also employ the BCE loss here. We multiply the classification losses of the cross-modal queries obtained in this step by the semantics-guided weight map to obtain the final loss of SQFA, which is formulated as follows:

$$L_{SQFA} = -\frac{1}{n}\sum_{i=1}^{n} w_i[dlogD_{cross}(\hat{q_i}) + (1-d)log(1-D_{cross}(\hat{q_i}))] \quad (3)$$

where w_i refers to the i-th output cross-modal query's corresponding weight. In summary, our proposed SQFA differentiates domain-specific queries from learnable query outputs and conducts fine-grained cross-modal feature alignment controlled by a semantics-guided weight map, thus reducing the domain shift in the cross-modal feature space.

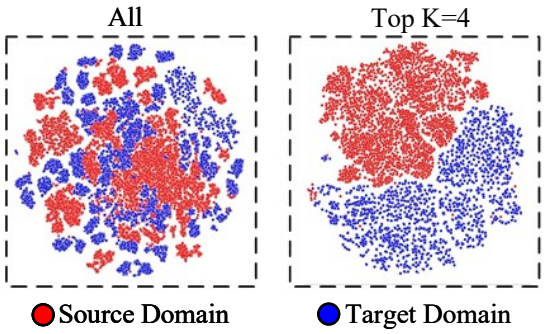

Figure 2. The t-SNE visualization results of all cross-modal queries as well as the TOP K=4 queries with the highest average semantic similarities in the AOKVQA → VQA-Abstract scenario. We can observe that queries with higher similarities exhibit more pronounced distribution differences between the two domains.

4.3. Pair-wise Domain-aware Prompt Strategy

A suitable hard prompt strategy can lead to significant performance improvements in both unsupervised zero-shot inference (target domain) and supervised fine-tuning (source domain) settings [3]. Therefore, it is crucial to investigate how different prompt strategies impact MLLMs' UDA fine-tuning. We first investigate the impacts of several of the most widely used prompt strategies on MLLMs' zero-shot inference and UDA fine-tuning and then devise a simple and practical pair-wise domain-aware prompt strategy to aid MLLMs' UDA finetuning.

Our devised pair-wise domain-aware prompt strategy consists of three components, which are demonstrated as follows: **(1)** Domain-agnostic task instruction. We use "Visual Question Answering" for VQA. **(2)** Domain-specific style description. We carefully choose a style description for the images in each dataset, such as "abstract images" for VQA-Abstract. Detailed descriptions of datasets are provided in the supplementary material. To our knowledge, we are the first to explore the impact of image style descriptions on zero-shot VQA inference. **(3)** Formatted textual input of each sample. We format the original question input by the

Table 3. The upper right part of the table displays the calculated Maximum Mean Discrepancy (MMD) of the visual features of the randomly selected samples of each pair of datasets, whereas the lower left part presents the MMD results of the textual features.

	OKVQA	AOKVQA	VQA-Abstract	CLEVR
OKVQA	-	0.063	0.096	0.256
AOKVQA	0.192	-	0.095	0.285
VQA-Abstract	0.224	0.062	-	0.288
CLEVR	0.194	0.064	0.065	-

strategy "Question: ⟨question⟩ Answer:"

When conducting UDA fine-tuning, we keep the task description unchanged between domains and use different domain-specific style descriptions for source and target domains, prompting MLLMs to discern the commonality of tasks and the distinctiveness of domains between diverse inputs. During inference, we use the corresponding prompt strategy of the target domain. This prompt strategy proves to be effective in both zero-shot inference and UDA fine-tuning and is also adaptable, as each component can be modified according to data changes.

