When Open-Vocabulary Visual Question Answering Meets Causal Adapter: Benchmark and Approach

Feifei Zhang[1*], **Zhaoyi Zhang**[1], **Xi Zhang**[2], **Changsheng Xu**[3, 4, 5]
[1]Tianjin University of Technology
[2]Alibaba Group
[3]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[4]School of Artificial Intelligence, University of Chinese Academy of Sciences
[5]Peng Cheng Laboratory
feifeizhang@email.tjut.edu.cn, zhaoyiz0606@gmail.com, zx443053@alibaba-inc.com, csxu@nlpr.ia.ac.cn

Abstract

Visual Question Answering (VQA) is a multifaceted task that integrates computer vision and natural language processing to produce textual answers from images and questions. Existing VQA benchmarks predominantly adhere to a closed-set paradigm, limiting their ability to address arbitrary, unseen answers, and thus falling short in real-world scenarios. To address this limitation, we introduce the Open-Vocabulary Visual Question Answering (OVVQA) benchmark, specifically designed to evaluate models under open-world conditions by assessing their performance on both base classes (seen, common answers) and novel classes (unseen, rare answers). In conjunction with this benchmark, we propose a model-agnostic Causal Adapter to combat the inherent bias found in current VQA tasks. Our approach leverages front-door adjustment to enhance causal reasoning, significantly improving model performance on novel categories while maintaining accuracy on base classes. Additionally, we introduce an adaptive transfer loss to facilitate the transfer of more knowledge from the pretrained model to our OVVQA task. Extensive experiments across multiple datasets validate the superiority of our method over existing state-of-the-art approaches, demonstrating its robust generalization and adaptability in open-world VQA scenarios.

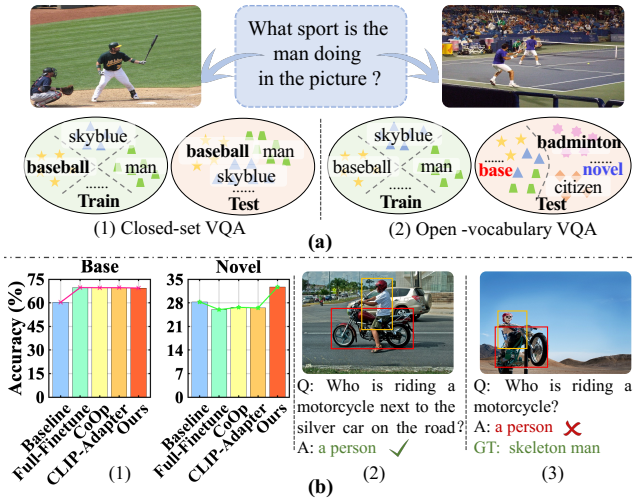

Figure 1: (a) Closed-set VQA *vs* Open-Vocabulary VQA; (b) Experimental results comparing baseline, existing fine-tuning models, and our method on base and novel classes (1); Biased predictions in OVVQA tasks (2-3).

Introduction

Visual Question Answering (VQA) is a typical multimodal task and has drawn increasing interest over the past few years (Khan and Fu 2024; Li et al. 2024a), which can automatically generate a textual answer given a question and an image. This field holds significant promise for improving our ability to integrate visual and textual information. The application of VQA in areas such as intelligent robotics and assistive technology can be particularly impactful, for example, in enhancing user interactions with automated systems.

While the VQA community has achieved substantial progress, the benchmarks employed thus far have predominantly adhered to the closed-set setting (Li et al. 2024b; Wang et al. 2024a). As depicted in Fig. 1(a), these benchmarks rely on a predefined set of candidate answers (e.g., baseball, skyblue), restricting models to select from this fixed list and impeding their ability to handle unseen concepts (e.g., badminton, citizen). This constraint diminishes model effectiveness in open-world scenarios, where answer distributions are highly varied and dynamic. Moreover, current closed-set approaches often treat each benchmark sample equivalently, assessing model performance based on overall average recognition rates. This methodology encourages models to prioritize common, high-frequency vocabulary, which may not reflect the challenges posed by real-world applications that usually involve rare or novel terms. Consequently, these benchmarks fail to fully assess a model's true performance and ability to generalize to practical applications. To overcome these limitations, we propose a new evaluation benchmark for VQA, termed Open-Vocabulary Visual Question Answering (**OVVQA**). OVVQA is designed to better align with open-world conditions and to provide a more accurate assessment of a model's multimodal reasoning abilities. Unlike traditional closed-set VQA benchmarks, as shown on the right of Fig. 1(a), OVVQA includes both base and novel classes.

*Corresponding author.

The base class, comprising common and high-frequency answers, evaluates fundamental model performance, while the novel class, containing unseen and low-frequency answers, tests model's capacity to generalize beyond the training set.

However, performing open-vocabulary VQA is non-trivial, as novel classes often exhibit different representation structures, typically being less compact, compared to those seen during training. Consequently, the prior knowledge about the test-time classes and distributions is absent. Recently, large-scale pretrained language models (Touvron et al. 2023) have yielded impressive results in natural language processing, leading many VQA studies (Li et al. 2023; Dong et al. 2024) to leverage these models through fine-tuning to benefit from the versatile knowledge acquired during pre-training. Building on this approach, we establish a series of baselines and adapt several well-known fine-tuning strategies (e.g., Full Finetune (Lin, Madotto, and Fung 2020), CoOp (Zhou et al. 2022), and CLIP-Adapter (Gao et al. 2024)) to train VQA models under our proposed OVVQA setting. However, as illustrated on the left of Fig. 1(b), our results reveal that fine-tuned VQA models show improved performance on base categories but deteriorate in novel categories. Specifically, models perform worse on novel classes after fine-tuning, indicating that these methods do not enhance generalization but merely memorize the data in the base class. We identify this phenomenon as a newly discovered bias, termed *canonical class overfitting* bias. This bias arises because existing VQA models typically employ empirical risk minimization as the optimization objective, which involves minimizing the loss between predicted and ground truth answers. As a result, the models tend to capture statistical regularities and exploit correlations within seen classes (base classes), hindering their ability to generalize to unseen classes (novel classes). As shown on the right of Fig. 1(b), predictions are accurate before fine-tuning but become erroneous afterward. This shift arises because concepts like 'person' and 'motorcycle' often co-occur in the base class. Consequently, when the model encounters the novel class 'skeleton man' during testing, it remains biased towards the base class, leading to incorrect predictions. To address biases in VQA tasks (e.g., language bias (Niu et al. 2021)), common solutions are to expand the dataset (Cho et al. 2023). However, this approach is ineffective for our *canonical class overfitting* bias in OVVQA, as novel categories must remain unseen during training.

