In-Context Compositional Generalization for Large Vision-Language Models

Chuanhao Li[1], Chenchen Jing[3], Zhen Li[1], Mingliang Zhai[1,2], 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
[3]School of Computer Science, Zhejiang University, Hangzhou, China
Correspondence: wuyuwei@bit.edu.cn

Abstract

Recent work has revealed that in-context learning for large language models exhibits compositional generalization capacity, which can be enhanced by selecting in-context demonstrations similar to test cases to provide contextual information. However, how to exhibit in-context compositional generalization (ICCG) of large vision-language models (LVLMs) is non-trival. Due to the inherent asymmetry between visual and linguistic modalities, ICCG in LVLMs faces an inevitable challenge—redundant information on the visual modality. The redundant information affects in-context learning from two aspects: (1) Similarity calculation may be dominated by redundant information, resulting in sub-optimal demonstration selection. (2) Redundant information in in-context demonstrations brings misleading contextual information to in-context learning. To alleviate these problems, we propose a demonstration selection method to achieve ICCG for LVLMs, by considering two key factors of demonstrations: content and structure, from a multimodal perspective. Specifically, we design a diversity-coverage-based matching score to select demonstrations with maximum coverage, and avoid selecting demonstrations with redundant information via their content redundancy and structural complexity. We build a GQA-ICCG dataset to simulate the ICCG setting, and conduct experiments on GQA-ICCG and the VQA v2 dataset. Experimental results demonstrate the effectiveness of our method.

1 Introduction

Compositional generalization, understanding unseen compositions by recombining known primitives, is a fundamental ability of human intelligence (Fodor and Pylyshyn, 1988; Lake et al., 2017). Deploying such ability in machine learning models has received increasing attention and significant progress in vision and language. Nonetheless, most

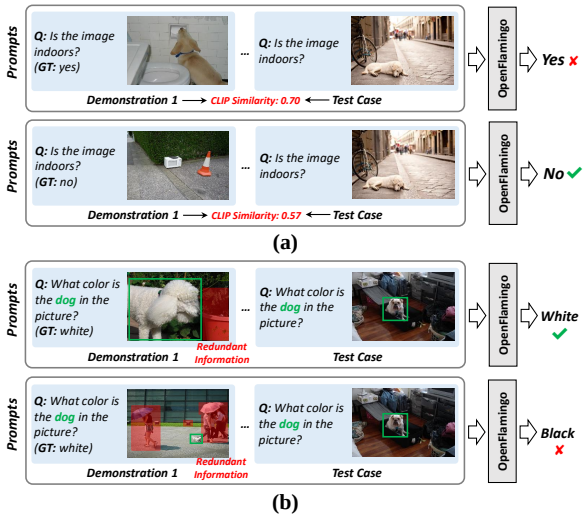

Figure 1: Illustration of the problems stemming from redundant information in ICCG for LVLMs. (a) Multimodal similarity is dominated by redundant information. (b) More redundant information in in-context demonstrations brings more difficulties in answering the sample.

work (Li et al., 2022, 2023b; Xu et al., 2023; Yang et al., 2023) focuses on boosting the compositional generalization ability of models by re-training or fine-tuning the models with explicit constraints, which is not applicable to large pre-trained models. Recently, in-context learning with large vision-language models (LVLMs) (Alayrac et al., 2022; Zhou et al., 2023) exhibits impressive performance on few-shot learning of various vision-language tasks. A question naturally arises: whether and how LVLMs exhibit in-context compositional generalization (ICCG)?

Recent work (Levy et al., 2022; An et al., 2023) has revealed that in-context learning for large language models (LLMs) exhibits compositional generalization by selecting in-context demonstrations similar to a test case. However, selecting demonstrations based on similarity is not applicable to large vision-language models for ICCG, because the inherent asymmetry between visual and linguis-

tic modalities leads to much redundant information in the visual modality. Such redundant information affects in-context learning in two aspects. Firstly, the calculation of multimodal similarity may be dominated by redundant information, resulting in sub-optimal demonstration selection. Secondly, redundant information in demonstrations brings misleading contextual information to in-context learning. For example, for a test case with an image that contains a dog in an indoor environment and a question "Is the image indoors?", the cosine similarity of CLIP features (Radford et al., 2021) is greater when a dog was included in the demonstration rather than an indoor environment, though the dog is redundant to answer the question, as shown in Figure 1 (a). Moreover, the more redundant information in the selected in-context demonstrations, the harder it will be for the model to answer the test case, as shown in Figure 1 (b).

In this paper, we consider the asymmetry characteristics of multimodal data from both the content and structure perspectives, and propose a demonstration selection method for LVLMs to achieve in-context compositional generalization. A diversity-coverage-based matching score is designed to select demonstrations with maximum coverage and less redundant information. Specifically, for either a demonstration or a test case, we extract its primitives and construct constituent trees for its visual and linguistic information as its content and structure information, respectively. The calculation of the proposed matching score between a demonstration and a test case falls into two terms: (1) The content intersection and structure intersection between the demonstration and the test case at different modalities. (2) The symmetric difference between visual and linguistic content and the depth of the constituent tree, which model content redundancy and structural complexity of a demonstration, respectively. The first term is used to select diverse demonstrations with maximum coverage of the test case, while the second term is used to select demonstrations with less redundant information. In doing so, the redundant information mixed in in-context learning is reduced.

To quantitatively evaluate the in-context compositional generalization ability of LVLMs, we build a GQA-ICCG dataset based on the GQA dataset (Hudson and Manning, 2019). We first filter out samples from the val-balanced split of GQA that contain novel compositions of primitives seen in the train-balanced split of GQA, to construct the test set of the GQA-ICCG. For each test case, the GQA-ICCG contains a candidate demonstration set. The candidate set is constructed by randomly selecting 10 samples that contain each primitive in the test sample. This ensures that for each test case, there are various ways to select a set of in-context demonstrations that fully cover its primitives to satisfy the compositional generalization setting. We evaluated our method with four LVLMs varies in parameters (3B to 9B), OpenFlamingo (Awadalla et al., 2023a), Otter (Li et al., 2023a), FROMAGe (Koh et al., 2023) and IDEFICS (Laurençon et al., 2024) on our GQA-ICCG dataset and the VQA v2 dataset (Goyal et al., 2017). Experimental results demonstrate the effectiveness of our method.

