Multi-Sourced Compositional Generalization in Visual Question Answering

Chuanhao Li[2,1*] , **Wenbo Ye**[2,1*] , **Zhen Li**[1] , **Yuwei Wu**[1,2†] , **Yunde Jia**[2,1]

[1]Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science & Technology, Beijing Institute of Technology, China
[2]Guangdong Laboratory of Machine Perception and Intelligent Computing,
Shenzhen MSU-BIT University, China
{lichuanhao, yewenbo, li.zhen, wuyuwei, jiayunde}@bit.edu.cn

Abstract

Compositional generalization is the ability of generalizing novel compositions from seen primitives, and has received much attention in vision-and-language (V&L) recently. Due to the multi-modal nature of V&L tasks, the primitives composing compositions source from different modalities, resulting in multi-sourced novel compositions. However, the generalization ability over multi-sourced novel compositions, *i.e.*, multi-sourced compositional generalization (MSCG) remains unexplored. In this paper, we explore MSCG in the context of visual question answering (VQA), and propose a retrieval-augmented training framework to enhance the MSCG ability of VQA models by learning unified representations for primitives from different modalities. Specifically, semantically equivalent primitives are retrieved for each primitive in the training samples, and the retrieved features are aggregated with the original primitive to refine the model. This process helps the model learn consistent representations for the same semantic primitives across different modalities. To evaluate the MSCG ability of VQA models, we construct a new GQA-MSCG dataset based on the GQA dataset, in which samples include three types of novel compositions composed of primitives from different modalities. The GQA-MSCG dataset is available at https://github.com/NeverMoreLCH/MSCG.

1 Introduction

Compositional generalization refers to the ability of generalizing novel compositions from seen primitives. In vision-and-language (V&L), the primitives of a composition come from either the linguistic modality or the visual modality, *i.e.*, compositions are multi-sourced. As shown in Figure 1, in the context of visual question answering (VQA), "white (linguistic modality) + dog (linguistic modality)" is an novel composition, and "white (linguistic modality) + 🐕 (visual modality)" and "🐕 (white + dog in visual modality)" are also novel compositions. However, prior works [Dankers *et al.*, 2022;

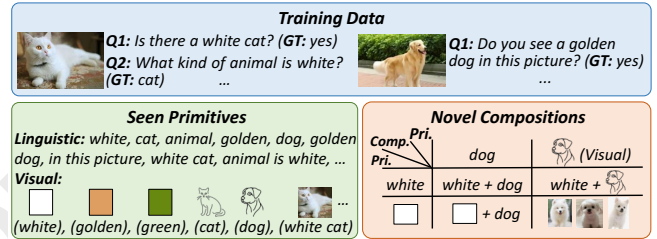† Corresponding author: Yuwei Wu.

Figure 1: Multi-sourced novel compositions in the context of VQA.

Li *et al.*, 2024b; Li *et al.*, 2023b] only consider novel compositions of primitives from a single modality (*e.g.*, "white + dog"). Whether the model's generalization ability to different modality primitives' novel compositions (*i.e.*, multi-sourced novel compositions) remains unexplored. To generalize over multi-sourced compositions, a model needs to not only have the ability to understand individual primitives but also the ability to align primitives of different modalities.

To address the above issue, we propose a retrieval-augmented training framework. The basic idea of the framework is to learn similar representations for primitives from different modalities by retrieving semantically equivalent primitives, thereby maintaining consistent generalization ability for multi-sourced compositions. Specifically, the framework consists of three key components: retrieval database construction, feature retrieval, and feature aggregation. For the retrieval database, we construct separate primitive databases for the linguistic and visual modalities, where words with the same prototype are treated as the same linguistic primitive, and visual entities with the same label are treated as the same visual primitive. For the same primitive, both the linguistic/visual primitive databases contain multiple instances of that primitive in different contexts. For example, for the linguistic primitive "**dog**", the linguistic primitive database contains instances of "Is the **dog** black?", "How many **dogs** are there?", "Do you see a golden **dog** in this picture?" in different contexts. During training, for each primitive in the training sample, semantically similar primitives from both the linguistic and visual primitive databases are retrieved. The retrieved features are then aggregated with the original primitive features to optimize the model. Since the representations in the primitive databases are up-

dated throughout the training, the model continuously refines its current representation by leveraging the retrieved relevant features to learn similar representations for the same semantic primitives across different modalities and contexts.

To quantitatively evaluate the generalization ability of VQA models for multi-sourced novel compositions, we construct the GQA-MSCG dataset based on the GQA dataset [Hudson and Manning, 2019]. We categorize compositions based on the modality of the primitives from the composition, resulting in three types of novel compositions: [Linguistic primitive, Linguistic primitive] (LL), [Visual primitive, Visual primitive] (VV), and [Linguistic primitive, Visual primitive] (LV). We construct three basic test splits, labeled LL, VV, and LV, each containing test samples of different types of novel compositions. To further explore how the co-occurrence of different types of novel compositions in samples affects model performance, we construct test splits, where each sample simultaneously contain two types of novel compositions: $LL + VV$, $LL + LV$, and $VV + LV$, as well as a test split, where each sample contain all three types of novel compositions: $LL + VV + LV$. Experimental results demonstrate that the proposed framework significantly improves VQA models' generalization ability to multi-sourced novel compositions while maintaining their independent and identically distributed (IID) generalization ability.

To sum up, our contributions are as follows:

- We are the first to explore the multi-sourced compositional generalization in V&L, which is critical for cross-modal understanding.