To address the aforementioned challenges, we propose a novel Causal Adapter to efficiently transfer causal knowledge from the pretrained model to OVVQA, which is designed in a plug-and-play manner for ease of integration. Our framework enables the model to identify the causal relationship between inputs (e.g., visual content) and outputs (e.g., answers), thereby mitigating biases caused by distribution discrepancies between base and novel classes in OVVQA. We begin by formulating OVVQA as a causal graph and analyzing how existing methods establish spurious associations between visual content and answers. Spurious effects are then eliminated through a causal intervention based on the front-door adjustment principle (Pearl 1995), which is incorporated into the adapter and does not require the assumption of any observed confounder. This makes the proposed Causal Adapter applicable across various domains where the adapter resides. Additionally, we introduce an adaptive transfer loss to improve the fine-tuning process, effectively leveraging knowledge from pretext tasks to enhance performance across base and novel classes.

The major contributions of this work are as follows: (1) We propose a pioneering, model-agnostic causal adapter that leverages front-door adjustment to mitigate biases arising from distribution shifts between base and novel classes in OVVQA. This adapter serves as a plug-and-play module, enhancing the efficiency of model fine-tuning. (2) We establish OVVQA, the inaugural open-vocabulary VQA benchmark, designed to evaluate pretrained models on both base and novel categories. This benchmark assesses multimodal reasoning capabilities and enhances the applicability of models in open-world scenarios. (3) We demonstrate the effectiveness of our causal adapter by integrating it into various pretrained models. Extensive experiments show that our method consistently enhances performance across multiple datasets, underscoring its robustness and adaptability.

Related Work

Visual Question Answering (VQA)

Current VQA models fall into two categories: task-specific models (Yuan, You, and Bao 2023; Chen and Zhao 2023; Dancette et al. 2023; Ganz et al. 2024) and models pretrained on large-scale datasets, then fine-tuned for VQA tasks (Dai et al. 2023; Dong et al. 2023; Wu et al. 2024). While these methods have shown promise, they primarily focus on closed-set VQA. To improve generalization in real-world scenarios, attention has shifted to zero-shot VQA (Guo et al. 2023; Lan et al. 2023), where models aim to answer questions about images without prior exposure to specific question types. Modern zero-shot VQA methods, often based on large vision-language models like CLIP (Radford et al. 2021) and Flamingo (Alayrac et al. 2022), leverage large-scale pretraining on diverse image-text datasets, enabling them to address a wide range of questions about novel contents. However, due to the lack of explicit task-specific supervision, these models struggle with questions requiring logical deduction or complex object relationships, limiting their practicality (Pourpanah et al. 2023). To address these limitations, we introduce the OVVQA benchmark. The most closely related work is by Ko et al. (Ko et al. 2023), which focuses on open-set video QA. Our OVVQA benchmark differs in key aspects: our benchmark focuses on image-based question answering, whereas (Ko et al. 2023) targets video question answering. Additionally, (Ko et al. 2023) requires unseen categories to be known and embedded during testing. In contrast, OVVQA treats any category outside the base set as novel without prior knowledge.

Causal Inference

Causal inference (Pearl, Glymour, and Jewell 2016) has recently gained traction in multimodal tasks, such as visual dialogue (Zhang, Ji, and Liu 2023; Su et al. 2024), video

moment localization (Lv, Su, and Wen 2023), image captioning (Cao et al. 2024), and VQA (Zhang, Zhang, and Xu 2023; Vosoughi et al. 2024), aiming to eliminate dataset bias (Bareinboim and Pearl 2012) and enhance models' ability to uncover true causality (Liu, Li, and Lin 2023; Nie et al. 2023). For instance, (Zang et al. 2023) decouples causal and non-causal features in visual and textual modalities, optimizing multimodal causal relationships through interventions to improve robustness and reusability. However, existing methods (Zhang et al. 2020; Zang et al. 2023) often assume observable confounders and rely on backdoor adjustments, which may limit their applicability, as confounders are often unobservable and elusive. To address this, we employ front-door adjustment (Pearl 1995), which mitigates dataset bias without assuming observed confounders. For example, Yang et al. (Yang et al. 2021) use front-door adjustment to eliminate confounding effects in vision-language models, showing promising results across various tasks. However, their approach discards previously learned knowledge from pretrained models. In contrast, we propose a causal adapter for efficient fine-tuning and knowledge transfer, allowing our method to retain the benefits of pretrained models while effectively adapting to novel scenarios, thereby enhancing OVVQA performance.

OVVQA Benchmark

Problem Definition

Traditional VQA benchmarks follow a closed-set paradigm, with predefined training and testing answers, marking any out-of-set answer as incorrect during evaluation. This setting limits the model's generalization, as it can only predict familiar answers, thereby introducing bias and neglecting unseen answers. OVVQA, by contrast, requires models to predict both seen (base) and unseen (novel) answers, offering a more rigorous evaluation of generalization capability. To construct the OVVQA benchmark, we consolidate the train and test sets from existing VQA datasets, partitioning unseen test classes from seen classes. The seen classes constitute the base class for train and testing, while the unseen classes form the novel class for testing only.