To sum up, our contributions are as follows: (1) To the best of our knowledge, we are the first to investigate the in-context compositional generalization for large vision-language models, which is a promising few-shot paradigm. (2) We propose a demonstration selection method for large vision-language models to achieve in-context compositional generalization, taking both the content and structure of demonstrations into consideration.(3) We present a GQA-ICCG dataset to quantitatively evaluate the in-context compositional generalization ability of large vision-language models.

2 Related Work

2.1 Compositional Generalization

Compositional generalization is crucial for simulating the fundamental compositionality of human cognition (Fodor and Pylyshyn, 1988). and has attracted much attention in vision and language. Early works (Hudson and Manning, 2018; Shi et al., 2019; Akula et al., 2021; Bogin et al., 2021; Yamada et al., 2022) perform explicit reasoning by structuring input text into serialized reasoning steps to achieve compositional generalization. Moreover, several works (Saqur and Narasimhan, 2020; Zhang et al., 2021, 2022a; Li et al., 2022) strengthen the coupling of concepts between two modalities through cross-graph reasoning. Recently, Li et al. (2023b) improved compositional generalization by using a self-supervised training framework to learn primitive effects. Xu et al. (2023) handled various levels of novel compositions by optimizing models on multiple virtual sets. Yang et al. (2023) learned compositional representations by using a ranking loss. Rahimi et al. (2023) investigated which factors can improve compositional general-

ization in training data design. These works focus on designing components or frameworks that can be incorporated into the training to improve compositional generalization. In contrast, we investigate how to improve compositional generalization for LVLMs in a training-free in-context learning paradigm, which is plug-and-play.

2.2 In-Context Learning

In-context learning is a promising few-shot paradigm for large language models (LLMs) (Brown et al., 2020; Hosseini et al., 2022; Wei et al., 2022; Chiang et al.; Touvron et al., 2023; Wu et al., 2023), where the models are provided with contextual information for each test case using a prompt with several demonstrations. Existing works (Alayrac et al., 2022; Peng et al., 2023; Zhou et al., 2023) have shown that this ability also exists in large vision-language models (LVLMs). There are several works that improve the in-context compositional generalization of LLMs. Qiu et al. (2022) investigated how compositional generalization is affected by the size of LLMs in fine-tuning, prompt tuning and in-context learning. Drozdov et al. (2022) proposed the dynamic least-to-most prompting technique to improve compositional generalization of LLMs in realistic semantic parsing tasks. Levy et al. (2022) improved compositional generalization for semantic parsing by leveraging in-context learning. An et al. (2023) explored in-context compositional generalization for LLMs and revealed several factors in selecting demonstrations. Differently, we are the first to investigate in-context compositional generalization for LVLMs, and alleviate the LVLMs-specific redundant information problem stemming from the asymmetry between visual and linguistic modalities.

2.3 Demonstration Selection

Demonstration selection has proven to be a key component of in-context learning. For large language models, Liu et al. (2022) utilized k-nearest neighbors as demonstrations. Ye et al. (2022) selected demonstrations that are both relevant and complementary by maximizing marginal relevance. Poesia et al. (2022) tuned target similarity for selecting demonstrations. Levy et al. (2022) and An et al. (2023) selected diverse demonstrations with a structure similar to test cases. Moreover, several works (Pasupat et al., 2021; Rubin et al., 2022; Zhang et al., 2022b; Li et al., 2023e) train a retriever to retrieve demonstrations based on the similarities learned by the retriever. For LVLMs, Alayrac et al. (2022) retrieved similar images in the support set to collect demonstrations. Yang et al. (2024) explored different types of image similarities to select demonstrations for image captioning. Zhou et al. (2024) proposed a visual in-context learning method for LVLMs. The studies above for both large language models and LVLMs use a well-designed single-modal similarity measure to select demonstrations for test cases. By contrast, we propose a demonstration selection method for LVLMs by considering the content and structure of demonstrations from a multimodal perspective. Moreover, several works (Yang et al., 2022; Li et al., 2023d) select demonstrations based on the multimodal content similarity for visual question answering. Differently, we consider both the content similarity and the structure similarity to select demonstrations, as we focus on in-context compositional generalization, where the structure similarity of visual concepts and textual semantics is crucial for understanding unseen compositions.

3 Method

3.1 Overview

The overview of our demonstration selection method is shown in Figure 2. By iteratively selecting the demonstration with the largest matching score to the test case, we collect a in-context demonstration set for each test case to perform in-context learning. The matching score is calculated by considering six terms: content diversity, content coverage, content redundancy, structural coverage, structural diversity, and structural complexity. Specifically, for a test case $X = (X_l, X_v)$, where X_l and X_v denotes the linguistic and visual information of X respectively, the matching score between X and a demonstration $X' = (X'_l, X'_v)$ is defined as

$$M(X, X') = w_c \cdot C(X, X') + S(X, X'), \qquad (1)$$

where $C(X, X')$ and $S(X, X')$ denote the matching score of content and structure, respectively, and w_c is a hyperparameter to balance the content matching score and structural matching score.

3.2 Content Matching

To measure whether X' matches X in content, we consider three factors: coverage, diversity and redundancy, and the content matching score of X