- We propose a retrieval-augmented training framework that improves the multi-sourced compositional generalization ability of VQA models by learning similar representations for primitives from different modalities.

- We present a GQA-MSCG dataset to evaluate the multi-sourced compositional generalization ability of VQA models with different types of novel compositions.

2 Related Work

Compositional generalization has garnered significant attention in various research fields. In natural language processing (NLP), works [Li *et al.*, 2021; Dankers *et al.*, 2022] focus on improving the compositional generalization ability of models for tasks like machine translation. Additionally, Chai et al. [2024] used text generation models for data augmentation during training to enhance compositional generalization in multi-label text classification tasks. In computer vision (CV), works [Naeem *et al.*, 2021; Jing *et al.*, 2024; Li *et al.*, 2024b] focus on improving compositional generalization for tasks like compositional zero-shot learning, specifically for novel compositions of [visual primitive, visual primitive]. In V&L, Pantazopoulos et al. [2022] constructed the first dataset to evaluate the compositional generalization ability of image captioning models. Works [Bahdanau *et al.*, 2018; Akula *et al.*, 2021; Yamada *et al.*, 2024; Li *et al.*, 2023a] focus on enhancing the compositional generalization ability of VQA models. Moreover, Li et al. [2022] constructed the Charades-CG and ActivityNet-CG datasets

for temporal video grounding (TVG) and used variational cross-graph reasoning to improve the compositional generalization ability of TVG models. These works primarily focus on improving the generalization ability of models for novel compositions involving the same modality of primitives, *e.g.*, works in NLP and V&L focus on [linguistic primitive, linguistic primitive] compositions, and works in CV focus on [visual primitive, visual primitive] compositions. In contrast, we explore the compositional generalization ability of V&L models for novel compositions involving different modality primitives, construct a new benchmark GQA-MSCG, and propose a retrieval-augmented training framework to improve the model's generalization ability for different types of novel compositions.

3 Framework

3.1 Overview

The overall framework of the proposed framework in the context of VQA is shown in Figure 2. For the training set $\mathcal{D}_t$, the first step is to construct linguistic and visual primitive databases, $\mathcal{D}_q$ and $\mathcal{D}_v$, respectively, for retrieval purposes. primitives with the same category (e.g., dog, cat) or attribute (e.g., blue, big) are treated as the same type of primitive. Both $\mathcal{D}_q$ and $\mathcal{D}_v$ contain multiple instances of each primitive in different contexts from $\mathcal{D}_t$. During training, for each training sample (Q, V), where Q represents the question and V represents the image, the retrieval module is used to retrieve similar primitives from $\mathcal{D}_q$ and $\mathcal{D}_v$ for the primitives in Q and V at feature level, respectively. The retrieved features are then aggregated with the original primitive features and used to replace the original features for training. Through this training process, the VQA model continuously refines its feature extractor by leveraging the retrieved relevant features, thereby learning similar features for the same semantic primitives across different modalities and contexts, thus improving the generalization ability of the model for multi-sourced novel compositions.

3.2 Retrieval Database Construction

To ensure fairness in the comparison, the retrieval database is constructed based on the training set $\mathcal{D}_t$ to avoid introducing external data during training.

Linguistic Primitive Database. For the linguistic primitive database, we first extract all words from the questions in $\mathcal{D}_t$ using the NLTK toolkit [Bird *et al.*, 2009]. All words are lemmatized, and words with part-of-speech tags as nouns, verbs, adjectives, and adverbs are considered as linguistic primitives. The set of all unique linguistic primitives is denoted as $\mathcal{S}_q$. For each linguistic primitive, T_q questions containing the primitive are sampled from $\mathcal{D}_t$. Although the meaning of the primitive remains the same across these T_q questions, the different questions provide different contexts. In different contexts, the features of the same linguistic primitive, extracted using a recurrent neural network, will differ. Thus, it is necessary to sample different questions for each linguistic primitive. To preserve the different contexts of the linguistic primitives, the linguistic primitive database stores the original questions, not just the linguistic primitive itself.
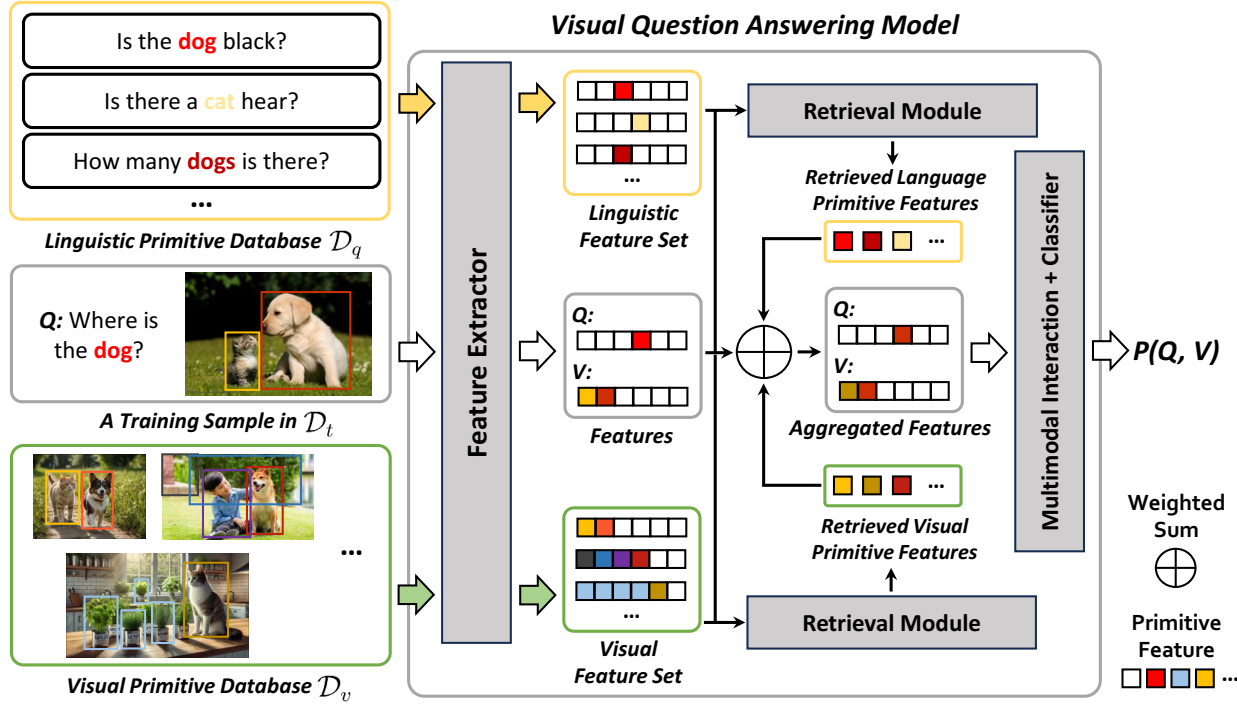
Figure 2: The overall framework of the proposed framework.