4.4. Overall Training Strategy

In this section, we illustrate our UDA fine-tuning strategy based on MLLMs. We first reconstruct the question input of the source and target datasets using our pair-wise domain-aware prompt strategy. In terms of the labeled source domain data, we ask the MLLM to give free-form answers and calculate the language modeling loss L_{lm} with the predictions and the corresponding ground-truth answers. For the target domain data ($d = 1$), since there are no ground-truth labels, $L_{lm}(F) = 0$. For data from both domains, we compute the corresponding losses for SCFA, DQFA, and SQFA and control the weight ratio between them through a hyperparameter α. The adversarial training loss L_{UDAM} can be formulated as follows:

$$L_{UDAM} = \alpha L_{SQFA} + (1 - \alpha)(L_{SCFA} + L_{DQFA}) \quad (4)$$

To summarize, the overall training objective is defined as:

$$\min_{F} \max_{D} L_{lm}(F) - \lambda L_{UDAM}(F, D) \quad (5)$$

where D represents the domain discriminators as well as f_t and f_i. F is a unified representation of the original trainable modules in the MLLM. λ stands for the weight of the adversarial training loss L_{UDAM}.

5. Experiments

5.1. Datasets

OKVQA [35] is currently the largest knowledge-based VQA dataset, having 14031 images and 14055 questions that require a variety of external knowledge resources to answer. **AOKVQA** [39] is the successor of OKVQA, which demands models' visual-grounding reasoning ability and also requires rich external knowledge to answer its questions. **CLEVR** [25] consists of synthetic images featuring various objects, shapes, colors, and spatial relationships. The dataset encourages models to understand and reason about complex scenes through questions that involve compositional structures and relationships. **VQA-Abstract** [16] is created with images depicting abstract scenes. Answering questions in this dataset emphasizes a comprehensive and detailed understanding of the abstract scenes. **ArtVQA** [13] requires not only great comprehension of the given paintings but also a profound knowledge of art history.

5.2. Cross-Domain Scenarios

OKVQA [35] and AOKVQA [39] are currently the most widely used knowledge-based VQA datasets. We utilize them as our source domain datasets because these two datasets are relatively small in scale and contain a significant amount of external knowledge.

We choose CLEVR [25] and VQA-Abstract [16] as our target domain datasets based on the following considerations: 1. They exhibit significant differences in the image domain compared to common VQA datasets, which presents challenges for UDA algorithms; 2. As shown by Figure 4, since they have a much larger number of samples than the source datasets, target performance gains can intuitively reflect the improvement of MLLMs' capabilities.

To study the domain shifts among the above four datasets, we utilize a pre-trained Qformer to extract features from 5,000 randomly selected samples in each dataset. Textual features are represented by the output of the [CLS] token when only questions are input. When only an image is provided, the average of all learnable query outputs is used as the visual features representation. We employ MMD to measure the differences in both textual and visual features between any two datasets, with the results depicted in Table 3. The results indicate that the four datasets exhibit certain domain shifts in the two modalities. Therefore, we design the following four cross-domain scenarios to carry out our experiments: (1) OKVQA → VQA-Abstract; (2) AOKVQA → VQA-Abstract; (2) OKVQA → CLEVR; (4) AOKVQA → CLEVR.

Table 4. The accuracies of fine-tuned BLIP2-Vicuna7B with different prompt strategies and prevalent UDA methods in four cross-domain scenarios. The numbers in parentheses after the dataset names indicate the number of samples in the training set of the dataset.

		BLIP2(source)	BLIP2+DANN	BLIP2+MMD	BLIP2+MK-MMD	BLIP2+UDAM
OKVQA(9009)→ VQA-Abstract(60000)						Metric: VQA Accuracy (%)
1	qa	53.91	54.35	54.43	54.59	54.66
2	task-qa	54.05	54.10	54.16	54.25	54.56
3	task-domain-qa	54.43	54.33	54.40	54.37	**55.11**
AOKVQA(17056)→ VQA-Abstract(60000)						Metric: VQA Accuracy (%)
4	qa	58.10	58.36	58.91	58.87	59.10
5	task-qa	58.23	58.42	58.34	58.27	58.74
6	task-domain-qa	58.35	58.98	58.95	58.76	**59.57**
OKVQA(9009)→ CLEVR(699989)						Metric: Exact Match (%)
7	qa	35.89	35.26	35.53	35.78	36.43
8	task-qa	35.71	36.52	36.41	36.02	36.60
9	task-domain-qa	37.42	37.32	37.36	37.23	**38.01**
AOKVQA(17056)→ CLEVR(699989)						Metric: Exact Match (%)
10	qa	36.91	37.22	36.33	36.87	37.60
11	task-qa	36.38	36.29	36.01	35.77	36.71
12	task-domain-qa	37.01	37.77	36.27	36.87	**38.21**

