Mimic In-Context Learning for Multimodal Tasks

Yuchu Jiang[1,2] kamichanw@seu.edu.cn
Jiale Fu[1,2] jiale.fu@seu.edu.cn
Chenduo Hao[1,2] 213201447@seu.edu.cn
Xinting Hu [3] xinting001@e.ntu.edu.sg
Yingzhe Peng[1,2] yingzhe.peng@seu.edu.cn
Xin Geng[1,2] xgeng@seu.edu.cn
Xu Yang[1,2*] xuyang_palm@seu.edu.cn

[1]Southeast University
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, (Southeast University), Ministry of Education
[3]Nanyang Technological University

Abstract

*Recently, In-context Learning (ICL) has become a significant inference paradigm in Large Multimodal Models (LMMs), utilizing a few in-context demonstrations (ICDs) to prompt LMMs for new tasks. However, the synergistic effects in multimodal data increase the sensitivity of ICL performance to the configurations of ICDs, stimulating the need for a more stable and general mapping function. Mathematically, in Transformer-based models, ICDs act as "shift vectors" added to the hidden states of query tokens. Inspired by this, we introduce **Mim**ic **I**n-**C**ontext Learning (**MimIC**) to learn stable and generalizable shift effects from ICDs. Specifically, compared with some previous shift vector-based methods, MimIC more strictly approximates the shift effects by integrating lightweight learnable modules into LMMs with four key enhancements: 1) inserting shift vectors after attention layers, 2) assigning a shift vector to each attention head, 3) making shift magnitude query-dependent, and 4) employing a layer-wise alignment loss. Extensive experiments on two LMMs (Idefics-9b and Idefics2-8b-base) across three multimodal tasks (VQAv2, OK-VQA, Captioning) demonstrate that MimIC outperforms existing shift vector-based methods. The code is available at https://github.com/Kamichanw/MimIC.*

1. Introduction

In-Context Learning (ICL) allows models to generalize from a few examples, known as in-context demonstrations (ICDs), enabling them to learn new tasks without explicit fine-tuning [3, 7, 29, 37, 55]. This approach has become a significant inference paradigm for both Large Language Models (LLMs) and Large Multimodal Models (LMMs) [36] and finds broad applications in areas such as recommendation systems [11, 40, 54] and point cloud understanding [16]. However, ICL in LMMs faces more limitations than LLMs due to the synergistic effects of integrating vision and language data [23, 53], making some strategies useful in LLM lose their efficacy. For example, while various studies in LLM show that using similar ICDs as the query is beneficial [26, 43], [50] found that in captioning tasks, using less similar images may actually improve performance when only low-quality in-context captions are available. This is because similar images can lead LMMs to copy captions through shortcut inference, rather than generalizing to new examples.

Due to synergistic effects, when implementing ICL, it is hard for LMMs to capture the general mapping from input-output pairs of complex multimodal ICDs as LLMs. Instead, [23] shows that LMMs tend to rely on the distribution of ICDs to narrow the prediction space, *e.g.*, in Visual Question Answering (VQA), a LMM might recognize the ICD answer format and respond an answer in the same format, rather than learning the correct function as in language QA. Consequently, the ICL performance in LMMs is more sensitive than in LLMs to ICD configurations [28, 58] and finding optimal ICD configurations in LMM is still an open question [7]. A straightforward approach to mitigate the high sensitivity issue is to use more ICDs to help LMMs recognize stable patterns for improved predictions. However, image inputs require more tokens than text, and increasing the number of ICDs significantly raises computational demands. Moreover, current LMMs, like open-source 8B models, typically support only up to 32-shot ICDs [18],