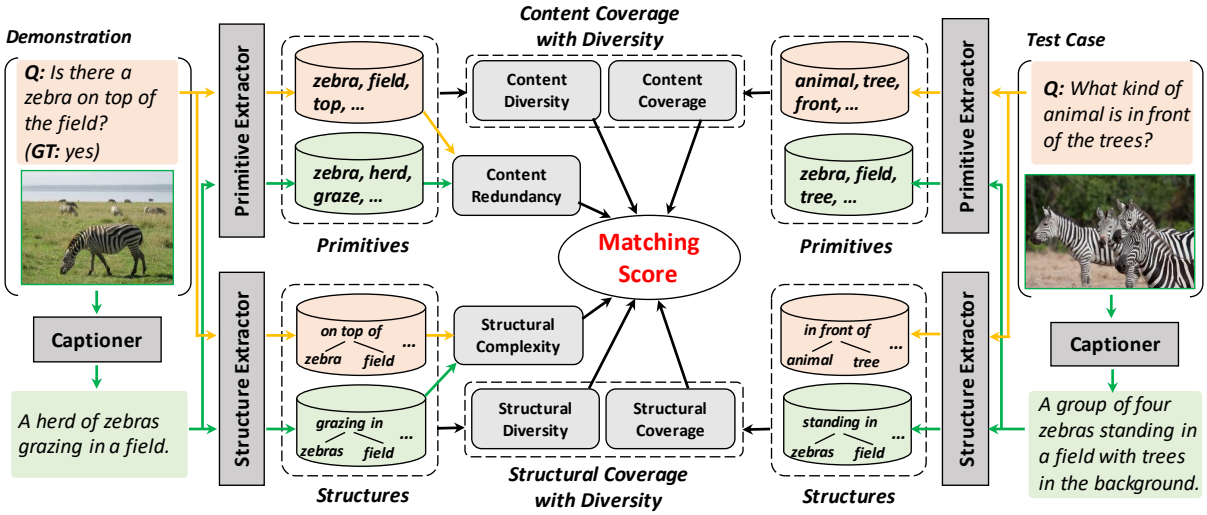

Figure 2: Overview of the proposed demonstration selection method. We design a matching score between demonstrations and test cases considering both their content and structure, to greedily select demonstrations.

and X' is defined as

$$C(X, X') = |\underbrace{P(X) \cap [P(X') - P(E)]}_{coverage\quad diversity}| - w_r \cdot \underbrace{R(X')}_{redundancy}, \tag{2}$$

where $P(\cdot)$ is a function that outputs the primitives of the input, E is the set of already selected in-context demonstrations, and $R(\cdot)$ is a function that computes the redundancy of the input sample.

Content Coverage aims to make the in-context demonstrations provide as much of the same information as possible to the test case. As primitives are the build blocks of compositions and are crucial for compositional generalization, we use the coverage degree of primitives to represent the coverage degree of content. For LVLMs, the primitives of samples come from both visual and linguistic modalities, and the primitives from different modalities play a different role. For linguistic information, we use the benepar toolkit (Kitaev and Klein, 2018; Kitaev et al., 2019) to extract its words and phrases as primitives. For visual information, we use the pre-trained BLIP-2 model (Li et al., 2023c) to generate captions for it, and then extract primitives in the same way as for linguistic information. As a result, we consider the primitives of the two modalities separately to ensure content coverage, and expand the $coverage$ term in Equation (2) as

$$P(X) \cap P(X') = \sum_{m \in \mathcal{M}} \sum_{m' \in \mathcal{M}} p_{m,m'} P(X_m) \cap P(X'_{m'}), \tag{3}$$

where $\mathcal{M} = \{l, v\}$ denotes the modality set, $p_{m,m'}$ is a binary weight whose value is either 1 or 0. If $p_{m,m'} = 1$, it means we consider using $P(X'_{m'})$ to cover $P(X_m)$ during content coverage.

Content Diversity focuses on achieving wider primitive coverage on the test case with in-context demonstrations. Similarly, we consider the differences between the two modalities to expand the $diversity$ term in Equation (2) as

$$\begin{aligned} &P(X) \cap [P(X') - P(E)] \\ &= \sum_{m \in \mathcal{M}} \sum_{m' \in \mathcal{M}} p_{m,m'} P(X_m) \cap [P(X'_{m'}) - P(E_{m'})]. \end{aligned} \tag{4}$$

Content Redundancy reflects the degree of redundancy of information in a demonstration, which should be low to provide little misleading contextual information for in-context learning. As the redundant information is caused by asymmetry between visual and linguistic modalities, we model the content redundancy as the size of the symmetric difference between visual and linguistic content. The size of the symmetric difference is computed by counting the number of primitives that are redundant in the two modalities:

$$R(X') = \left| P(X'_l) \cup P(X'_v) - P(X'_l) \cap P(X'_v) \right|. \tag{5}$$

3.3 Structural Matching

For a test case, when two demonstrations meet the same content matching, the one with a more similar structure can provide more appropriate contextual information. Therefore, both content matching and structural matching need to be considered.

Similar to content matching, we consider both coverage and diversity for structural matching. Furthermore, we consider the structural complexity of demonstrations. The structural matching score of

X and X' is defined as

$$S(X, X') = |\underbrace{T(X) \cap [T(X')}_{coverage} - \underbrace{T(E)]}_{diversity}| - w_d \cdot \underbrace{D(X')}_{complexity}, \tag{6}$$

where $T(\cdot)$ is a function that extracts the sub-structures of the constituent tree for the input, E has the same meaning as it in Equation (2), and $D(X')$ denotes a function to compute the complexity of the input sample.

Structural Coverage concerns the structural similarities between demonstrations and test cases. For each demonstration and test case, we first use the benepar toolkit to construct constituent trees for their linguistic information and visual captions. Then we count sub-trees of each constituent tree with a depth of no more than 3 as its sub-structures, *i.e.*, the output of $T(\cdot)$. We define the structural coverage score of X and X' as

$$T(X) \cap T(X') = \sum_{m \in \mathcal{M}} \sum_{m' \in \mathcal{M}} t_{m,m'} T(X_m) \cap T(X'_{m'}), \tag{7}$$

where $t_{m,m'}$ is a binary weight whose value is either 1 or 0, and play a role similar to $p_{m,m'}$ in Equation (3).

Structural Diversity focuses on the repetitiveness of structures among demonstrations, which has been proved to affect the ability of compositional generalization (Oren et al., 2021; An et al., 2023). We achieve high structural diversity by removing structures from already selected in-context demonstrations when performing structural coverage, and the structural matching score used in this process is written as

$$\begin{aligned} &T(X) \cap [T(X') - T(E)] \\ &= \sum_{m \in \mathcal{M}} \sum_{m' \in \mathcal{M}} t_{m,m'} T(X_m) \cap [T(X'_{m'}) - T(E_{m'})]. \end{aligned} \tag{8}$$

Structural Complexity measures the difficulty in understanding a demonstration. A demonstration with a more complex structure often requires stronger reasoning capabilities to understand, which provides more obscure contextual information. We compute the structural complexity by considering visual and linguistic modalities simultaneously, and define it as

$$D(X') = \sum_{m \in \mathcal{M}} d_m \cdot depth(X'_m), \tag{9}$$

where $depth(\cdot)$ computes the depth of the constituent tree of the input sentence as its complexity, and d_m is a binary weight to measure whether the complexity of modality m is considered.