The resulting linguistic primitive database is represented as $\mathcal{D}_q = \bigcup_{p \in \mathcal{S}_q} f_q(p)$, where $f_q(p) = \{p_i\}_{i=1}^{T_q}$ represents the T_q sampled questions containing the linguistic primitive p.

Visual Primitive Database. The process of constructing the visual primitive database relies on the scene graph annotations, which contains information about object categories and attributes in the images. Objects in the images are treated as visual primitives, with category and attribute information serving as the labels of these primitives. After removing duplicates from the images in $\mathcal{D}_t$, the resulting set of labels is denoted as $\mathcal{S}_v$. Similar to linguistic primitives, even though different visual primitives may share the same label, they represent different visual expressions of objects with the same type/attribute. Therefore, for each label, T_v images containing the visual primitive with the label are sampled. Since different VQA models use different visual encoders, the input forms are not limited to object-level features. To address this, the visual primitive database stores the original images, not just the visual primitives. The resulting visual primitive database is represented as $\mathcal{D}_v = \bigcup_{l \in \mathcal{S}_v} f_v(l)$, where $f_v(l) = \{p_i^{(l)}\}_{i=1}^{T_v}$ represents the T_v sampled images containing visual primitives with label l.

3.3 Feature Retrieval and Aggregation

To perform retrieval and aggregation at the feature level, both the training samples and all primitives in the two primitive databases ($\mathcal{D}_q$ and $\mathcal{D}_v$) need to pass through the feature extractors of the VQA model. For a question Q, we obtain its feature $\boldsymbol{h}_q = g_q(Q) = \{\boldsymbol{h}_q^i\}_{i=1}^N$, where $g_q(\cdot)$ denotes the linguistic feature extractor of the VQA model. Here, $\boldsymbol{h}_q^i$ rep-

resents the feature of the i-th word, and N is the number of words. For an image V, the visual feature extractor $g_v(\cdot)$ typically produces two types of features: "object-level" features or "patch-level" features, which can be represented as $\boldsymbol{h}_v = g_v(V) = \{\boldsymbol{h}_v^i\}_{i=1}^M$, where $\boldsymbol{h}_v^i$ represents the feature of the i-th object/patch, and M is the number of objects/patches.

After obtaining the primitive-level features (word-level, object-level, patch-level), for each primitive feature in the training sample, retrieval is performed on the primitive feature sets of the questions and images in $\mathcal{D}_q$ and $\mathcal{D}_v$ (processed by $g_q(\cdot)$ and $g_v(\cdot)$). Specifically, for a primitive feature $\boldsymbol{p}$ in the training sample, the top K_q most similar primitive features $\{\boldsymbol{p}_q^{(i)}\}_{i=1}^{K_q}$ are retrieved from the primitive feature set in $\mathcal{D}_q$, and the top K_v most similar primitive features $\{\boldsymbol{p}_v^{(i)}\}_{i=1}^{K_v}$ are retrieved from the primitive feature set in $\mathcal{D}_v$. We use cosine similarity $\cos(\cdot, \cdot)$ to measure the similarity between two primitive features. The features are aggregated using a weighted average

$$\boldsymbol{p}_a = \boldsymbol{p} + w_q \cdot \frac{\sum_{i=1}^{K_q} \cos(\boldsymbol{p}, \boldsymbol{p}_q^{(i)})}{K_q} + w_v \cdot \frac{\sum_{i=1}^{K_v} \cos(\boldsymbol{p}, \boldsymbol{p}_v^{(i)})}{K_v}, \quad (1)$$

where the hyperparameters w_q and w_v control the contributions of the different modality primitive databases. We use the aggregated feature $\boldsymbol{p}_a$ to replace the original primitive feature $\boldsymbol{p}$ for training.

3.4 Optimization

The proposed framework is applicable to different VQA baseline models, using the same optimization approach as the baseline model without introducing additional training losses

or constraints. Therefore, for a VQA baseline model using the training loss $\mathcal{L}$, for a training sample (Q, V) with ground-truth A, the training loss for both the baseline model and the model with our framework is the same:

$$\mathcal{L} = \text{loss}(P(Q, V), A), \tag{2}$$

where $P(Q, V)$ represents the output of the VQA model (*e.g.*, a distribution vector over the number of categories), and $\text{loss}(\cdot, \cdot)$ represents the training loss function, such as the cross-entropy loss used in the UpDn model [Anderson *et al.*, 2018].