*Corresponding Author

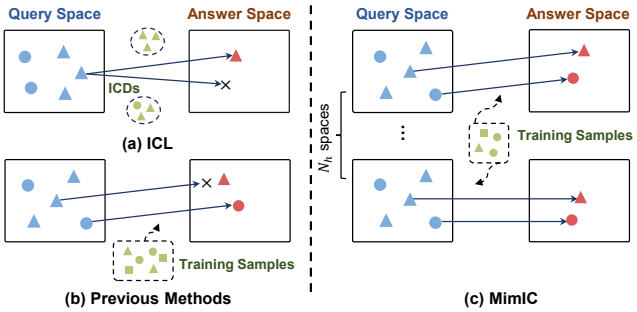

Figure 1. Sketches of shift effects from query space to answer space. (a) Traditional ICL induces the shift vector by ICDs, which is sensitive to ICD configurations, *i.e.*, changing one ICD make prediction incorrect. (b) Previous shift vector-based methods insert a query-independent shift vector learned from a large training set, causing equal shift magnitude for diverse queries, which may make prediction incorrect. (c) MimIC assigns a unique query-dependent shift vector learned from fewer training samples after each attention head layer, shifting diverse magnitude for different queries, thus achieving stronger generalization ability.

rendering this approach impractical at larger scales.

In this way, we may wonder whether it is possible to learn a general mapping function from ICDs and then directly incorporate this function into LMMs to enhance ICL performance. Interestingly, from a mathematical perspective, the role of ICDs can be seen as adding shift vectors to the hidden representations of query tokens in LLMs/LMMs [13, 27, 35, 44]. Motivated by this, researchers propose to find a general shift vector as a general mapping function to transform the query space into the answer space. The early methods [13, 27, 44] used heuristic-based approaches to generate shift vectors, which were effective in simple NLP tasks but proved insufficient for more complex multimodal tasks. To address this, a recent study, LIVE [35], introduces a training-based method to learn the shift vector from a large supporting set, outperforming these heuristic-based methods.

Although these mathematically inspired methods improve ICL performance, closer examination in Sec. 3.1 reveals that they use incomplete approximations. First, the formula suggests that the shift vector should be applied after the attention layers, but current methods incorrectly place it after the feed-forward network (FFN) layers. This misuse leads to a second issue: since Transformers use multi-head attention, where each head may have a distinct representation space, applying the shift vector after the FFN overlooks the need for separate shifts for each attention head, reducing the effectiveness of the mapping. Third, the formula implies that the shift magnitude should depend on the query, but these methods use a query-independent shift magnitude. Consequently, as shown in Fig. 1 (b), during inference, the fixed shift magnitude can lead to poor predictions for diverse queries.

In this work, we propose to approximate the shift effect more rigorously to better **Mim**ic **I**n-**C**ontext Learning (**MimIC**), offering four key improvements for better adaptability and efficiency. First, we position the shift vector after the attention layers instead of the FFN layers. This change allows each attention head to learn a unique shift vector, capturing distinct representation shifts as illustrated in Fig. 1 (c), which is our second enhancement. Third, we make the scaling factor of the shift query-dependent, enabling dynamic adjustment of the shift magnitude during inference. Finally, we implement a layer-wise alignment loss to ensure that hidden states in zero-shot setting closely align with those of ICL, which allows our method to achieve ICL-like performance with minimal training data.

We validate the effectiveness of MimIC on three foundational multimodal tasks: VQAv2, OK-VQA and Image Captioning (IC) and on two prominent open-source LMMs: Idefics1 [18] and Idefics2 [19], which represent cross-attention and fully autoregressive architectures, respectively. The results show MimIC surpasses standard ICL, *e.g.*, it achieves a 3.46% accuracy/3.57% accuracy/9.00 CIDEr improvement than 32-shot ICL on VQAv2/OKVQA/IC on Idefics1. Moreover, MimIC's generalization allows it to match 32-shot ICL performance with only 1-shot ICL guidance. Compared to previous methods, MimIC achieves superior results, *e.g.*, it improves 4.04% accuracy /4.99% accuracy/2.13 CIDEr than the second-best method on VQAv2/OKVQA/IC. Furthermore, comprehensive ablation studies and analyzes confirm the effectiveness of our four proposed enhancements, showing that MimIC requires fewer training samples and achieves a better approximation to ICL compared to other trainable methods.

In summary, we have the following contributions:

- Mathematically, we show the flaws of the approximations in previous shift vector-based methods and propose a more rigorous approximation, offering a stronger mathematical assurance for implementation.
- Guided by the mathematical formula, we propose a feasible method that achieves approximation by adding fewer learnable parameters.
- The results of the experiment show that MimIC achieves consistent improvements compared to the original ICL, previous shift vector-based methods, and LoRA in three multimodal tasks on two LMMs.