To sum up, our demonstration selection method considers two key factors of demonstrations: content and structure, from a multimodal perspective, to select diverse in-context demonstrations that: (1) share same primitives and similar structures with the test case; (2) have less redundant information and low complexity.

4 Experiments

4.1 Experimental Settings

Datasets. To enable the evaluation of the ICCG ability, we introduce a GQA-ICCG dataset based on the GQA dataset (Hudson and Manning, 2019), a realistic large-scale visual question answering dataset for compositional reasoning. We build GQA-ICCG following three steps: (1) We use the benepar toolkit (Kitaev and Klein, 2018; Kitaev et al., 2019) to obtain primitives and primitive compositions for each sample in the train-balanced and val-balanced split of GQA. (2) To satisfy the compositional generalization setting, we filter out samples from the val-balanced split of GQA that contain novel compositions of primitives seen in the train-balanced split of GQA, as test cases of GQA-ICCG. (3) For each primitive in each test case, we random select 10 samples from the train-balanced split of GQA that contains the primitive, ensuring that sufficient contextual information can be found in the candidate demonstrations. Finally, there are 10, 000 test cases and 48, 103 candidate demonstrations in the GQA-ICCG dataset. We evaluate the proposed method on GQA-ICCG and the VQA v2 dataset (Goyal et al., 2017). The reason for choosing VQA v2 is to evaluate the compatibility of our method in improving ICCG and independent and identically distributed (IID) generalization.

Baseline Models. We test in-context compositional generalization with four LVLMs varies in parameters (3B to 9B), OpenFlamingo (OF) (Awadalla et al., 2023a), Otter (Li et al., 2023a), FROMAGe (Koh et al., 2023) and IDEFICS (Laurençon et al., 2024). OpenFlamingo is an open-source replication of DeepMind's Flamingo models (Alayrac et al., 2022), and exhibits good in-context learning capabilities. We do not test models with larger scales, as previous work suggests that ∼7B scale are large enough to demonstrate in-context learning capabilities and can generalize to larger model sizes (Koh et al., 2023). For all baselines, we use "*<image1> Question: {question1} Short answer: {answer1} <|endofchunk|>, ..., <image> Question:*

Method	OpenFlamingo-4B		OpenFlamingo-9B		Otter-7B		FROMAGe-6.7B		IDEFICS-9B	
	4-shot	8-shot	4-shot	8-shot	4-shot	8-shot	4-shot	8-shot	4-shot	8-shot
Random	38.89	38.52	38.22	43.58	37.43	37.96	36.35	36.39	49.59	49.96
Q Similarity	43.85	43.34	45.36	46.79	45.16	46.71	36.94	38.33	50.82	51.98
BCCS	46.13	47.64	47.67	47.86	41.12	44.01	35.03	37.15	50.63	51.58
RICES (Yang et al., 2022)	41. 03	42.88	41.24	45.12	44.89	45.37	35.87	37.09	50.73	51.32
Cover-LS (Levy et al., 2022)	46.54	48.20	46.30	48.55	45.71	47.91	36.63	39.18	51.38	52.35
SDC (An et al., 2023)	46.25	46.37	47.39	48.56	48.94	49.89	39.89	40.69	51.74	52.39
TOPK+MDL (Wu et al., 2023)	48.22	47.91	48.06	48.80	48.44	49.87	40.81	42.87	51.23	52.64
UnsupPR (Zhang et al., 2023)	39.04	38.62	38.67	43.89	37.73	38.28	36.78	36.92	49.66	50.36
VICL (Zhou et al., 2024)	39.12	39.78	39.13	43.88	37.61	38.45	37.12	37.10	50.33	50.57
Ours	**49.72**	**51.65**	**48.10**	**49.83**	**49.01**	**51.04**	**43.55**	**45.41**	**52.88**	**53.83**

Table 1: Accuracy (%) of state-of-the-art methods on GQA-ICCG.

{question} Short answer:" as prompts and set the width of beam search to 3. To reduce the recency bias (Zhao et al., 2021), we order the demonstrations in a prompt by increasing order of matching scores for all demonstration selection methods, such that in-context demonstrations with higher similarity are placed closer to the test case.

Implementation Details. To investigate the influence of each factor including content coverage, content diversity, content redundancy, structural coverage, structural diversity and structural complexity, we set the hyperparameters as follows: (1) We set $w_c = 100$ in Equation (1) to ensure the content coverage is satisfied firstly, which is critical for compositional generalization. (2) When investigating content/structural coverage, we set $w_r/w_d = 0$ in Equation (2)/(6), and exclude $P(E)/T(E)$ term. (3) When investigating content/structural diversity, we also set $w_r/w_d = 0$, but keep the $P(E)/T(E)$ term. (4) When investigating content redundancy, we ensure the priority of content coverage by setting $w_r = 0.05$ in Equation (2) to satisfy $|w_r| \cdot \max(R(X')) < 1$, as $\max(R(X')) = 19$ in GQA-ICCG. (5) For a reason similar to (4), we set $w_d = 0.04$ in Equation (6) that satisfies $|w_d| \cdot \max(D(X')) < 1$ to ensure the priority of structural coverage when investigating structural complexity, as $\max(D(X')) = 21$.

4.2 In-Context Compositional Generalization Performance

We compare our method with several demonstration selection methods, including: (1) Random—randomly select demonstrations from all the candidate demonstrations for each test case. (2) Q Similarity—use cosine similarities between CLIP features (Radford et al., 2021) of the questions from demonstrations and test cases to greedily select demonstrations. (3) BCCS—incorporate the similarity between captions obtained from BLIP-2 (Li et al., 2023c) into Q Similarity, for demonstration selection. (4) RICES (Yang et al., 2022)—use the average of question similarity and image similarity calculated using CLIP features to select demonstrations. (5) Cover-LS (Levy et al., 2022), SDC (An et al., 2023), and TOPK+MDL (Wu et al., 2023)—several state-of-the-art methods to select demonstrations for LLMs from a linguistic modality perspective. (6) UnsupPR (Zhang et al., 2023) and VICL (Zhou et al., 2024)—seveal state-of-the-art methods for visual in-context learning.