We construct OVVQA using three standard datasets commonly employed in closed-set VQA: VQA v2 (Goyal et al. 2017) with 0.65 million image-question pairs, GQA (Hudson and Manning 2019) with 1.1 million pairs for visual reasoning and compositional question answering, and OKVQA (Marino et al. 2019) with 14,055 pairs requiring external knowledge for answer reasoning.

OVVQA

OVVQA presents two key requirements: (1) *Ensuring novel classes in the test set remain unseen during training:* this is crucial for accurately assessing the model's ability to generalize to new categories without prior exposure. (2) *Ensuring the benchmark validates models' comprehensive capability:* the benchmark must meticulously test models' adaptability by including a diverse array of unseen scenarios.

For the first requirement, we utilize existing closed-set VQA datasets (e.g., VQA v2, GQA, and OKVQA), merging their train and test sets and categorizing each sample into *seen* and *unseen* categories based on the answer. Categorizing by answer offers several advantages: First, the primary goal of VQA is to determine the correct answer, making it essential for models to handle previously unseen answers effectively. Second, answers are more concise and straightforward compared to images, which often contain redundant information, and questions, which may include irrelevant details. Additionally, we consider all candidate answers in each training sample for labeling and categorization, ensuring that unseen categories are not included within incorrect answers, thus preserving the integrity of the evaluation.

Data Split →		Classes		Samples	
Dataset ↓		Base	Novel	Base	Novel
OV-VQA$_{v2}$	Train	2743	-	596265	-
	Test	2743	386	52252	9594
OV-GQA	Train	1022	-	1062339	-
	Test	1022	821	12293	13008
OV-OKVQA	Train	14040	-	9009	-
	Test	14040	1000	4345	701

Table 1: Main statistics of our three OVVQA benchmarks.

For the second requirement, we employ a two-step strategy for data partition and effective evaluation. First, we merge and rank the train and test samples from each dataset based on answer frequency, dividing them into a head set and a tail set. The tail set (e.g., answers with frequencies below N) forms the *unseen* set, while the head set forms the *seen* set. We set N to 30 for VQA v2 and GQA, and 10 for OKVQA. The unseen set then becomes the *novel* classes in the test set. This division better reflects real-world scenarios, where data is typically sparse due to its diversity. In the second step, to preserve the original data distribution, samples from the *seen* set that originally belonged to the test set remain as test samples, forming the *base* classes in the test set. The remaining *seen* classes are used for training. This approach ensures that our OVVQA benchmark effectively validates the model's generalization and reusability by emphasizing its ability to handle rare and diverse scenarios while maintaining the inherent data distribution. Tab. 1 presents the number of classes and samples in the train and test sets across our three reconstructed OVVQA benchmarks: OV-VQA$_{v2}$, OV-GQA, and OV-OKVQA.

Evaluation Metrics

Two evaluation metrics are adopted to assess the model's generalization and transferability: arithmetic mean and harmonic mean. The former computes the average accuracy across all tasks, defined as $Avg = \frac{Acc_{base}+Acc_{novel}}{2}$, where Acc_{base} and Acc_{novel} represent the prediction accuracies on the base and novel classes, respectively. While straightforward, this metric can be skewed by higher accuracy on the base class, failing to capture the trade-offs between the two classes. To address this, we also use the harmonic mean, a metric commonly applied in open-set tasks such as classi-

fication and detection (Zhang et al. 2024; Yao et al. 2023). The harmonic mean provides a more balanced evaluation by accounting for the trade-off between base and novel classes and is calculated as $H = \frac{(Acc_{base} \times Acc_{novel}) \times 2}{Acc_{base} + Acc_{novel}}$.

Method

Causal View of OVVQA

Problem Formulation: Given an image $i \in I$ and a question $q \in Q$, the objective of OVVQA is to predict an answer $y \in \{Y_b, Y_n\}$, where Y_b and Y_n represent the base (seen) and novel (unseen) classes in the test set, respectively. During training, the model is exposed only to the base classes Y_b. To prevent data leakage, we use the widely recognized VL-T5 and VL-BART models (Cho et al. 2021) as baselines, ensuring they have not been exposed to the datasets used in this study. As depicted on the left side of Fig. 2, both models employ an encoder-decoder architecture, and any similar architecture could serve as our baseline. Our causal adapter, a plug-and-play module, can be seamlessly integrated between the encoder and decoder, enabling efficient model fine-tuning. This integration reduces the impact of OVVQA bias, enhancing the model's robustness and generalization.

Causal Graph Construction: We formalize cross-modal causality for the OVVQA task using a structural causal model (Pearl, Glymour, and Jewell 2016), which defines the high-level causal dependencies (edges) among data variables (nodes). In Fig. 2(a) (depicted on the right), we represent key factors in OVVQA (image I, question Q, joint representation X, and label Y=$\{Y_b, Y_n\}$) as components in a causal graph, with connecting links representing their relationships. Although simple and intuitive, this process is prone to canonical class overfitting bias, caused by statistical regularities in the seen classes. To address this, we propose a new causal graph (Fig. 2(b)), modeling the bias as a confounder C that influences both visual content I and prediction Y. The causality in the new graph is as follows:

$I \rightarrow X \leftarrow Q$ denotes feature extraction and alignment of multimodal data (e.g., image and question).

$X \rightarrow Y_b/Y_n$ represents the direct causal effect from the fused feature to the answer, crucial for both base and novel (only emerges in the test stage) class predictions.