2. Related Work

2.1. In-Context Learning.

In-Context learning (ICL) refers to a model's ability to perform new tasks by conditioning on a sequence of input-output examples without requiring updates to its parameters [3, 7]. This mechanism, widely adopted in Large Lan-

guage Models (LLMs), allows models to enhance downstream task performance with minimal labeled data [33, 46]. However, practical applications of ICL face two prominent challenges. First, its performance is highly sensitive to the selection [9, 23, 24, 39, 52, 58] and ordering [17, 28, 47] of in-context demonstrations (ICDs). Methods that select and utilize high-quality ICDs, such as similarity-based retrieval [4, 12, 26, 43, 51], are often computationally expensive and not scalable in data-scarce scenarios. Second, an excessive number of demonstrations can result in long context windows, which significantly slows down inference [48].

2.2. ICL in Large Multimodal Models.

In the realm of multimodal models, several approaches have incorporated ICL capabilities by training on interleaved image-text datasets [1, 18, 19, 21, 22, 42]. However, integrating ICL into multimodal models introduces unique challenges that are often underexplored [59]. Leveraging the inference strengths of LLMs, Large Multimodal Models (LMMs) such as Idefics [18] and Idefics2 [19] exhibit ICL capabilities by using multiple samples as contextual information during training. Nonetheless, the inherent complexity and diversity of multimodal tasks exacerbate existing challenges in ICL, making it more difficult to fully harness the potential of multimodal ICL [2, 23, 30, 49, 50].

A recent study by [34] shows that for some LMMs, similarity-based retrieval methods for selecting ICDs can perform worse than random selection. This highlights the difficulty of identifying high-quality ICDs in multimodal tasks, which remains an open problem. Moreover, current mainstream multimodal architectures, such as those used in [19, 25, 45], typically concatenate image tokens directly with text tokens. Since one image can be equivalently encoded to thousands of text tokens, incorporating multiple image tokens significantly increases the context length, leading to substantial slowdowns in ICL inference.

2.3. Understanding ICL Mechanisms.

Understanding the underlying mechanisms of ICL is critical for improving its effectiveness in guiding the inference processes of LLMs. Various approaches have been proposed to explain why ICL works. For instance, [32] suggests that ICL performance depends not only on the accuracy of true labels but also on factors such as label space representation, input distribution, and sequence format. Additionally, [6] argues that Transformer attention mechanisms in ICL operate similarly to gradient descent, framing the process as implicit fine-tuning. Another perspective comes from the concept of a "task vector" [13], which posits that ICL compresses training data into a single vector that guides the model's outputs. Similarly, the "Function Vector" [44] idea identifies a compact neural encoding of input-output functions within autoregressive language models.

A prevailing approach interprets ICL through the framework of shift vector, where ICL is understood as encoding task context within a learnable vector representation that modulates model behavior. For example, LIVE [35] employs a self-distillation strategy to optimize a learnable ICV directly, which enhances ICL's performance in multimodal settings. Similarly, the Multimodal Task Vector (MTV) method [15] averages activations across multiple attention heads, encoding task-specific information as task vectors to enable robust few-shot multimodal ICL. Our work aligns with this methodology but introduces a unique and innovative approach. Unlike previous methods, MimIC achieves superior data efficiency and enhanced learning capability.

3. Mimicking In-Context Learning

Our objective is to mimic in-context learning (ICL) from the perspective of space shift induced by ICL with fewer trainable parameters and training samples. We begin by analyzing the behavior of in-context demonstrations (ICDs) within the self-attention mechanism in Sec. 3.1. This analysis reveals that the output of self-attention can be decomposed into two components: one affected by the ICDs and the other independent of them. We then detail how to approximate the component influenced by the ICDs to capture the general shift effect in Sec. 3.2.

3.1. Mathematic Analyses

ICL allows large language models (LLMs) or large multimodal models (LMMs) to generalize to new tasks by providing a few ICDs directly in the input. Formally, the prompt context is defined as $\boldsymbol{C} = \{\boldsymbol{X}_D, \boldsymbol{X}\}$, where $\boldsymbol{X}_D = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k\} \in \mathbb{R}^{l_D \times d}$ represents the concatenation of k ICDs, and $\boldsymbol{X} \in \mathbb{R}^{l_q \times d}$ is the query input. Here, l_D and l_q denote the number of tokens in $\boldsymbol{X}_D$ and $\boldsymbol{X}$, respectively, and d is the embedding dimension.