5.3. Baselines

UDA for VQA based on MLLMs is unexplored to the best of our knowledge. Inspired by previous works on UDA for VQA [6, 51, 58, 59] which mainly study the impacts of MMD [46] and DANN [12] on various VQA models, we implement the following baselines for fair comparisons: 1. BLIP2 trained only on source data, showcasing the inherent cross-domain capabilities of BLIP2; 2. BLIP2+DANN. This baseline conducts feature alignment on the final cross-modal queries with a domain discriminator, representing a straightforward adversarial-based feature alignment method. 3. BLIP2+MMD. Inspired by Zhang et al. [58], we curate a baseline that minimizes the Maximum Mean Divergence (MMD) of the cross-modal queries between two domains, representing a discrepancy-based feature alignment method. 4. BLIP2+MK-MMD. Influenced by Long et al. [33], we extended baseline 3 to implement a version featuring multi-kernel MMD. We also compare the effects of different prompt strategies on MLLMs' UDA fine-tuning with various UDA methods.

5.4. Evaluations

To evaluate VQA performance, we utilize the standard VQA score [35]. Given a predicted answer a for a single question q, assuming there are N ground-truth answers for q, with $S(a)$ of them being identical to the answer a, the accuracy of a for the question q can be calculated as $Acc(a, q) = \min(\frac{S(a)}{3}, 1)$. For VQA datasets [25] that only have one answer for each question, we use the percentage of questions with predicted answers that exactly match the unique standard answer for VQA accuracy (*i.e.* the Exact Match metric). Before performing string matching, we preprocess predicted answers following Li et al. [29].

5.5. Implementation Details

We mainly use the Vicuna7B-based BLIP2 model and initialize model weights from official pretrained checkpoints. We adopt most of the hyperparameters mentioned in [29] for fine-tuning VQA, except for the following differences: we adopt a batch size of 4, with two samples from each domain in every batch. Following [9], we only fine-tune the parameters of the Q-Former and the FC Layer. We use the image resolution of 224 $\times$ 224 during fine-tuning. We set λ and α to 0.003 and 0.9 in our experiments. All experiments are conducted under the same random seed settings and executed on a single V100 GPU.

5.6. Key Findings from Our Experiments

Q1: How do different hard prompt strategies benefit the model's zero-shot inference? We conduct zero-shot VQA inference experiments using various hard prompt strategies on four different datasets based on two BLIP2 variants, and the results are shown in Table 2. We first test three common prompt strategies on two BLIP2 variants and find that the "qa" and "short-qa" strategies show the best performance on Vicuna7B-based BLIP2 and FlanT5XL-based BLIP2. We can find that incorporating task instruction into the "qa" and "short-qa" strategies generally leads to performance improvements. Further enhancement is observed with the addition of domain-specific style descriptions. Therefore, given the zero-shot inference experiment results on two variants of BLIP2, we have validated that each component of our domain-aware prompt strategy effectively boosts BLIP2's zero-shot inference performance.

Q2: How do different hard prompt strategies benefit the model's UDA fine-tuning? We further perform UDA fine-tuning with three different hard prompt strategies, as shown in Table 4. We first fine-tune BLIP2 merely on the la-

beled source data to explore how different prompt strategies benefit BLIP2's inherent cross-domain ability, as shown in the "BLIP2(source)" column in Table 4. Under this setup, our domain-aware prompt strategy achieves the best post-finetuning performance across all cross-domain scenarios. In the OKVQA $\rightarrow$ CLEVR scenario, using "task-domain-qa" leads to a performance improvement of 1.71% compared to "task-qa" and 1.53% compared to "qa". The results of UDA finetuning with different prompt strategies also demonstrate the necessity and effectiveness of each component in our proposed prompt strategy.

Q3: How do different feature alignment methods enhance the model's target performance? From Table 4, we observe that the three baselines struggle to provide consistent performance improvements across different scenarios and prompt strategies, sometimes even resulting in decreased performance in certain cases. However, UDAM achieves performance gains in most cases.