$I \leftarrow C \rightarrow Y_b \hookrightarrow Y_n$ indicates that bias, acting as a confounder C, creates a spurious link from image I to label Y_b, which in turn affects the prediction of Y_n. This bias is due to statistical regularities in the seen classes. For example, in Fig. 1, the frequent co-occurrence of 'person' and 'motorcycle' in the base class hinders the model's ability to correctly predict novel classes like 'skeleton man' during testing.

Front-Door Causal Intervention

To address spurious correlations, many approaches employ backdoor adjustment (Zhao et al. 2022; Wang et al. 2024b). However, in practice, data bias is complex and variable, making it challenging to identify and disentangle confounders. The canonical class overfitting bias, introduced in this paper, exemplifies these challenges, as quantifying the impact of learning on seen classes on the model's generalization to unseen classes is difficult. The front-door adjustment offers a feasible approach to mitigate bias when explicit representation of the confounder is not possible. As shown in Fig. 2(c), it introduces an additional mediator, M, between X and Y, creating a path $I \rightarrow X \rightarrow M \rightarrow Y$ that facilitates knowledge transmission. Consequently, the answer prediction process then transforms as follows:

$$P(Y|I,Q) = \sum_m P(M=m|I,Q)P(Y|M=m), \quad (1)$$

where m denotes the selected knowledge from I and Q. To mitigate the bias, Eq.(1) is then transformed into an interventional probabilistic form:

$$P(Y|do(I),Q) = \sum_m P(M=m|do(I),Q)P(Y|do(M=m)), \quad (2)$$

where $do(\cdot)$ represents a causal intervention (Pearl 2009), which means that the variable is given a specific value so that it is no longer affected by its parent nodes. Eq.(2) delineates the answer prediction process into two phases: $\{I,Q\} \rightarrow X \rightarrow M$ and $M \rightarrow Y$. This sequential decomposition structures the influence paths in the causal framework. Next, we detail the learning process for each component and describe our causal adapter, designed to enhance the model's generalizability by capturing authentic causal relationships.

As illustrated in Fig. 2(c), the backdoor path $I \leftarrow C \rightarrow Y \leftarrow M$ between I and M is severed by the collider junction $C \rightarrow Y \leftarrow M$. Consequently, the probability $P(M=m|do(I),Q)$ can be computed as:

$$P(M=m|do(I),Q) = P(M=m|I,Q). \quad (3)$$

This intervention ensures that the estimation of M is not confounded by I, leading to more reliable and causally valid inferences. Furthermore, to eliminate spurious correlations caused by hidden confounders, and estimate the true causal effect, we must block the back-door path between M and Y. This can be achieved by stratifying the input variable I into different cases $\{i\}$ and then measuring the average causal effects of M on Y as follows:

$$P(Y|do(M=m)) = \textstyle\sum_i P(I=i)P(Y|I=i,M=m). \quad (4)$$

In summary, the true causal effect between I, Q and Y can be computed by applying Eq.(3) and Eq.(4) into Eq.(2):

$$P(Y|do(I),Q) = \textstyle\sum_m P(M=m|I,Q)\sum_i P(I=i)P(Y|I=i,M=m). \quad (5)$$

Causal Adapter

In this section, we detail the implementation of causal interventions within our causal adapter to enhance model capabilities. This approach minimizes reliance on spurious correlations and improves robustness across seen and unseen classes, creating a more reliable and generalizable framework for VQA. Specifically, we first parameterize $P(Y|I,M)$ using an adapter $Adapter(\cdot)$, with L_A standard transformer layers followed by the decoder $Decoder(\cdot)$ described in Section 4.1. This can be formulated as:

$$P(Y|I,M) = Decoder_*(Adapter(I,M)). \quad (6)$$

To facilitate rapid adaptation of large pretrained models to the VQA task, we fix the parameters in the

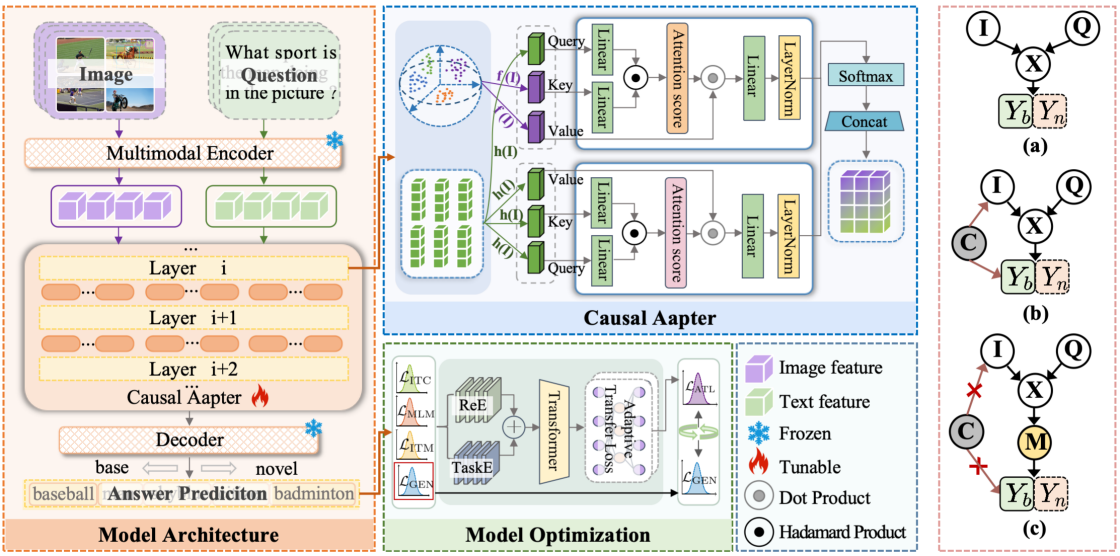

Figure 2: **Left**: Our framework comprises two key components: a causal adapter and an adaptive transfer loss. The causal adapter minimizes the model's reliance on spurious correlations by uncovering causal relationships between different modalities. The adaptive transfer loss leverages knowledge from pretext tasks to balance performance between base and novel classes. **Right**: Causal graphs for existing VQA (a), with confounder C (b), and for our method (c).