Multi-head self-attention applies the self-attention (SA) mechanism over N_h heads, each parameterized by weight matrices $\boldsymbol{W}_k, \boldsymbol{W}_q, \boldsymbol{W}_v \in \mathbb{R}^{d \times d_h}$ to project $\boldsymbol{C}$ into keys $\boldsymbol{K}_C$, queries $\boldsymbol{Q}_C$, and values $\boldsymbol{V}_C$. Typically, d_h is set to d/N_h to reduce parameter usage by operating each attention head in a lower-dimensional space. For a specific head, the key mapping is defined as:

$$\boldsymbol{K}_C = \boldsymbol{C}\boldsymbol{W}_k = \begin{bmatrix} \boldsymbol{X}_D \\ \boldsymbol{X} \end{bmatrix} \boldsymbol{W}_k = \begin{bmatrix} \boldsymbol{K}_D \\ \boldsymbol{K} \end{bmatrix}. \tag{1}$$

Similarly, we compute the corresponding $\boldsymbol{Q}_D, \boldsymbol{Q}$, and $\boldsymbol{V}_D, \boldsymbol{V}$ using $\boldsymbol{W}_q$ and $\boldsymbol{W}_v$, respectively. For each query

vector $\boldsymbol{q} \in \boldsymbol{Q}$, the single-head self-attention operation is[1] :

$$
\begin{aligned}
&\mathrm{SA}\left(\boldsymbol{q}, \begin{bmatrix}\boldsymbol{K}_D \\ \boldsymbol{K}\end{bmatrix}, \begin{bmatrix}\boldsymbol{V}_D \\ \boldsymbol{V}\end{bmatrix}\right) \\
&= \mathrm{softmax}\left(\left[\boldsymbol{q}\boldsymbol{K}_D^\top, \boldsymbol{q}\boldsymbol{K}^\top\right]\right)\begin{bmatrix}\boldsymbol{V}_D \\ \boldsymbol{V}\end{bmatrix} \\
&= \left[\frac{\exp\left(\boldsymbol{q}\boldsymbol{K}_D^\top\right)}{Z_1+Z_2}, \frac{\exp\left(\boldsymbol{q}\boldsymbol{K}^\top\right)}{Z_1+Z_2}\right]\begin{bmatrix}\boldsymbol{V}_D \\ \boldsymbol{V}\end{bmatrix} \\
&= \frac{Z_2}{Z_1+Z_2}\frac{\exp(\boldsymbol{q}\boldsymbol{K}^\top)}{Z_2}\boldsymbol{V} + \frac{Z_1}{Z_1+Z_2}\frac{\exp(\boldsymbol{q}\boldsymbol{K}_D^\top)}{Z_1}\boldsymbol{V}_D \\
&= \frac{Z_2}{Z_1+Z_2}\mathrm{softmax}(\boldsymbol{q}\boldsymbol{K}^\top)\boldsymbol{V} + \frac{Z_1}{Z_1+Z_2}\exp(\boldsymbol{q}\boldsymbol{K}_D^\top)\boldsymbol{V}_D \\
&= (1-\mu)\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) + \mu\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D) \qquad (2)
\end{aligned}
$$

where μ is a scalar representing the normalized attention weights over the ICDs:

$$\mu(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{K}) = \frac{Z_1(\boldsymbol{q}, \boldsymbol{K}_D)}{Z_1(\boldsymbol{q}, \boldsymbol{K}_D) + Z_2(\boldsymbol{q}, \boldsymbol{K})}, \qquad (3)$$

where $Z_1(\boldsymbol{q}, \boldsymbol{K}_D) = \sum_{i=1}^{l_D}\exp(\boldsymbol{q}\boldsymbol{K}_D^\top)_i$ and $Z_2(\boldsymbol{q}, \boldsymbol{K}) = \sum_{j=1}^{l_q}\exp(\boldsymbol{q}\boldsymbol{K}^\top)_j$.