4 GQA-MSCG Dataset

In order to quantitatively evaluate the generalization ability of VQA models to multi-sourced novel compositions, we construct the GQA-MSCG dataset based on the GQA dataset [Hudson and Manning, 2019]. The process of constructing the GQA-MSCG dataset consists of three main steps: composition extraction, sample filtering, and sample classification.

Composition Extraction. We use the same steps as in Section 3.2 to extract the linguistic and visual primitives for all samples in the train_balanced split $\mathcal{D}_t$ of the GQA dataset. The linguistic and visual primitives for a sample $s \in \mathcal{D}_t$ are represented by $\mathcal{P}_s^q = \{p_s^{q_i}\}_{i=1}^N$ and $\mathcal{P}_s^v = \{p_s^{v_i}\}_{i=1}^M$, respectively, where N and M denote the number of linguistic and visual primitives in the sample. The compositions in sample s are represented as $\mathcal{C}_s = \{[p_1, p_2] | p_1 \in \mathcal{P}_s^q \cup \mathcal{P}_s^v, p_2 \in \mathcal{P}_s^q \cup \mathcal{P}_s^v, p_1 \neq p_2\}$. Thus, the complete set of primitives in $\mathcal{D}_t$ can be expressed as $\mathcal{P}_{\mathcal{D}_t} = \bigcup_{s \in \mathcal{D}_t} \mathcal{P}_s^q \cup \mathcal{P}_s^v$, and the complete set of compositions can be expressed as $\mathcal{C}_{\mathcal{D}_t} = \bigcup_{s \in \mathcal{D}_t} \mathcal{C}_s$. Similarly, the compositions in the val_all split $\mathcal{D}_v$ of the GQA dataset can be processed to obtain all compositions. We denote the set of all compositions as $\mathcal{C}_{\mathcal{D}_v}$.

Sample Filtering. For a sample $s \in \mathcal{D}_v$, if the sample satisfies the following conditions simultaneously, it is added to the candidate test sample set $\mathcal{D}_c$:

- All primitives in the sample must appear in the training set — $\forall p \in \mathcal{P}_s^q \cup \mathcal{P}_s^v, p \in \mathcal{P}_{\mathcal{D}_t}$.
- The sample must contain at least one composition not seen in the training set (i.e., an novel composition) — $\exists c \in \mathcal{C}_s, c \notin \mathcal{C}_{\mathcal{D}_t}$.

Sample Classification. Based on the modality of the primitives in a composition, we categorize three types of novel compositions: [Linguistic primitive, Linguistic primitive] (LL), [Visual primitive, Visual primitive] (VV), and [Linguistic primitive, Visual primitive] (LV). Test samples in $\mathcal{D}_c$ are classified into seven categories: (1) LL: Samples containing only novel compositions of the type LL. (2) VV: Samples containing only novel compositions of the type VV. (3) LV: Samples containing only novel compositions of the type LV. (4) $LL+VV$: Samples containing both novel compositions of the type LL and VV. (5) $LL+LV$: Samples containing both novel compositions of the type LL and LV. (6) $VV+LV$: Samples containing both novel compositions of the type VV and LV. (7) $LL+VV+LV$: Samples containing novel compositions of all three types: LL, VV, and LV. Samples in LL, VV and LV are used to evaluate the multi-sourced compositional generalization ability, while samples in $LL+VV$,
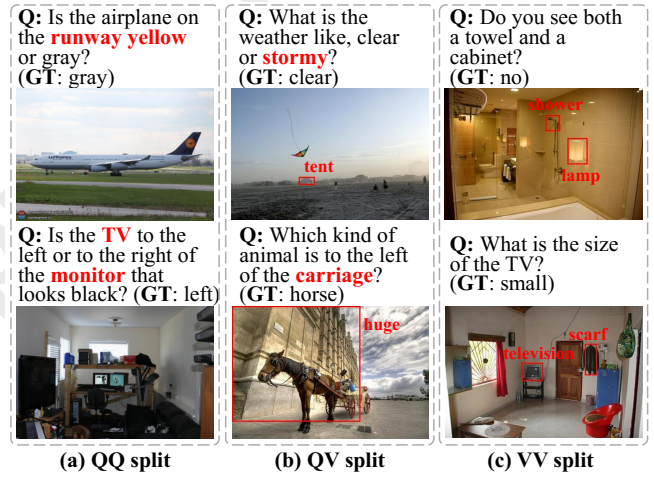

(a) QQ split   **(b) QV split**   **(c) VV split**

Figure 3: Level-1 samples in the GQA-MSCG dataset.