$Decoder_*(\cdot)$ and make the $Adapter(\cdot)$ trainable. This allows the adapter to learn task-specific adjustments without altering the robust features learned during pre-training, optimizing performance and efficiency. Furthermore, to compute $P(Y|do(I), Q)$ in Eq.(5), we need to sample I and M. However, forward propagation for all samples is computationally expensive. To address this, we apply the Normalized Weighted Geometric Mean (Srivastava et al. 2014) optimization to sample outer samples, as shown in Eq.(7):

$$\begin{aligned} &P(Y|do(I), Q) \approx Decoder_*(CausalA(\hat{I}, \hat{M})) \\ &= Decoder_*\Big[CausalA\Big(\sum_i P(I=i|f(I))i, \\ &\qquad \sum_m P(M=m|h(I,Q))m\Big)\Big], \end{aligned} \tag{7}$$

where $\hat{M}$ represents in-sample sampling, with m denoting selected knowledge from I and Q, and $\hat{I}$ represents cross-sample sampling, as it is derived from other samples. Both $\hat{M}$ and $\hat{I}$ can be computed using attention networks. Consequently, we introduce a novel causal adapter, $CausalA(\cdot)$, with an updated attention module that jointly estimates $\hat{M}$ and $\hat{I}$. This enhanced module improves the representation of causality-aware multimodal features, enabling the adapter to better capture and utilize causal relationships, thereby enhancing model performance in complex VQA tasks. Consider the estimation of $\hat{I}$ as an example. We apply a Q-K-V operation with a cross-sample sampling strategy:

$$\begin{aligned} &Query = h(I), Key = f(I), Value = f(I), \\ &\hat{I} = \alpha \cdot Value, \quad \alpha = Softmax(\frac{Key^\top Query}{\sqrt{D_k}}), \end{aligned} \tag{8}$$

where $f(I)$ is the visual feature selector using k-means to identify distinctive features from other training samples, and $h(I)$ is the feature extractor that enhances representation for querying. The estimation of $\hat{M}$ follows a similar process, with $Query = Key = Value = h(I, Q)$. Finally, $\hat{I}$ and $\hat{M}$ are concatenated to estimate $P(Y|do(I), Q)$.

Objective Function

To implement the model for OVVQA, we train the causal adapter by minimizing the negative log-likelihood of the ground truth answer tokens as formalized below:

$$\mathcal{L}_{GEN} = -\sum_{j=1}^{|Y_b|} logP(Y_j|Y_{<j}, \hat{I}, \hat{M}). \tag{9}$$

To enhance the model's generalization, we extend beyond traditional fine-tuning, which typically focuses on a single downstream task. Instead, we propose an adaptive transfer loss that nonlinearly combines multiple auxiliary losses to improve knowledge transfer from pretext tasks to OVVQA, thereby enhancing generalization to novel answers. Specifically, utilizing the originally provided image captions in the dataset, we include image-text matching loss $\mathcal{L}_{ITM}$, image-text contrastive learning loss $\mathcal{L}_{ITC}$, and masked language model loss $\mathcal{L}_{MLM}$, which have been used in pre-training the baseline models (VL-T5, VLBART) (Cho et al. 2021). We then combine the losses with $\mathcal{L}_{GEN}$ into a unified loss using an automated and nonlinear Transformer-based approach:

$$\mathcal{L}_{ATL} = Transformer(ReE(\mathcal{L}), TaskE(t)), \tag{10}$$

where $ReE(\mathcal{L}) = MLP(\mathcal{L}_{GEN}, \mathcal{L}_{ITM}, \mathcal{L}_{ITC}, \mathcal{L}_{MLM})$ projects each loss value to an embedding, and $TaskE(t) = Embedding(t)$ embeds each task, functioning as positional encoding. The model is then fine-tuned iteratively using $\mathcal{L}_{GEN}$ and $\mathcal{L}_{ATL}$. These carefully designed losses focus on

Data Split →	OV-VQA$_{v2}$							OV-GQA			
Method ↓	Base	Novel				Avg	H	Base	Novel	Avg	H
		Y/N	Num.	Other	Overall						
LXMERT	69.68	82.59	4.15	19.63	19.09	44.39	29.97	60.97	18.33	39.65	28.19
ALBEF	69.70	76.43	5.55	24.12	23.20	46.45	34.81	56.73	18.07	37.40	27.41
RepARe	71.31	71.43	5.34	29.67	28.15	49.73	40.37	56.86	21.23	39.05	30.92
Baseline$_{VL-T5}$	60.37	78.39	5.48	28.93	28.29	44.33	38.53	40.52	18.49	29.51	25.39
Full-Finetune	69.70	75.54	4.97	27.34	25.99	47.85	37.86	59.35	12.46	35.91	20.60
+CoOp	69.61	79.55	5.22	27.94	26.67	48.14	38.56	60.17	13.76	36.97	22.40
+CLIP-Adapter	69.63	80.45	5.21	27.73	26.49	48.06	38.38	60.70	18.13	39.42	27.92
+Ours	69.40	80.18	5.68	34.54	**32.69**	**51.05**	**44.44**	60.20	23.34	**41.77**	**33.64**
Baseline$_{VL-BART}$	59.38	79.17	5.13	26.85	28.39	43.12	36.98	41.37	18.31	29.84	25.38
Full-Finetune	68.18	78.39	6.93	25.80	24.86	46.52	36.43	59.59	14.39	36.99	23.18
+CoOp	68.22	77.77	6.38	26.55	25.48	46.85	37.10	60.38	15.84	38.11	25.10
+CLIP-Adapter	68.93	79.20	6.12	27.01	25.89	47.41	37.64	59.65	18.17	38.91	27.86
+Ours	68.79	80.80	5.59	32.94	**31.24**	**50.02**	**42.97**	59.75	25.11	**42.43**	**35.36**

Table 2: Comparisons with state-of-the-art methods on OV-VQA$_{v2}$ and OV-GQA. Avg: arithmetic mean; H: harmonic mean.

domain-specific features and contexts, ensuring that they do not introduce noise or cause data leakage in novel categories. During inference, the trained model takes the given image and question as input, and the decoder outputs the answer, which may belong to a base or unseen class.