Eq. (2) shows that the self-attention over the prompt context $\boldsymbol{C}$ can be decomposed into two terms. For the former "standard attention", it is the self-attention over the query tokens, which is independent of the ICDs. While for the latter "shift vector", it is the shift effects caused by the ICDs to shift the query space into the answer space, and such effects is calculated as the attention between the ICDs and the query $\boldsymbol{q}$. This shift is governed by the attention difference term $\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D) - \mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V})$ and the scalar $\mu(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{K})$, both of which depend on the ICDs.

Now, we show which terms in Eq. (2) are affected by ICDs and the fluctuation of ICDs changes these terms, making the predictions sensitive to the ICD configurations. Moreover, directly applying self-attention in LLMs or LMMs over long ICD inputs costs substantial computation burdens. In the next section, we describe how to approximate the ICD-affected terms in Eq. (2) to capture the general shift effect, improving robustness in ICL and significantly increasing inference efficiency.

3.2. Mimicking ICD Affected Terms

From Eq. (2) and Eq. (3), we observe that only $\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D)$ and $Z_1(\boldsymbol{q}, \boldsymbol{K}_D)$ are affected by ICDs. To mimic ICL, we approximate these terms by inserting a few lightweight modules into the attention heads of LMM and name them as **MimIC Attention Heads** as in Fig. 2(a).

To approximate $Z_1(\boldsymbol{q}, \boldsymbol{K}_D)$, we note that it is a positive scalar dependent solely on the current query token $\boldsymbol{q}$ and the ICD keys $\boldsymbol{K}_D$. Therefore, we use a simple mapping: a trainable linear layer $f(\cdot) : \mathbb{R}^{d_h} \to \mathbb{R}$ to approximate $\log Z_1$.ntion difference term $\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D) - \mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V})$, as Fig. 2(a) shows, we insert a learnable vector $\boldsymbol{v} \in \mathbb{R}^{d_h}$ in each attention head to capture the general shift effect for this head. Then, the output of Eq. (2) in MimIC attention head is computed as $\mathrm{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V}) + \tilde{\mu}(\boldsymbol{q}, \boldsymbol{K})\boldsymbol{v}$, where $\tilde{\mu}(\boldsymbol{q}, \boldsymbol{K}) = \tilde{Z}_1(\boldsymbol{q})/(\tilde{Z}_1(\boldsymbol{q}) + Z_2(\boldsymbol{q}, \boldsymbol{K}))$ and $\tilde{Z}_1(\boldsymbol{q}) = \exp(f(\boldsymbol{q}))$. After obtaining the outputs from all MimIC attention heads, they are concatenated, flattened, and passed through the matrix $\boldsymbol{W}_o \in \mathbb{R}^{d\times d}$ and FFN layer.

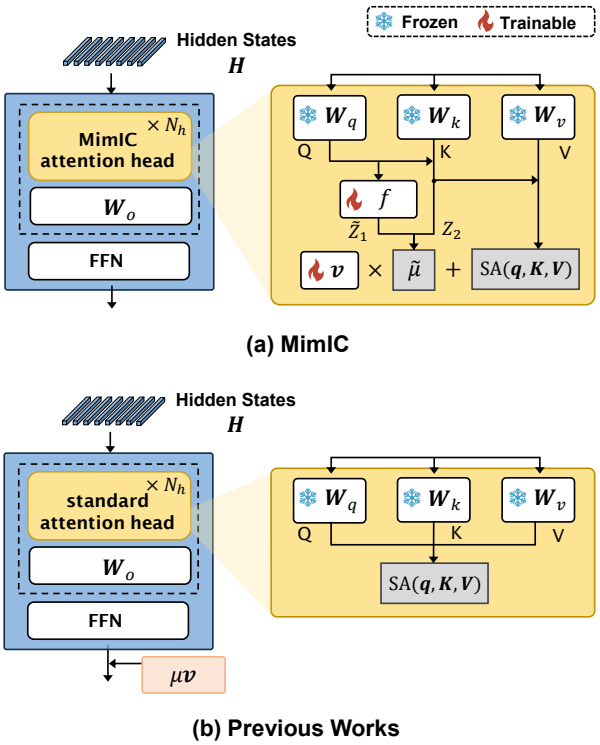

Figure 2. Comparison of MimIC and previous shift vector based methods. (a) MimIC changes the attention mechanism for each head, which inserts a learnable shift vector $\boldsymbol{v}$ with a query-dependent magnitude μ. (b) Previous methods insert the pre-calculated or learnable shift vector with a query-independent μ after FFN layer without changing the attention mechanism.