The experimental results on GQA-ICCG are listed in Table 1, where Ours denotes our method with $w_c = 100$, $w_r = 0.05$, $w_d = 0.04$, $P(E)$ and $T(E)$ activated, which performs best among the settings in Implementation Details (provided in the **appendix**). From the results, we have the following observations: (1) Our method achieves the best in-context compositional performance under different k-shots. On the baseline model IDEFICS-9B, our method achieves the best accuracy 52.88% and 53.83% for $k = 4$ and $k = 8$. This indicates that our method is more suitable for demonstration selection for LVLMs. (2) Our method significantly improves the in-context compositional performance for different baseline models. Take $k = 8$ as an example, our method achieves relative improvements of 5.28%, 1.27%, 1.15%, 0.36% and 1.44% in accuracy for different baseline models compared with SDC. These results demonstrate the generality of our method for improving different baseline models. For experimental results on more settings of k, please refer to the **appendix**.

4.3 Independent and Identically Distributed Generalization Performance

Experimental results of OpenFlamingo and IDEFICS at 4-shot paradigm on the VQA v2 dataset (Goyal et al., 2017) are listed in Table 2. We observe from the table that our method im-

Method	OF-3B	OF-4B	OF-9B	IDEFICS-9B
Random	40.36	45.43	47.84	53.84
Q Similarity	44.54	48.67	57.77	53.75
RICES (Yang et al., 2022)	45.70	49.00	54.80	55.40
SDC (An et al., 2023)	47.04	48.81	58.35	59.44
Ours	**47.55**	**49.63**	**59.30**	**60.71**
SQA† (Li et al., 2023d)	-	-	60.12	-

Table 2: Accuracy (%) of state-of-the-art methods at 4-shot on VQA v2. † represents the methods rely on ground-truth of the test case for demonstration selection, we do not compare with them for fair comparison.

Content Matching			Structural Matching			Accuracy
Cover	*Div*	*Red*	*Cover*	*Div*	*Comp*	
-	-	-	-	-	-	38.52
✓	-	-	-	-	-	45.09
✓	✓	-	-	-	-	45.65
✓	✓	✓	-	-	-	45.78
✓	✓	✓	✓	-	-	48.55
✓	✓	✓	✓	✓	-	51.15
✓	✓	✓	✓	✓	✓	**51.65**

Table 3: Ablation studies on GQA-ICCG. We use OF-4B at 8-shot as baseline model, whose performance is shown in the first line.

Hyperparameters				Accuracy	
$p_{l,l}$	$p_{v,v}$	$p_{v,l}$	$p_{l,v}$	*Cover*	*Div*
-	-	-	-	38.52	38.52
✓	-	-	-	43.69	44.33
-	✓	-	-	38.62	35.79
-	-	✓	-	39.51	36.86
-	-	-	✓	39.64	37.87
✓	✓	-	-	**45.09**	45.18
✓	-	✓	-	44.06	**45.65**
✓	-	-	✓	44.99	45.41
✓	✓	✓	-	44.08	45.28
✓	✓	✓	✓	43.92	44.54

Table 4: Accuracy (%) under different content coverage and content diversity settings on GQA-ICCG. We use OF-4B at 8-shot as baseline model, whose performance is shown in the first line.

proves the accuracy of all four baselines varies in parameters (3B to 9B), *e.g.*, 0.95% and 1.27% improvements on OpenFlamingo-9B and IDEFICS-9B compared with SDC. The experiment results demonstrate that our method works in not only the ICCG setting but also the regular IID setting.

4.4 Ablation Studies

To validate the effectiveness of different factors of our method, we evaluate different variants of our method by ablating certain factors. We use OpenFlamingo-4B as the baseline model, and experimental results on GQA-ICCG at 8-shot paradigm are shown in Table 3. We first investigate the effect of content matching from three aspects including coverage (Cover), diversity (Div) and redundancy (Red). We can observe that: (1) By gradually activating different factors, we observe better performance (45.09% $\rightarrow$ 45.65% $\rightarrow$ 45.78%). (2) Even only using content coverage, our method achieves significant performance improvements over the baseline model (ours 45.09% vs. baseline's 38.52%). Then we incorporate the structural matching into the content matching to investigate whether there is a further improvement, which also includes three aspects: coverage, diversity and complexity (Comp). We can observe that: (1) Using structural matching in addition to content matching can further improve model performance. (2) Structural diversity plays an important role in improving the performance of structural matching, which achieves 2.65% relative gains compared to only using structural coverage in accuracy. These observations suggest that all factors of our method are effective and complementary to each other for improving in-context compositional generalization.

4.5 Analysis of Content Matching

As the input of LVLMs contains information from both visual and linguistic modalities, which brings multiple types of content matching between demonstrations and test cases, rather than just match in a single modality, such as content matching between the visual information of a demonstration and the linguistic information of a test case. In this section, we focus on how information from different modalities affects the in-context compositional generalization of content matching, and provide experimental results for analysis.

Content Coverage. There are four binary hyperparameters that affect the content coverage in Equation (2), including $p_{l,l}$, $p_{v,v}$, $p_{v,l}$ and $p_{l,v}$. The results are listed in Table 4, where $p_{m,m'} = 1$ if there is a ✓ for $p_{m,m'}$ otherwise $p_{m,m'} = 0$. We can observe that: (1) Covering the test cases using only the visual content of the demonstration will not bring significant improvements, *e.g.*, $p_{l,v} = 1$ and $p_{v,v} = 1$. The reason is that the key semantics in the samples of visual question answering are mainly dominated by questions. (2) LVLMs achieve the best in-context compositional generalization when the linguistic content of in-context demonstrations covers both the visual and linguistic content of the test case as much as possible, *i.e.*, satisfying $p_{l,l} = 1$ and $p_{v,v} = 1$ simultaneously. These observations suggest that using information from inappropriate modalities to perform content

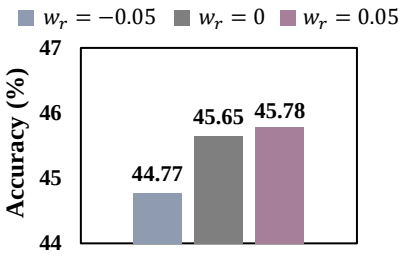

Figure 3: Accuracy (%) of OF-4B under different content redundancy settings at 8-shot on GQA-ICCG.

matching may degrade performance.