Experiments

Experimental Setup

Implementation Details. Our causal adapter is a plug-and-play module, allowing seamless integration into existing encoder-decoder models without altering their architectures. In our experiments, the number of layers L_A is set to 3. For each image-question pair, visual and textual embeddings are extracted using Faster R-CNN (Girshick 2015) and WordPiece tokenization, respectively. These embeddings are concatenated and passed through an encoder to generate a contextualized joint representation. The decoder then predicts the probability distribution over future answer tokens by attending to previously generated tokens and the encoder outputs. To adapt pretrained models to the OVVQA task, we freeze the encoder and decoder modules and integrate our causal adapter between them in VL-T5 and VL-BART. The parameters from our causal adapter and Eq.(10) are optimized using Adam with a learning rate of $5e$-5. Batch sizes are set to 80 for VL-T5 and 128 for VL-BART.

Experimental Results

Compared Methods. To demonstrate the effectiveness and generalizability of our causal adapter, we integrate it into two renowned pretrained models (Cho et al. 2021), VL-T5 and VL-BART, and evaluate its performance against three fine-tuning methods: a full finetune approach (Lin, Madotto, and Fung 2020) and two parameter-efficient methods, CoOP (Zhou et al. 2022) and CLIP-Adapter (Gao et al. 2024). Additionally, we further adapt another three prominent pretrained models (LXMERT (Tan and Bansal 2019), ALBEF (Li et al. 2021), and RepARe (Prasad, Stengel-Eskin, and Bansal 2023)) to our OVVQA task to assess the broader impact of our approach.

Results Analysis. Tab. 2 and Tab. 3 present the experimental results on our reconstituted OVVQA datasets: OV-VQA$_{v2}$, OV-GQA, and OV-OKVQA. We report performance across several aspects, including results for base and novel classes, arithmetic mean (Avg), and harmonic mean (H). For OV-VQA$_{v2}$, we further subdivide the novel class into three subcategories: Yes/No (Y/N), Number (Num.), and Other. From the results, we can draw the following conclusions: (1) Comparing baseline methods with fine-tuning strategies (Full-Finetune, CoOp, and CLIP-Adapter) reveals that while these methods enhance base class performance, they significantly degrade performance on the novel class, highlighting the challenge of extending fine-tuned models to unseen categories. (2) Our causal adapter not only matches the performance of fine-tuning methods on the base class but also significantly improves performance on the novel class. For example, using VL-T5 as the baseline, our approach boosts novel class performance from 6.02% to 6.7% on OV-VQA$_{v2}$, 5.21% to 10.88% on OV-GQA, and 1.65% to 2.48% on OV-OKVQA. With VL-BART as the baseline, it achieves enhancements from 5.35% to 6.38%, 6.94% to 10.72%, and 1.39% to 2.36% on the respective datasets. These results demonstrate the robustness and adaptability of our approach in diverse, unseen VQA scenarios. (3) Compared to other VQA models like LXMERT, ALBEF, and RepARe, our approach consistently outperforms them across all datasets in both average and harmonic mean metrics. Although RepARe, which is also enhanced by image captions, performs well, our method shows improvements of 4.07%, 2.72%, and 12.79% for H, and 1.32%, 2.72%, and 12.5% for Avg on OV-VQA$_{v2}$, OV-GQA, and OV-OKVQA, respectively, using VL-T5 as the baseline. With VL-BART as the

Data Split →	OV-OKVQA			
Method ↓	Base	Novel	Avg	H
LXMERT	31.74	15.90	23.82	21.19
ALBEF	29.47	15.13	22.30	19.99
RepARe	33.93	17.78	25.86	23.33
Baseline$_{VL-T5}$	38.17	27.91	33.04	32.24
Full-Finetune	47.36	27.45	37.41	34.76
+CoOp	47.45	26.62	37.04	34.11
+CLIP-Adapter	47.41	27.10	37.26	34.49
+Ours	47.61	29.10	**38.36**	**36.12**
Baseline$_{VL-BART}$	38.23	27.87	33.05	32.24
Full-Finetune	46.49	27.08	36.79	34.22
+CoOp	46.58	26.31	36.45	33.63
+CLIP-Adapter	46.77	26.11	36.44	33.51
+Ours	46.90	28.47	**37.69**	**35.43**

Table 3: Comparisons with state-of-the-art methods on OV-OKVQA. Avg: arithmetic mean; H: harmonic mean.

baseline, our approach continues to achieve notable gains across all datasets under both metrics. These results highlight the effectiveness of our causal fine-tuning strategy in enhancing the model's ability to discern true causal relationships, thereby improving generalization to novel classes.