Given this MimIC attention head, we replaces all the self-attention heads of the original LMM to get the **MimIC LMM** as in Fig. 3(b). Then we hope Mimic LLM can handle a single query $\boldsymbol{X}$ in the same way the original LMM implements ICL, *i.e.*, using the ICDs $\boldsymbol{X}_D$ to produce the result for $\boldsymbol{X}$. To achieve this, given a training set of n samples, we randomly select k samples as ICDs $\boldsymbol{X}_D$ and one sample as the query $\boldsymbol{X}$. As Fig. 3(a) shows, for the original LMM, we input the context $\boldsymbol{C} = \{\boldsymbol{X}_D, \boldsymbol{X}\}$ into it to get the hidden states at each layer, which are recorded as $\boldsymbol{\mathcal{H}}' = \{\boldsymbol{H}'_1, \ldots, \boldsymbol{H}'_N\}$. For MimIC LMM, we only input

[1]For simplicity, we illustrate the method using a single head, though each head typically has distinct weight matrices.

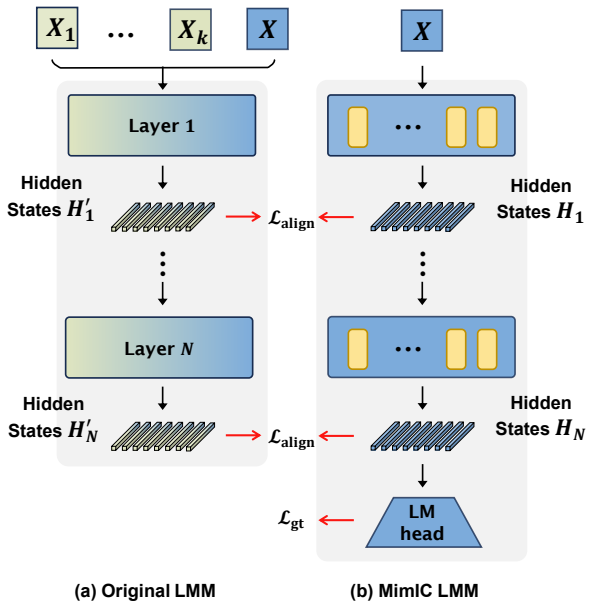

Figure 3. Overall training framework of MimIC. (a) The original LMM processes k ICDs and query input as conventional ICL, generating hidden states $\boldsymbol{H}'_1$ to $\boldsymbol{H}'_N$ at each layer. (b) In MimIC LMM, only a single query input $\boldsymbol{X}$ is processed, producing shifted hidden states $\boldsymbol{H}_1$ to $\boldsymbol{H}_N$, which are aligned with the original hidden states via the alignment loss $\mathcal{L}_{\text{align}}$. Additionally, the logits of language modeling head is used to compute ground truth loss $\mathcal{L}_{\text{gt}}$. The yellow blocks represents MimIC attention heads.

$\boldsymbol{X}$ into it to get the hidden states $\boldsymbol{\mathcal{H}} = \{\boldsymbol{H}_1, \ldots, \boldsymbol{H}_N\}$. To make MimIC LMM behave similar as the original LMM, we set an alignment loss $\mathcal{L}_{\text{align}}$ to make $\boldsymbol{\mathcal{H}}$ be close to $\boldsymbol{\mathcal{H}}'$. Specifically, $\mathcal{L}_{\text{align}}$ is computed as the average L_2 distance between the hidden states at each layer, ensuring a layer-wise contribution:

$$\mathcal{L}_{\text{align}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{l_q}\left\|\boldsymbol{h}_{i,j} - \boldsymbol{h}'_{i,j}\right\|_2^2. \tag{4}$$

In addition, we employ the language modeling loss $\mathcal{L}_{\text{gt}}$ to enhance the model's performance on downstream tasks, allowing it to learn task-specific features more effectively. Thus, the total loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{align}} + \lambda\mathcal{L}_{\text{gt}}, \tag{5}$$

where λ is a hyperparameter that controls the trade-off between alignment and task-specific loss.