(a) QQ+QV split   **(b) QQ+VV split**   **(c) VV+QV split**
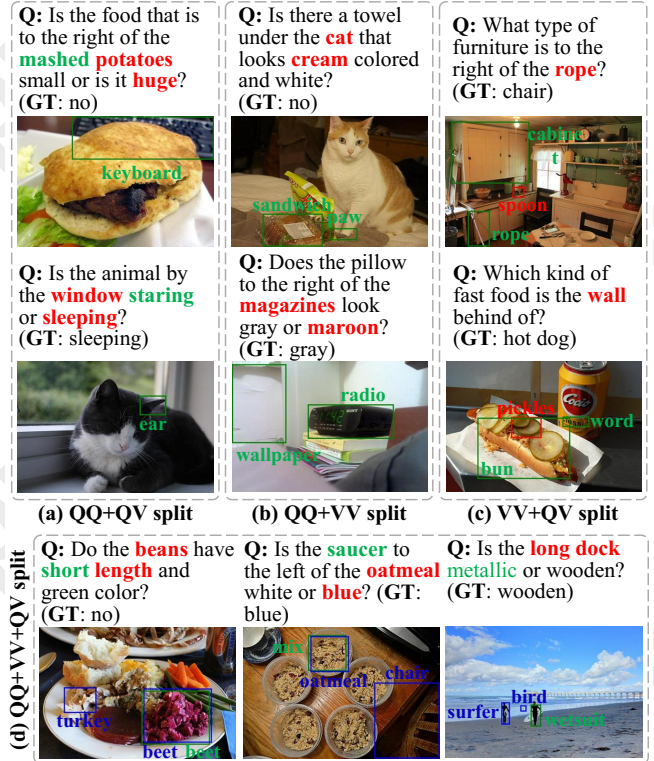
Figure 4: Level-2 samples and Level-3 samples in the GQA-MSCG dataset.

$LL+LV$, $VV+LV$ and $LL+VV+LV$ are used to further evaluate the impact of the co-occurrence of different types of novel compositions on model performance.

For each category of test samples, we randomly sample 5,000 samples from $\mathcal{D}_c$, resulting in a total of 35,000 samples for the GQA-MSCG dataset. Samples containing x types of novel compositions are referred to as Level-x samples, and the difficulty of the samples increases as x increases. For example, samples in LL, VV and LV are Level-1 samples,

| Type | Model | Level-1 | | | Level-2 | | | Level-3 | Overall |
		LL	VV	LV	LL+VV	LL+LV	VV+LV	LL+VV+LV	
Small Model ($\leq$ 0.2B)	CFR [Nguyen *et al.*, 2022]	74.78	72.18	73.36	72.26	73.56	70.94	69.78	72.41
	+ RAG (Ours)	**76.28**	**73.88**	**75.64**	**73.58**	**75.66**	**73.18**	**71.22**	**74.21**
Large Model ($\geq$ 7B)	LLaVA-1.5 [Liu *et al.*, 2023]	70.22	70.34	69.62	67.04	68.84	68.40	64.14	68.37
	LLaVA-1.6 [Liu *et al.*, 2024a]	72.36	71.52	72.60	69.70	69.50	69.98	67.60	70.46
	Qwen-VL [Bai *et al.*, 2023]	72.24	65.54	69.06	68.18	72.08	66.04	66.32	68.49
	+ RAG (Ours)	**74.68**	**68.90**	**71.98**	**71.50**	**74.82**	**68.92**	**69.88**	**71.53**

Table 1: Accuracy (%) on the GQA-MSCG dataset.

$LL + VV$, $LL + LV$ and $VV + LV$ are Level-2 samples, $LL + VV + LV$ are Level-3 samples. As the level increases, the co-occurrence count of novel composition categories in the test samples continues to rise, making the difficulty of the test samples increasingly higher. The test samples of GQA-MSCG are shown in Figure 3 and Figure 4, where the words or visual regions of the same color in a sample represent a pair of novel compositions in the sample.

5 Experiments

5.1 Experimental Setup

Dataset. We evaluate proposed frameworks on the VQA task, three datasets are selected to validate the effectiveness of the proposed frameworks: the GQA dataset [Hudson and Manning, 2019], the VQA v2 dataset [Goyal *et al.*, 2017] and our GQA-MSCG dataset. The GQA dataset is a widely used large-scale dataset in VQA, containing a large number of template-based compositional questions. The VQA v2 dataset is an extended balanced version of the VQA v1 dataset [Antol *et al.*, 2015], where questions are human-made. These two datasets are often used to test the generalization ability of VQA models to IID data as well as their ability for compositional reasoning. Our GQA-MSCG dataset is used to evaluate the VQA model's ability to generalize consistently to multi-sourced novel compositions, which refers to the VQA model's ability to generalize to novel compositions of primitives from different modalities.

Baseline Models. We use CFR [Nguyen *et al.*, 2022] and Qwen-VL [Bai *et al.*, 2023] as baseline models. The baseline models combined with the proposed framework are referred to as CFR+RAG and Qwen-VL+RAG, respectively. CFR is a representative small model in the VQA task with a parameter size less than 0.2B. Qwen-VL is a popular open-sourced multimodal large model (with more than 7B parameters), which can be applied to various vision-and-language tasks such as VQA, image captioning, visual dialogue, and others. For both CFR and Qwen-VL, we reimplemented them based on their officially released code.

Implementation Details. For experiments on all three datasets including GQA, GQA-MSCG and VQA v2, we fine-tune Qwen-VL and Qwen-VL+RAG with LoRA [Hu *et al.*, 2022] with a maximum of 2 epochs. For CFR+RAG and Qwen-VL+RAG, we set $w_q = 0.6$ and $w_v = 0.4$.