In all cross-domain scenarios, when employing our proposed prompt strategy for UDA fine-tuning, all three baselines experience a decline in performance, whereas our proposed UDAM still manages to enhance the performance further. When using VQA-Abstract and CLEVR as target domains, in conjunction with our domain-aware prompt strategy, UDAM results in large performance improvements of 1.22% and 1.20%. Zhang et al. [58] manage to bring a 1.3% improvement from a baseline of 40.1% in the VQAv2 [16] $\rightarrow$ VQA-abstract scenario. On the contrary, leveraging AOKVQA, which has only 1/37 of VQAv2's sample size, we achieve a comparable level of improvement from a better baseline (1.22% from 58.35%).

Q4: How do different labeled source domain datasets benefit UDA finetuning? We observed that when the model undergoes UDA fine-tuning with AOKVQA as the source dataset, no matter what the target domain dataset is, the final target performance tends to be higher, and the gains from UDAM are also more significant. We attribute this to the fact that AOKVQA has a larger data size and offers more external knowledge than OKVQA. We believe that in UDA finetuning, the larger the source domain dataset is and the more knowledge it provides, the higher the final target performance and the improvements brought by UDAM.

Ablation studies. Table 5 provides details on the results of the ablation studies. From the 3^{nd}-4^{th} row in Table 5, it can be observed that SCFA and DQFA bring performance improvements of 0.36% and 0.65%, respectively. Their combination leads to a performance gain of 0.74%. It can be observed from the 5^{th}- 7^{th} rows that SQFA also contributes to improving BLIP2's target performance. Without employing SCFA or DQFA, directly using SQFA leads to a performance improvement of 0.78%. Ultimately, UDAM achieves a further improvement of 1.22% on a fine-tuning baseline with an accuracy of 58.35%, surpassing the zero-shot inference accuracy of our domain-aware prompt strategy by 5.64%. Compared to the zero-shot performance of BLIP2 [29] using the "qa" strategy, UDAM significantly boosts BLIP2's performance on VQA-Abstract and CLEVR by 11.03% and 14.66% at a very low cost, demonstrating the feasibility of adapting MLLMs to an unlabeled new domain through UDA finetuning.

Table 5. Ablation studies of UDAM in the AOKVQA $\rightarrow$ VQA-Abstract scenario using our domain-aware prompt strategy. The 2^{nd} is the performance after fine-tuning on only the source domain.

	SCFA	DQFA	SQFA	VQA Accuracy (%)
1	zero-shot inference result			53.93
2				58.35
3	✓			58.71 (+0.36)
4		✓		59.00 (+0.65)
5	✓	✓		59.09 (+0.74)
6			✓	59.13 (+0.78)
7	✓	✓	✓	**59.57 (+1.22)**

Extending to more MLLMs and more tasks We further apply UDAM to BLIVA [21], which is a more performant MLLM, to demonstrate its effectiveness and applicability. The results are shown in Table 6. UDAM is fundamentally applicable to all multimodal tasks that MLLM can handle, and we will explore more tasks in our future work.

Table 6. We present the VQA accuracy of BLIVA-Vicuna7B, trained with different components of UDAM in the AOKVQA $\rightarrow$ VQA-Abstract scenario. The results demonstrate that UDAM effectively enhances BLIVA's performance.

	VQA Accuracy (%)
BLIVA(source)	60.57
BLIVA + SCFA + DQFA	61.45 (+0.88)
BLIVA + SQFA	61.52 (+0.95)
BLIVA + UDAM	61.96 (+1.29)

6. Conclusion

This paper presents a comprehensive investigation into the underexplored area of UDA for VQA based on MLLMs in terms of feature alignment methods and prompt engineering strategies and proposes the first systematic approach called UDAM. UDAM provides a feasible way to adapt MLLMs to unlabeled new domains with very limited data.

7. Acknowledgements

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), SSTIC Grant (KJZD20230923115106012, KJZD20230923114916032, GJHZ20240218113604008), Beijing Key Lab of Networked Multimedia and the National Natural Science Foundation of China under grants 62436003.