During training, since the ICDs are randomly selected in each step, MimIC LMM is encouraged to capture the most general shift pattern from the fluctuated shifts brought by various random configurations of ICDs. After training, the attention difference term $\text{SA}(\boldsymbol{q}, \boldsymbol{K}_D, \boldsymbol{V}_D) - \text{SA}(\boldsymbol{q}, \boldsymbol{K}, \boldsymbol{V})$ captures the general shift direction across various ICD configurations, while $Z_1(\boldsymbol{q}, \boldsymbol{K}_D)$ adjusts the shift magnitude based on the query input $\boldsymbol{q}$. Consequently, when using MimIC to inference, ICDs are no longer required, leading to a significant improvement in inference speed.

3.3. Design Difference from Previous Methods

Fig. 2 compares the differences between MimIC and previous shift vector-based methods. First, previous methods insert the shift vector $\boldsymbol{v}$ after FFN layer, while we insert $\boldsymbol{v}$ into each attention head. In this way, each vector can learn suitable shift direction for the corresponding head representation space, leading to more powerful shift effects. Second, previous methods use query-independent shift magnitude μ, while we set μ be depended on the query, enabling dynamic adjustment of the shift magnitude to achieve better performance. Although these differences seem subtle, experiments will show that the devil is in the details. The findings presented in Sec. 4.3 highlight that MimIC's multi-head, query-dependent magnitude is essential for capturing general shift effects from distinct representation space.

4. Experiment

4.1. Setup

Models, datasets, and metrics. We evaluate MimIC on two large-scale multimodal models (LMMs), Idefics-9b [18] and Idefics2-8b-base [19], referred to as Idefics1 and Idefics2, across three datasets: VQAv2 [10], OK-VQA [31], and COCO Caption [5]. Idefics1 is based on a cross-attention architecture, while Idefics2 employs a fully autoregressive architecture. These models represent two popular architectures for vision-language models. For each dataset, we randomly select 1,000 samples for training. We follow the evaluation protocol of previous works [23, 35], using 10,000 validation samples from VQAv2 and the full validation splits for OK-VQA and COCO. We present more results on various datasets in Appendix.

Implementation details. During each training step, we randomly select 32 samples as ICDs for Idefics1 and 8 for Idefics2, with one additional distinct sample as the query input. We employ the AdamW optimizer with a learning rate of 5×10^{-3}, coupled with a cosine annealing scheduler with warmup, allocating 10% of the total steps for warmup. The value of λ in Eq. (5) is set to 0.5. All results are reported from the best-performing epoch. Additional implementation details are provided in the Appendix.

4.2. Comparison with Existing Methods

Compared methods. We compare MimIC with the following methods: (1) **In-Context Learning (ICL)** is evaluated under three settings: zero-shot, few-shot, and Retrieval-based In-Context Examples Selection (RICES). For few-shot ICL, we use 32/8-shot for Idefics1 and Idefics2, re-

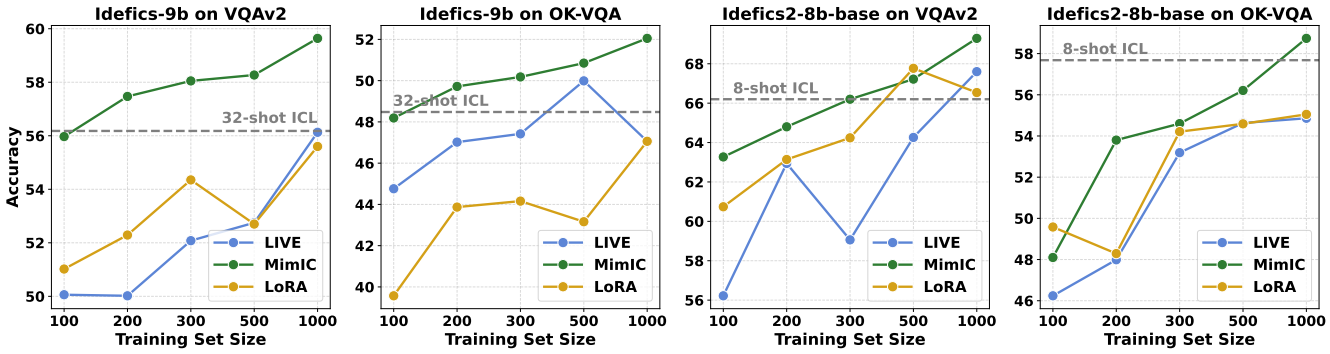

Figure 4. Performance comparisons of trainable methods on two LMMs across VQAv2/OK-VQA with fewer training set size.