For experiments on the GQA dataset and the GQA-MSCG dataset, we fine-tune CFR, Qwen-VL, CFR+RAG, and Qwen-VL+RAG using the train_balanced split of the GQA dataset and selected the best-performing model weights on the val_balanced split of GQA. Using these model weights, we present the experimental results on the test-dev split of the GQA dataset and all seven test splits of our GQA-MSCG dataset. The maximum number of epochs for fine-tuning CFR and CFR+RAG was set to 12. The sampled number T_q and T_v for constructing $\mathcal{D}_q$ and $\mathcal{D}_q$ are set to 8 and 32, respectively. Moreover, we use the scene graph provided by GQA to construct $\mathcal{D}_v$. The number of aggregated primitive K_q and K_v are set to 4 and 16, respectively. Distinctively, for experiments on the VQA v2 dataset, we set $T_q = 1$, $T_v = 32$, $K_q = 4$ and $K_v = 4$. For constructing $\mathcal{D}_v$, we use the object categories detected by Faster R-CNN [Ren *et al.*, 2016], ignoring the object attributes.

5.2 Multi-Sourced Compositional Generalization Performance

We evaluate the MSCG ability of VQA models on the proposed GQA-MSCG dataset. We compare with multimodal large models, including LLaVA-1.5 [Liu *et al.*, 2023] and LLaVA-1.6 [Liu *et al.*, 2024a], with the experimental results shown in Table 1. We can observe that: (1) As the co-occurrence count of novel composition categories in the test samples increases (Level-1 $\rightarrow$ Level-3), the model's performance gradually decreases. For example, LLaVA-1.5 has an accuracy of 70.22% in the LL split, 67.04% in the $LL + VV$ split, and 64.14% in the $LL + VV + LV$ split. (2) Different VQA models exhibit significant differences in their ability to generalize to novel compositions of different modality primitives. For example, LLaVA-1.5 performs better on the VV split than on the LL split, while Qwen-VL shows the opposite. (3) Both small VQA models (*e.g.*, CFR) and large VQA models (*e.g.*, Qwen-VL) benefit significantly from the proposed framework, which enhances their generalization ability to multi-sourced novel compositions.

As a result, based on the experimental results, the following conclusions can be drawn: (1) Existing VQA models still have shortcomings in handling complex compositional questions. (2) It is necessary to specifically consider the multi-sourced compositional generalization ability for VQA models. (3) The proposed framework is highly versatile and can be applied to different baseline models, improving their generalization ability to novel compositions of primitives from different modalities.

Type	Model	Accuracy
Attention-based	MAC [Hudson and Manning, 2018]	52.43
Graph-based	LCGN [Hu *et al.*, 2019]	55.63
NMN-based	MMN [Chen *et al.*, 2021]	59.14
Pretrain-based (zero-shot)	BLIP-2 (FlanT5$_{XXL}$) [Li *et al.*, 2023c]	44.70
	MiniGPT-4 [Zhu *et al.*, 2023]	43.50
	LLaVA-1.5 [Liu *et al.*, 2023]	61.93
	LLaVA-1.6 [Liu *et al.*, 2024a]	64.26
Pretrain-based (fine-tuned)	CFR [Nguyen *et al.*, 2022]	70.27
	+ RAG (Ours)	**71.70**
	Qwen-VL [Bai *et al.*, 2023]	54.98
	+ RAG (Ours)	**56.11**

Table 2: Accuracy (%) on the test-dev split of the GQA dataset.

Type	Model	Accuracy
Small Model ($\leq$ 0.2B)	DLR [Jing *et al.*, 2020]	57.96
	CF-VQA [Niu *et al.*, 2021]	63.73
	CLS [Mao *et al.*, 2024]	63.94
	ASS [Li *et al.*, 2024a]	64.00
	KDAR [Peng and Wei, 2024]	65.54
Large Model ($\geq$ 7B)	PNP-VQA [Tiong *et al.*, 2022]	63.30
	BLIP-2 (FlanT5$_{XXL}$) [Li *et al.*, 2023c]	65.20
	Qwen-VL [Bai *et al.*, 2023]	69.04
	+ RAG (Ours)	**69.82**

Table 3: Accuracy (%) on the val split of the VQA v2 dataset.

5.3 Independent and Identically Distributed Generalization Performance

We verify the improvement of our framework in IID generalization performance on the test-dev split of GQA [Hudson and Manning, 2019] and the val split of VQA v2 [Goyal *et al.*, 2017]. For GQA, we compare to five different types of VQA models, including attention-based MAC [Hudson and Manning, 2018], graph-based LCGN [Hu *et al.*, 2019], neural modular network based (NMN-based) MMN [Chen *et al.*, 2021], zero-shot pretrain-based BLIP-2 [Li *et al.*, 2023c], MiniGPT-4 [Zhu *et al.*, 2023], LLaVA-1.5 [Liu *et al.*, 2024b], LLaVA-1.6 [Liu *et al.*, 2023], and fine-tuned pretrain-based CFR [Nguyen *et al.*, 2022], Qwen-VL [Bai *et al.*, 2023]. For VQA v2, we compare to VQA models with less than 0.3B parameters (DLR [Jing *et al.*, 2020], CF-VQA [Niu *et al.*, 2021], CLS [Mao *et al.*, 2024], ASS [Li *et al.*, 2024a], KDAR [Peng and Wei, 2024]) and VQA models with more than 7B parameters (PNP-VQA [Tiong *et al.*, 2022], BLIP-2 [Li *et al.*, 2023c], Qwen-VL [Bai *et al.*, 2023]).