Content Diversity. For content diversity, The hyperparameters that affect it are the same as in content coverage. We investigate the improvements brought by introducing content diversity at different types of content matching, and the improvements are shown in Table 4. The results demonstrate that high content diversity brings gains on content coverage, and the gain is more significant when $p_{l,l} = 1$. The observations show that our method achieves best performance when $p_{l,l} = 1$ and $p_{v,l} = 1$ simultaneously.

Content Redundancy. For each demonstration, content redundancy is calculated independently according to Equation (5) and is independent of test cases. We test the effect of content redundancy by incorporating the redundancy term into the subversion of our method that performs best when investigating content diversity, *i.e.*, $p_{l,l} = 1$ and $p_{v,l} = 1$ simultaneously. The results are shown in Figure 3, where we achieve low and high content redundancy by setting $w_r = 0.05$ and $w_r = -0.05$, respectively. We can observe that: (1) The performance is further improved when demonstrations have low content redundancy. (2) The performance decreases when the content of demonstrations is relatively redundant. These observations suggest that reducing content redundancy is critical for in-context compositional generalization.

4.6 Analysis of Structural Matching

In this section, we analyze how information about different modalities affects the improvements of structural matching in details.

Structural Coverage. Similar to content coverage, structural coverage are affected by $t_{l,l}$, $t_{v,v}$, $t_{v,l}$ and $t_{l,v}$. The experimental results under different hyperparameter settings are listed in Table 5. We observe that LVLMs achieve best performance when satisfying $t_{l,l} = 1$ and $t_{l,v} = 1$ simultaneously. Moreover, structural coverage provides a clear gain over content matching, which suggests that structural coverage among demonstrations is essential for ICCG besides content matching.

Hyperparameters				Accuracy	
$t_{l,l}$	$t_{v,v}$	$t_{v,l}$	$t_{l,v}$	*Cover*	*Div*
-	-	-	-	45.78	45.78
✓	-	-	-	48.30	49.33
-	✓	-	-	44.03	45.94
-	-	✓	-	45.22	46.49
-	-	-	✓	45.82	46.31
✓	✓	-	-	46.36	48.05
✓	-	✓	-	48.12	**51.15**
✓	-	-	✓	**48.55**	48.80
✓	✓	✓	-	46.56	48.55
✓	✓	✓	✓	46.82	48.55

Table 5: Accuracy (%) under different structural coverage and structural diversity settings on GQA-ICCG. We use OF-4B with content matching at 8-shot as baseline model, whose performance is shown in the first line.

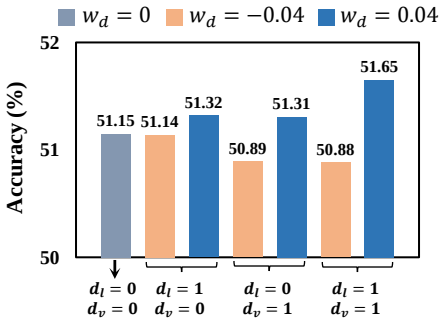

Figure 4: Accuracy (%) of OF-4B under different structural complexity settings at 8-shot on GQA-ICCG.

Structural Diversity. Structural diversity is affected by $t_{l,l}$, $t_{v,v}$, $t_{v,l}$ and $t_{l,v}$. We validate their effectiveness by ablating certain hyperparameters, and experimental results are listed in Table 5. We observe different performance gains over structural coverage when setting different values to the four hyperparameters, *e.g.*, 1.91% and 0.49% relative performance gains when $t_{v,v} = 1$ and $t_{l,v} = 1$, respectively. And the improvement is more significant when $t_{l,l} = 1$ and $t_{v,l} = 1$ simultaneously.

Structural Complexity. The structural complexity is calculated by Equation (9), where d_l and d_v decide the structural complexity of which modalities are considered during demonstration selection. To validate the effect of structural complexity on ICCG, we set $w_d = 0.04$ and $w_d = -0.04$ in Equation (6) to simulate low and high structural complexity, respectively. Experimental results are shown in Figure 4, which demonstrate that: (1) Low structural complexity of in-context demonstra-