Ablation Studies

Effect of Each Component. To assess the impact of each component in our approach, we construct several ablated versions, with results shown in Tab. 4. The results are reported on OV-VQA$_{v2}$. The variants $w/ATT\text{-}Adapter$ and $w/Causal\text{-}Adapter$ both integrate a three-layer Transformer as an adapter between the encoder and decoder in VL-T5 and VL-BART. The $w/ATT\text{-}Adapter$ employs a standard Transformer, while $w/Causal\text{-}Adapter$ incorporates causal mechanisms into the Transformer's attention, as detailed in Section 4.3. Comparing the baseline with these variants shows that while both approaches improve performance on seen classes, a poorly designed adapter can degrade performance on unseen classes. As shown in Tab. 4, compared to the baselines, $w/ATT\text{-}Adapter$ results in a decline in novel class performance by 2.54% on VL-T5 and 1.67% on VL-BART. In contrast, our proposed method not only improves base class performance but also increases novel class performance by 1.86% and 1.63%, respectively. This demonstrates that our causal adapter effectively captures true causal relationships across classes, enhancing performance on unseen categories. Additionally, the $w/\mathcal{L}_{ATL}$ variant incorporates the adaptive transfer loss from Eq. (10) into the baseline. Compared to the baseline, $w/\mathcal{L}_{ATL}$ shows performance gains on both base and novel classes, indicating that the auxiliary loss helps the model retain more generalizable knowledge from pretext tasks. Our full model, as presented in the last row, surpasses all variants, underscoring the effectiveness of our multi-modal reasoning approach.

Impact of Hyperparameter. We investigate the effect of the size of distinctive features K (i.e., the number of k-means cluster centers) used in our causal adapter. We train models with $K \in \{0, 50, 100, 150, 200, 300\}$, where K= 0 denotes random initialization instead of k-means clustering. The results are shown in Fig. 3. As demonstrated, models initialized without k-means perform worse, highlighting the importance of learning effective distinctive features for the causal adapter's success. Additionally, K= 100 consistently yields the best performance across both VL-T5 and VL-BART, so we set K= 100 in our experiments.

Method ↓	Base	Novel	Avg	H
Baseline$_{VL-T5}$	60.37	28.29	44.33	38.53
$w/$ ATT-Adapter	69.44	25.75	47.60	37.57
$w/$ Causal-Adapter	69.12	30.15	49.64	41.99
$w/\mathcal{L}_{ATL}$	67.38	28.97	48.18	40.52
Ours	69.40	32.69	**51.05**	**44.44**
Baseline$_{VL-BART}$	59.38	26.85	43.12	36.98
$w/$ ATT-Adapter	67.96	25.18	46.57	36.75
$w/$ Causal-Adapter	68.53	28.48	48.51	40.24
$w/\mathcal{L}_{ATL}$	66.19	27.01	46.60	38.36
Ours	68.79	31.24	**50.02**	**42.97**

Table 4: Ablation Studies on OV-VQA$_{v2}$ dataset using VL-T5 and VL-BART as baseline models.

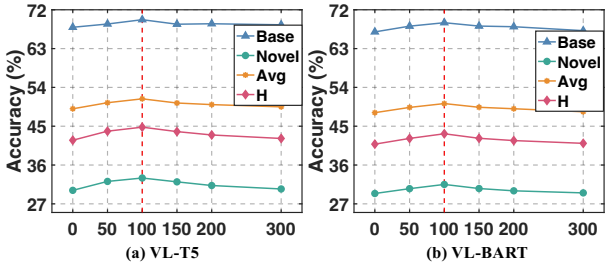

Figure 3: Performance variation on OV-VQA$_{v2}$ with respect to different sizes of distinctive features.

Conclusion

In this paper, we introduced the OVVQA to address the limitations of existing VQA benchmarks, specifically designed for open-world conditions. Unlike traditional benchmarks, OVVQA includes both base and novel classes, providing a more rigorous assessment of a model's generalization capabilities. We also proposed a model-agnostic Causal Adapter that leverages front-door adjustment to mitigate biases in current VQA tasks. Our approach, combined with an adaptive transfer loss, significantly improves accuracy on novel classes while maintaining strong performance on base classes. Extensive experiments show that our method consistently outperforms state-of-the-art approaches, demonstrating robust generalization and adaptability in real-world scenarios. Future work will focus on refining the causal adapter and applying it to other multimodal tasks.

Acknowledgments

This work was supported in part by the National Key Research and Development Plan of China under Grant 2021ZD0112200; in part by the National Natural Science Foundation of China under Grant 62376196, Grant 62036012, Grant U23A20387, Grant 62106262, Grant 62202331, Grant 62206200, and Grant 62276118; and in part by Tianjin Natural Science Foundation under Grant 24JCJQJC00190 and Grant 22JCYBJC00030.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: A visual language model for few-shot learning. In *NIPS*, 23716–23736.

Bareinboim, E.; and Pearl, J. 2012. Controlling selection bias in causal inference. In *AAAI*, 100–108.

Cao, Q.; Chen, X.; Song, R.; Wang, X.; Huang, X.; and Ren, Y. 2024. See or Guess: Counterfactually regularized image captioning. In *ACM MM*.

Chen, S.; and Zhao, Q. 2023. Divide and conquer: Answering questions with object factorization and compositional reasoning. In *CVPR*, 6736–6745.

Cho, J.; Lei, J.; Tan, H.; and Bansal, M. 2021. Unifying vision-and-language tasks via text generation. In *ICML*, 1931–1942.

Cho, J. W.; Kim, D.-J.; Ryu, H.; and Kweon, I. S. 2023. Generative bias for robust visual question answering. In *CVPR*, 11681–11690.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards general-purpose vision-Language models with instruction tuning. In *NIPS*, 49250–49267.

Dancette, C.; Whitehead, S.; Maheshwary, R.; Vedantam, R.; Scherer, S.; Chen, X.; Cord, M.; and Rohrbach, M. 2023. Improving selective visual question answering by learning from your peers. In *CVPR*, 24049–24059.

Dong, J.; Zhang, Q.; Zhou, H.; Zha, D.; Zheng, P.; and Huang, X. 2024. Modality-aware integration with large language models for knowledge-based visual question answering. In *ACL*, 2417–2429.