Experimental results on the test-dev split of the GQA dataset are shown in Table 2. From the table, we can observe that: (1) Pretrain-based models outperform other types of models (such as attention-based, graph-based, and NMN-based frameworks) in terms of IID generalization performance. (2) The proposed framework further enhances the IID generalization performance of baseline models (*e.g.*, 1.43%

Model	$\mathcal{D}_q$	$\mathcal{D}_v$	Level-1		
			LL	VV	LV
CFR [Nguyen *et al.*, 2022]	-	-	74.78	72.18	73.36
+ RAG (Ours)	✓	-	76.18	73.46	74.24
	-	✓	75.66	73.66	74.80
	✓	✓	**76.28**	**73.88**	**75.64**

Table 4: Ablation studies on the GQA-MSCG dataset.

and 1.13% absolute performance gains in accuracy for CFR and Qwen-VL, respectively). These experimental results demonstrate that the proposed framework is applicable to different task scenarios, including multi-sourced compositional generalization scenario and independent and identically distributed generalization scenarios.

Experimental results on the val split of the VQA v2 dataset are shown in the Table 3. The results demonstrate that our framework is inoffensive for IID generalization. The reason why the performance gains of the proposed framework on VQA v2 are less than on GQA is that the questions in GQA are more compositional and thus are more suitable to be improved by our framework. Such experimental results further prove the effectiveness of the proposed framework for improving the IID generalization ability of VQA models.

5.4 Ablation Studies

The results of ablation studies on GQA-MSCG using CFR as the baseline model are shown in Table 4. From the experimental results, the following conclusions can be drawn: (1) Conducting retrieval only on the question retrieval database $\mathcal{D}_q$ or the image retrieval database $\mathcal{D}_v$ during training can also improve the baseline model's performance, especially for novel compositions of specific modal primitives. For example, when retrieval is performed only on $\mathcal{D}_v$, improvements are more evident for compositions of visual modal primitives (the *VV* split). (2) Overall, using both the question and image retrieval databases simultaneously results in the best performance (highest average accuracy across all splits). These ablation study results confirm the necessity of multi-sourced retrieval during training and demonstrate that the linguistic primitive database $\mathcal{D}_q$ or visual primitive database $\mathcal{D}_v$ are complementary, helping to enhance the baseline model's generalization ability across different compositions.

5.5 Parameter Analysis

We analyze the influences of w_q and w_v on the MSCG ability of our framework, which denote the explicit contributions of the linguistic primitive database and the visual primitive database, respectively. Figure 5 shows the performance variations of the proposed framework with changing values of w_q and w_v. First, we conduct an initial analysis of w_q and w_v by setting them to the same value. The experimental results are shown in Figure 5 (a). It can be seen that as the values increase, the performance of the proposed framework peaks when $w_q = w_v = 0.4$, and further increasing w_q and w_v does not improve the performance. Furthermore, we analyze the effectiveness of the proposed framework with $w_q \neq w_v$ by

(a) Experimental results with $w_q = w_v$.

(b) Experimental results with $w_q = 0.4$ and varying w_v.

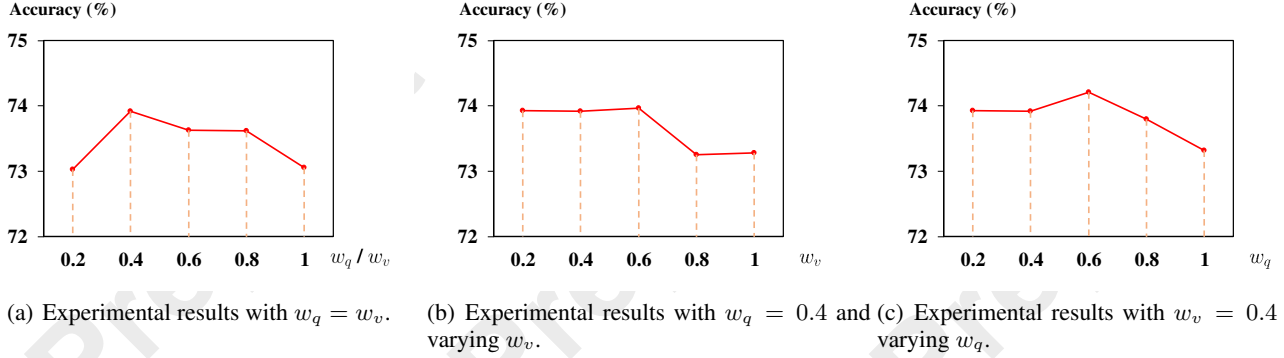(c) Experimental results with $w_v = 0.4$ and varying w_q.

Figure 5: Parameter analysis using CFR as the baseline model on the GQA-MSCG dataset.

Figure 6: The qualitative comparison between CFR+RAG (Ours) and CFR on the GQA-MSCG dataset.

fixing w_q or w_v at 0.4 and adjusting the other parameter (*i.e.*, w_v or w_q), as shown in Figures 5 (b) and (c). We can observe that: (1) The performance of CFG+RAG fluctuates obviously when adjusting either w_q or w_v. (2) CFR+RAG performs best with setting $w_q = 0.6$ and $w_v = 0.4$ simultaneously. Based on the above experimental results, we set w_q to 0.6 and w_v to 0.4 for all experiments.

5.6 Qualitative Analysis

For each split of Level-1, we provide two qualitative examples of CFR+RAG (Ours) and CFR (the baseline model) on the GQA-MSCG dataset in Figure 6. In the figure, novel compositions of linguistic primitives are highlighted in red text, and visual primitives are enclosed in red boxes. It can be observed that for test samples with novel compositions composed of primitives from different modalities, the proposed framework helps the baseline model make more accurate predictions. For example, in the first sample from the GQA-MSCG dataset's LL split, the question is "Are the jeans different in color than the flowers?", which includes the novel composition of linguistic primitives "jeans + flowers". The baseline model CFR gives an incorrect answer, "no". In contrast, after incorporating CFR into the proposed framework, it correctly predicts "yes". Moreover, for test samples in the VV and LV splits, the proposed framework still provides accurate answers, thanks to the alignment of primitives with the same semantics across different contexts and modalities. These qualitative examples demonstrate that the proposed framework enhances the baseline model's generalization ability to multi-sourced novel compositions, proving the effectiveness of the framework.