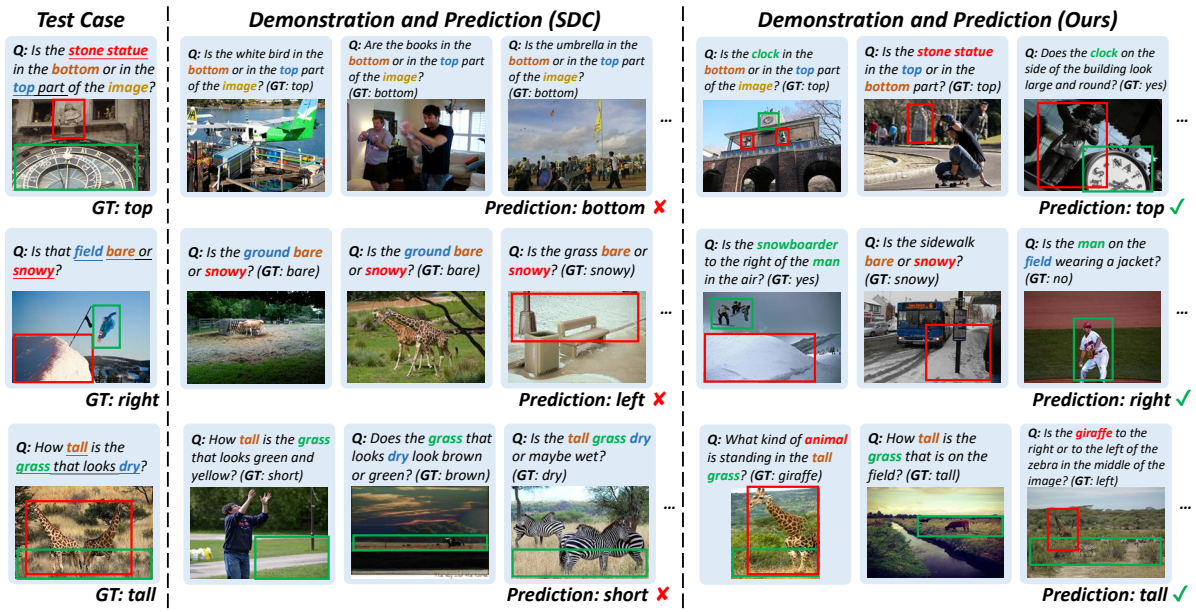

Figure 5: Qualitative comparisons between SDC and our method on GQA-ICCG, where we use OF-4B at 8-shot as baseline model. The same colored words and image regions in an example indicate that they are similar information. Underlined words in the test cases denote novel compositions that were not present in candidate demonstrations.

tions is more helpful to achieve in-context compositional generalization. (2) Structural complexity of both visual and linguistic modalities should be considered during demonstration selection.

4.7 Qualitative Analysis

We provide several qualitative examples to further compare with SDC (An et al., 2023) in Figure 5, where the selected in-context demonstrations with top-3 matching scores for each test case are visualized. We obverse that: (1) The performance of LVLMs is sensitive to demonstrations. (2) The model with demonstrations of our method makes predictions accurately. In the first example, the test case is about a "stone statue", and the image mainly consists of a "stone statue" and a "clock". For this test case, the demonstrations of SDC are only relevant to the test case on linguistic modality, but the visual content is completely irrelevant. Differently, the demonstrations of our method are relevant to the test case on both visual and linguistic modalities while having less redundant information. Therefore, our method can provide more suitable contextual information for in-context learning. More examples are provided in the **appendix**.

5 Conclusion

This work has presented a demonstration selection method to select in-context demonstrations, which can improve the ICCG ability of LVLMs. A diversity-coverage-based matching score is designed to reduce the impact of redundant visual information arising from the inherent asymmetry between visual and linguistic modalities. We have constructed a GQA-ICCG dataset to enable the quantitative evaluation for the ICCG ability of LVLMs. Experimental results demonstrate that content coverage/diversity/redundancy and structural coverage/diversity/complexity play an important role in demonstration selection.

6 Limitations

In our implementation, the demonstration selection method is separate from the LVLMs, which may affect the performance considering the differences in training data, architecture design, and abilities of LVLMs. In the future, we will consider incorporating the six factors into a trainable demonstration selection module, to train the module and the LVLMs in an end-to-end manner.

Acknowledgments This work was supported by the Natural Science Foundation of China (NSFC) under Grants No. 62176021 and No. 62172041, Natural Science Foundation of Shenzhen under Grant No. CYJ20230807142703006, Key Research Platforms and Projects of the Guangdong Provincial Department of Education under Grant No. 2023ZDZX1034.

References

Arjun Akula, Varun Jampani, Soravit Changpinyo, and Song-Chun Zhu. 2021. Robust visual reasoning via language guided neural module networks. pages 11041–11053.

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

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? pages 11027–11052.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023a. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023b. Openflamingo.

Ben Bogin, Sanjay Subramanian, Matt Gardner, and Jonathan Berant. 2021. Latent compositional representations improve systematic generalization in grounded question answering. 9:195–210.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. 33:1877–1901.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. 2023. *URL https://lmsys. org/blog/2023-03-30-vicuna*, 1(2):3.

Andrew Drozdov, Nathanael Schärli, Ekin Akyürek, Nathan Scales, Xinying Song, Xinyun Chen, Olivier Bousquet, and Denny Zhou. 2022. Compositional semantic parsing with large language models.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. pages 6904–6913.

Arian Hosseini, Ankit Vani, Dzmitry Bahdanau, Alessandro Sordoni, and Aaron Courville. 2022. On the compositional generalization gap of in-context learning. *arXiv preprint arXiv:2211.08473*.

Drew A Hudson and Christopher D Manning. 2018. Compositional attention networks for machine reasoning.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. pages 6700–6709.

Nikita Kitaev, Steven Cao, and Dan Klein. 2019. Multilingual constituency parsing with self-attention and pre-training. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3499–3505, Florence, Italy. Association for Computational Linguistics.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. 2024. Obelics: An open web-scale filtered dataset of interleaved image-text documents. 36.

Itay Levy, Ben Bogin, and Jonathan Berant. 2022. Diverse demonstrations improve in-context compositional generalization. *arXiv preprint arXiv:2212.06800*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Chuanhao Li, Zhen Li, Chenchen Jing, Yunde Jia, and Yuwei Wu. 2023b. Exploring the effect of primitives for compositional generalization in vision-and-language. pages 19092–19101.

Juncheng Li, Junlin Xie, Long Qian, Linchao Zhu, Siliang Tang, Fei Wu, Yi Yang, Yueting Zhuang, and Xin Eric Wang. 2022. Compositional temporal grounding with structured variational cross-graph correspondence learning. pages 3032–3041.

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

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. 2023d. How to configure good in-context sequence for visual question answering. *arXiv preprint arXiv:2312.01571*.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023e. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114.

Inbar Oren, Jonathan Herzig, and Jonathan Berant. 2021. Finding needles in a haystack: Sampling structurally-diverse training sets from synthetic data for compositional generalization. pages 10793–10809.

Panupong Pasupat, Yuan Zhang, and Kelvin Guu. 2021. Controllable semantic parsing via retrieval augmentation. pages 7683–7698.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Gabriel Poesia, Oleksandr Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models.

Linlu Qiu, Peter Shaw, Panupong Pasupat, Tianze Shi, Jonathan Herzig, Emily Pitler, Fei Sha, and Kristina Toutanova. 2022. Evaluating the impact of model scale for compositional generalization in semantic parsing. pages 9157–9179.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. pages 8748–8763.

Amir Rahimi, Vanessa D'Amario, Moyuru Yamada, Kentaro Takemoto, Tomotake Sasaki, and Xavier Boix. 2023. D3: Data diversity design for systematic generalization in visual question answering. *arXiv preprint arXiv:2309.08798*.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671.