Dong, R.; Han, C.; Peng, Y.; Qi, Z.; Ge, Z.; Yang, J.; Zhao, L.; Sun, J.; Zhou, H.; Wei, H.; et al. 2023. Dreamllm: Synergistic multimodal comprehension and creation. In *ICLR*, 871–887.

Ganz, R.; Kittenplon, Y.; Aberdam, A.; Ben Avraham, E.; Nuriel, O.; Mazor, S.; and Litman, R. 2024. Question aware vision transformer for multimodal reasoning. In *CVPR*, 13861–13871.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 132(2): 581–595.

Girshick, R. 2015. Fast r-cnn. In *CVPR*, 1440–1448.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 6904–6913.

Guo, J.; Li, J.; Li, D.; Tiong, A. M. H.; Li, B.; Tao, D.; and Hoi, S. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *CVPR*, 10867–10877.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 6700–6709.

Khan, Z.; and Fu, Y. 2024. Consistency and Uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *CVPR*, 10854–10863.

Ko, D.; Lee, J. S.; Choi, M.; Chu, J.; Park, J.; and Kim, H. J. 2023. Open-Vocabulary Video Question Answering: A new benchmark for evaluating the generalizability of video question answering models. In *CVPR*, 3101–3112.

Lan, Y.; Li, X.; Liu, X.; Li, Y.; Qin, W.; and Qian, W. 2023. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *ACM MM*, 4389–4400.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 19730–19742.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NIPS*, 9694–9705.

Li, L.; Peng, J.; Chen, H.; Gao, C.; and Yang, X. 2024a. How to configure good in-context sequence for visual question answering. In *CVPR*, 26710–26720.

Li, P.; Si, Q.; Fu, P.; Lin, Z.; and Wang, Y. 2024b. Object attribute matters in visual question answering. In *AAAI*, 18545–18553.

Lin, Z.; Madotto, A.; and Fung, P. 2020. Exploring versatile generative language model via parameter-efficient transfer learning. In *ACL*, 441–459.

Liu, Y.; Li, G.; and Lin, L. 2023. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(10): 11624–11641.

Lv, Z.; Su, B.; and Wen, J.-R. 2023. Counterfactual cross-modality reasoning for weakly supervised video moment localization. In *ACM MM*, 6539–6547.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 3195–3204.

Nie, W.; Wen, X.; Liu, J.; Chen, J.; Wu, J.; Jin, G.; Lu, J.; and Liu, A.-A. 2023. Knowledge-enhanced causal reinforcement learning model for interactive recommendation. *IEEE Transactions on Multimedia (TMM)*, 26: 1129–1142.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *CVPR*, 12700–12710.

Pearl, J. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. Causal inference in statistics: A primer. 2016. *Internet resource*.

Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C. P.; Wang, X.-Z.; and Wu, Q. M. J. 2023. A review of generalized zero-shot learning methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(4): 4051–4070.

Prasad, A.; Stengel-Eskin, E.; and Bansal, M. 2023. Rephrase, augment, reason: Visual grounding of questions for vision-language models. In *ICLR*, 96–111.

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.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research (JMLR)*, 15(1): 1929–1958.

Su, Y.; Wei, Y.; Nie, W.; Zhao, S.; and Liu, A. 2024. Dynamic Causal Disentanglement Model for Dialogue Emotion Detection. *IEEE Transactions on Affective Computing (ITAC)*.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 5100–5111.

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

Vosoughi, A.; Deng, S.; Zhang, S.; Tian, Y.; Xu, C.; and Luo, J. 2024. Cross modality bias in visual question answering: A causal view with possible worlds VQA. *IEEE Transactions on Multimedia (TMM)*.

Wang, J.; Zheng, Z.; Chen, Z.; Ma, A.; and Zhong, Y. 2024a. Earthvqa: Towards queryable earth via relational reasoning-based remote sensing visual question answering. In *AAAI*, 5481–5489.

Wang, L.; He, Z.; Dang, R.; Shen, M.; Liu, C.; and Chen, Q. 2024b. Vision-and-Language navigation via causal learning. In *CVPR*, 13139–13150.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T.-S. 2024. NExT-GPT: any-to-any multimodal LLM. In *ICML*.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021. Causal attention for vision-language tasks. In *CVPR*, 9847–9857.

Yao, L.; Han, J.; Liang, X.; Xu, D.; Zhang, W.; Li, Z.; and Xu, H. 2023. Detclipv2: Scalable open-vocabulary object detection pre-training via word-region alignment. In *CVPR*, 23497–23506.

Yuan, B.; You, S.; and Bao, B.-K. 2023. Self-PT: Adaptive Self-Prompt Tuning for Low-Resource Visual Question Answering. In *ACM MM*, 5089–5098.

Zang, C.; Wang, H.; Pei, M.; and Liang, W. 2023. Discovering the real association: Multimodal causal reasoning in video question answering. In *CVPR*, 19027–19036.

Zhang, D.; Zhang, H.; Tang, J.; Hua, X.-S.; and Sun, Q. 2020. Causal intervention for weakly-supervised semantic segmentation. In *NIPS*, 655–666.

Zhang, X.; Zhang, F.; and Xu, C. 2023. Reducing vision-answer biases for multiple-choice VQA. *IEEE Transactions on Image Processing (TIP)*.

Zhang, Y.; Zhang, C.; Yu, K.; Tang, Y.; and He, Z. 2024. Concept-guided prompt learning for generalization in vision-language models. In *AAAI*, 7377–7386.

Zhang, Z.; Ji, Y.; and Liu, C. 2023. Knowledge-aware causal inference network for visual dialog. In *ACM MM*, 253–261.

Zhao, H.; Ma, C.; Dong, X.; Luu, A. T.; Deng, Z.-H.; and Zhang, H. 2022. Certified robustness against natural language attacks by causal intervention. In *ICML*, 26958–26970.

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