6 Conclusion

In this paper, we explored the multi-sourced compositional generalization ability of models in the context of VQA. We have presented a retrieval-augmented training framework to encourage VQA models to learn unified representations for the same semantic compositions by aligning semantically equivalent primitives across different modalities at the feature level. The proposed framework can be seamlessly incorporated into existing VQA models to improve their multi-sourced compositional generalization ability. We extend the GQA dataset to construct a GQA-MSCG dataset, which enables the quantitative evaluation of the multi-sourced compositional generalization ability for VQA models. Experimental results demonstrate that our framework can improve not only the multi-sourced compositional generalization ability, but also the IID generalization ability.

Acknowledgements

Contribution Statement

Chuanhao Li and Wenbo Ye contributed equally to this work.

References

[Akula *et al.*, 2021] Arjun Akula, Varun Jampani, Soravit Changpinyo, and Song-Chun Zhu. Robust visual reasoning via language guided neural module networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11041–11053, 2021.

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

[Antol *et al.*, 2015] 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/CVF International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015.

[Bahdanau *et al.*, 2018] Dzmitry Bahdanau, Shikhar Murty, Michael Noukhovitch, Thien Huu Nguyen, Harm de Vries, and Aaron Courville. Systematic generalization: what is required and can it be learned? *arXiv preprint arXiv:1811.12889*, 2018.

[Bai *et al.*, 2023] 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.

[Bird *et al.*, 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[Chai *et al.*, 2024] Yuyang Chai, Zhuang Li, Jiahui Liu, Lei Chen, Fei Li, Donghong Ji, and Chong Teng. Compositional generalization for multi-label text classification: A data-augmentation approach. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 17727–17735, 2024.

[Chen *et al.*, 2021] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 655–664, 2021.

[Dankers *et al.*, 2022] Verna Dankers, Elia Bruni, and Dieuwke Hupkes. The paradox of the compositionality of natural language: A neural machine translation case study. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4154–4175, 2022.

[Goyal *et al.*, 2017] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017.

[Hu *et al.*, 2019] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10294–10303, 2019.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[Hudson and Manning, 2018] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[Hudson and Manning, 2019] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019.

[Jing *et al.*, 2020] Chenchen Jing, Yuwei Wu, Xiaoxun Zhang, Yunde Jia, and Qi Wu. Overcoming language priors in vqa via decomposed linguistic representations. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11181–11188, 2020.

[Jing *et al.*, 2024] Chenchen Jing, Yukun Li, Hao Chen, and Chunhua Shen. Retrieval-augmented primitive representations for compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, pages 2652–2660, 2024.

[Li *et al.*, 2021] Yafu Li, Yongjing Yin, Yulong Chen, and Yue Zhang. On compositional generalization of neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4767–4780, 2021.

[Li *et al.*, 2022] Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3032–3041, 2022.

[Li *et al.*, 2023a] Chuanhao Li, Zhen Li, Chenchen Jing, Yunde Jia, and Yuwei Wu. Exploring the effect of

primitives for compositional generalization in vision-and-language. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19092–19101, 2023.

[Li *et al.*, 2023b] Juncheng Li, Siliang Tang, Linchao Zhu, Wenqiao Zhang, Yi Yang, Tat-Seng Chua, and Fei Wu. Variational cross-graph reasoning and adaptive structured semantics learning for compositional temporal grounding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2023.

[Li *et al.*, 2023c] 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.

[Li *et al.*, 2024a] Chuanhao Li, Chenchen Jing, Zhen Li, Yuwei Wu, and Yunde Jia. Adversarial sample synthesis for visual question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 20(12):1–24, 2024.

[Li *et al.*, 2024b] Yun Li, Zhe Liu, Hang Chen, and Lina Yao. Context-based and diversity-driven specificity in compositional zero-shot learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17037–17046, 2024.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.

[Liu *et al.*, 2024b] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

[Mao *et al.*, 2024] Aihua Mao, Feng Chen, Ziying Ma, and Ken Lin. Overcoming language priors in visual question answering with cumulative learning strategy. *Neurocomputing*, 608:128419, 2024.

[Naeem *et al.*, 2021] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 953–962, 2021.

[Nguyen *et al.*, 2022] Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering. pages 4557–4565, 2022.

[Niu *et al.*, 2021] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12700–12710, 2021.

[Pantazopoulos *et al.*, 2022] George Pantazopoulos, Alessandro Suglia, and Arash Eshghi. Combine to describe: Evaluating compositional generalization in image captioning. In *Proceedings of the Association for Computational Linguistics: Student Research Workshop (ACLSRW)*, pages 115–131, 2022.

[Peng and Wei, 2024] Daowan Peng and Wei Wei. Overcoming language priors for visual question answering based on knowledge distillation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[Ren *et al.*, 2016] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 39(6):1137–1149, 2016.

[Tiong *et al.*, 2022] 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. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967, 2022.

[Yamada *et al.*, 2024] Moyuru Yamada, Vanessa D'Amario, Kentaro Takemoto, Xavier Boix, and Tomotake Sasaki. Transformer module networks for systematic generalization in visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[Zhu *et al.*, 2023] 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.