References

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

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

[3] Rabiul Awal, Le Zhang, and Aishwarya Agrawal. Investigating prompting techniques for zero-and few-shot visual question answering. *arXiv preprint arXiv:2306.09996*, 2023.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[5] Qi Cai, Yingwei Pan, Chong-Wah Ngo, Xinmei Tian, Lingyu Duan, and Ting Yao. Exploring object relation in mean teacher for cross-domain detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11457–11466, 2019.

[6] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Cross-dataset adaptation for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5725, 2018.

[7] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[9] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[10] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021.

[11] Yifan Du, Junyi Li, Tianyi Tang, Wayne Xin Zhao, and Ji-Rong Wen. Zero-shot visual question answering with language model feedback. *arXiv preprint arXiv:2305.17006*, 2023.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[13] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 92–108. Springer, 2020.

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

[15] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

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

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

[18] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023.

[19] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018.

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[21] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2256–2264, 2024.

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

[23] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

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

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

[26] Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, and Meng 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*, pages 2691–2701, 2024.

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

[28] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[30] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions. *arXiv preprint arXiv:2308.04152*, 2023.

[31] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[32] Qijiong Liu, Jieming Zhu, Yanting Yang, Quanyu Dai, Zhaocheng Du, Xiao-Ming Wu, Zhou Zhao, Rui Zhang, and Zhenhua Dong. Multimodal pretraining, adaptation, and generation for recommendation: A survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6566–6576, 2024.

[33] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015.

[34] Bhavitvya Malik, Abhinav Ramesh Kashyap, Min-Yen Kan, and Soujanya Poria. Udapter–efficient domain adaptation using adapters. *arXiv preprint arXiv:2302.03194*, 2023.

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

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

[37] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28, 2015.

[38] Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Arafat Sultan, and Christopher Potts. Udapdr: Unsupervised domain adaptation via llm prompting and distillation of rerankers. *arXiv preprint arXiv:2303.00807*, 2023.

[39] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022.

[40] Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[41] Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. Pmg: Personalized multimodal generation with large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 3833–3843, 2024.

[42] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. *Domain adaptation in computer vision applications*, pages 153–171, 2017.

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

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

[45] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[46] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[47] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[48] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[49] Wen Wang, Yang Cao, Jing Zhang, Fengxiang He, Zheng-Jun Zha, Yonggang Wen, and Dacheng Tao. Exploring sequence feature alignment for domain adaptive detection

transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1730–1738, 2021.

[50] Weixi Weng and Chun Yuan. Mean teacher detr with masked feature alignment: A robust domain adaptive detection transformer framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5912–5920, 2024.

[51] Yiming Xu, Lin Chen, Zhongwei Cheng, Lixin Duan, and Jiebo Luo. Open-ended visual question answering by multimodal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 367–376, 2020.

[52] Zenan Xu, Xiaojun Meng, Yasheng Wang, Qinliang Su, Zexuan Qiu, Xin Jiang, and Qun Liu. Learning summary-worthy visual representation for abstractive summarization in video. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5242–5250. International Joint Conferences on Artificial Intelligence Organization, 8 2023. Main Track.

[53] Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*, 2024.

[54] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 520–530, 2023.

[55] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[56] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

[57] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

[58] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. How to practice vqa on a resource-limited target domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4451–4460, 2023.

[59] Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, and Rebecca Hwa. Domain-robust vqa with diverse datasets and methods but no target labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7046–7056, 2021.

[60] Zhengkun Zhang, Wenya Guo, Xiaojun Meng, Yasheng Wang, Yadao Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. HyperPELT: Unified parameter-efficient language model tuning for both language and vision-and-language tasks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11442–11453, Toronto, Canada, July 2023. Association for Computational Linguistics.

[61] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[62] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[63] Jieming Zhu, Xin Zhou, Chuhan Wu, Rui Zhang, and Zhenhua Dong. Multimodal pretraining and generation for recommendation: A tutorial. In *Companion Proceedings of the ACM on Web Conference 2024*, pages 1272–1275, 2024.