Raeid Saqur and Karthik Narasimhan. 2020. Multimodal graph networks for compositional generalization in visual question answering. pages 3070–3081.

Jiaxin Shi, Hanwang Zhang, and Juanzi Li. 2019. Explainable and explicit visual reasoning over scene graphs. pages 8376–8384.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1423–1436.

Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, and Jun Liu. 2023. Meta compositional referring expression segmentation. pages 19478–19487.

Moyuru Yamada, Vanessa D'Amario, Kentaro Takemoto, Xavier Boix, and Tomotake Sasaki. 2022. Transformer module networks for systematic generalization in visual question answering. *arXiv preprint arXiv:2201.11316*.

Lijin Yang, Quan Kong, Hsuan-Kung Yang, Wadim Kehl, Yoichi Sato, and Norimasa Kobori. 2023. Deco: Decomposition and reconstruction for compositional temporal grounding via coarse-to-fine contrastive ranking. pages 23130–23140.

Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. 2024. Exploring diverse in-context configurations for image captioning. 36.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Ves Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2022. Complementary explanations for effective in-context learning. *arXiv preprint arXiv:2211.13892*.

Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. 2022a. Magic: Multimodal relational graph adversarial inference for diverse and

unpaired text-based image captioning. volume 36, pages 3335–3343.

Wenqiao Zhang, Haochen Shi, Siliang Tang, Jun Xiao, Qiang Yu, and Yueting Zhuang. 2021. Consensus graph representation learning for better grounded image captioning. volume 35, pages 3394–3402.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning. pages 9134–9148.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. pages 12697–12706.

Yang Zhou, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Prompting vision language model with knowledge from large language model for knowledge-based vqa. *arXiv preprint arXiv:2308.15851*.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. 2024. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*.

A Additional Ablation Studies

We provide ablation studies of OpenFlamingo-4B (Awadalla et al., 2023b) on our GQA-ICCG dataset at 4-shot paradigm. The experimental results are listed in Table 6. From the table, we can obtain an observation similar to the ablation studies performed at 8-shot paradigm in the manuscript: all six factors including content coverage, content diversity, content redundancy, structural coverage, structural diversity, and structural complexity have a certain impact on improving the in-context compositional generalization ability of LVLMs. The observation further demonstrates the validity of our demonstration selection method, where factors are complementary to improve in-context compositional generalization.

Content Matching			Structural Matching			Accuracy
Cover	*Div*	*Red*	*Cover*	*Div*	*Comp*	
-	-	-	-	-	-	38.89
✓	-	-	-	-	-	44.43
✓	✓	-	-	-	-	46.49
✓	✓	✓	-	-	-	47.00
✓	✓	✓	✓	-	-	47.96
✓	✓	✓	✓	✓	-	49.67
✓	✓	✓	✓	✓	✓	**49.72**

Table 6: Ablation studies on GQA-ICCG. We use OF-4B at the 4-shot as the baseline model, whose performance is shown in the first line.

B Additional Evaluation of ICCG Performance

The experimental results of more settings of k on the proposed GQA-ICCG dataset are listed in Table 7. From the table, we can observe that our method performs best among different few-shot paradigms compared to state-of-the-art methods. For instance, take Otter-7B (Li et al., 2023a) as baseline for comparisons, our method achieves 51.04% and 50.71% at 8-shot and 16-shot, which has 1.15% and 1.42% relative improvements over SDC (An et al., 2023). Another observation is that the improvement on Otter-7B at 4-shot is limited, probably due to the insufficient contextual information when the demonstration number is small. These observations show that: 1) The in-context compositional generalization ability of large vision-language models (LVLMs) is sensitive to the number of demonstrations. 2) Our demonstration selection method is effective in improving the in-context compositional generalization ability of different LVLMs at different few-shot paradigms.

C Additional Qualitative Examples

We visualize several qualitative examples from the GQA-ICCG dataset at the 8-shots setting in Figure 6. In the figure, we use Otter-7B as baseline model, and compare our demonstration selection method with SDC. Although the test cases have different question types (*e.g.*, four test cases ask about "color", "material", "position", and "height", respectively), the baseline model using our method can make correct answers for all these test cases, since our method provides more diverse contextual information that are similar to test cases but not redundant. For example, for the first qualitative example, the test case contains "sofa" and "red or gray" in the question, the demonstrations selected by SDC only contains "red or gray", but our method can find more useful demonstrations that contain both "sofa" and "red or gray". Moreover, the demonstrations selected by our method can cover more primitives in the image of the test case, thus help Otter understand the vision content and answer the question correctly.

Model	Method	k-Shot				
		$k=0$	$k=4$	$k=8$	$k=12$	$k=16$
OF-3B (Awadalla et al., 2023b)	Random	35.77	37.57	37.16	36.13	34.74
	Similarity		42.42	42.97	42.08	41.77
	SDC (An et al., 2023)		45.54	45.62	44.48	44.59
	Ours		**46.87**(+1.33)	**47.54**(+1.92)	**46.48**(+2.00)	**44.66**(+0.07)
OF-4B (Awadalla et al., 2023b)	Random	28.29	38.89	38.52	37.19	40.95
	Similarity		43.85	43.34	43.41	43.87
	SDC (An et al., 2023)		46.25	46.37	45.34	45.57
	Ours		**49.72**(+3.47)	**51.65**(+5.28)	**48.89**(+3.55)	**48.22**(+2.65)
Otter-7B (Li et al., 2023a)	Random	32.67	37.43	37.96	39.06	39.43
	Similarity		45.16	46.71	47.88	47.73
	SDC (An et al., 2023)		48.94	49.89	49.73	49.63
	Ours		**49.01**(+0.07)	**51.04**(+1.15)	**50.71**(+0.98)	**51.05**(+1.42)

Table 7: Accuracy (%) of the state-of-the-art methods on GQA-ICCG. The words in red font indicate the relative improvement of our method compared with SDC.

Figure 6: Qualitative comparisons between SDC and our method on GQA-ICCG. We use Otter-7B at 8-shot as baseline model. The same colored words and image regions in an example indicate that they are similar information. Underlined words in the test cases denote novel compositions that were not present in candidate demonstrations.