Model	Method	# Params (M)	VQAv2	OK-VQA	COCO
Idefics-9b	Zero-shot	-	29.25	30.54	63.06
	32-shot ICL	-	56.18	48.48	105.89
	RICES	-	58.07	51.11	110.64
	FV	-	30.21	31.02	74.01
	TV	-	43.68	32.68	84.72
	LIVE	0.13 (×0.5)	53.71	46.05	<u>112.76</u>
	LoRA	25.0 (×96.2)	<u>55.60</u>	<u>47.06</u>	97.75
	MimIC	0.26 (×1.0)	**59.64**	**52.05**	**114.89**
Idefics2-8b-base	Zero-shot	-	55.39	43.08	40.00
	8-shot ICL	-	66.20	57.68	122.51
	RICES	-	66.44	55.73	111.44
	FV	-	36.47	34.58	75.24
	TV	-	47.12	38.27	87.61
	LIVE	0.13 (×0.5)	<u>67.60</u>	54.86	<u>126.04</u>
	LoRA	17.6 (×67.7)	66.54	<u>55.05</u>	116.69
	MimIC	0.26 (×1.0)	**69.29**	**58.74**	**132.87**

Table 1. Results of VQAv2, OK-VQA, and COCO on Idefics-9b and Idefics2-8b-base. **Bold numbers**/<u>underlined numbers</u> represent the best/second-best results, respectively.

spectively[2]. RICES [51] retrieves similar images from the support set for each query image by comparing visual features extracted from a frozen pretrained visual encoder. (2) **Task Vector (TV)** [13] and **Function Vector (FV)** [44] extract compact vectors from a set of demonstrations, which are added to the hidden states of the last token in one or more layers. We evaluate these methods across different layers and report the configuration yielding the best performance. (3) **Learnable In-Context Vector (LIVE)** [35] introduces learnable vectors after each FFN layer, trained with a pipeline similar to MimIC's under the same few-shot setting. (4) **LoRA** [14] fine-tunes the model by adding low-rank adapters to the attention weights. We apply the widely used configuration, modifying $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v$, and $\boldsymbol{W}_o$ in all attention layers of both vision and language models.

[2]We implement up to 8-shot ICL on Idefics2, as it requires more image tokens compared to Idefics1, which far exceeds our computational capacity.

All trainable methods are trained using 1000 samples. All methods are evaluated using the optimal hyper-parameters recommended in their respective original works, ensuring a fair comparison.

Results analysis. Tab. 1 presents the results of MimIC compared to various baselines across two LMMs and three datasets. In ICL, the performance of RICES significantly differs from the random selection of ICDs, indicating that the choice of ICDs has a substantial impact on ICL performance. Although RICES outperforms random selection across all three datasets on Idefics1, its performance on Idefics2, OK-VQA, and COCO is inferior to that of the random selection method. This suggests that effective ICD selection strategies differ across models.

For non-trainable methods, while they outperform zero-shot baselines on Idefics1, there is still a significant gap compared to 32-shot ICL performance. On Idefics2, these non-trainable methods fail to surpass the zero-shot baseline. This indicates that non-trainable methods are not only ineffective at capturing essential task-specific information but also perform poorly across different LMMs.

Trainable methods consistently improve performance across both LMMs, approaching the effectiveness of few-shot ICL. LIVE performances similar to LoRA with fewer parameters, while its reliance on a fixed shift magnitude during inference limits its ability to generalize across different queries, making it less effective than MimIC. However, for MimIC, on Idefics1, the performance on VQA/captioning improved by an average of 3.52/9.00 compared to 32-shot ICL, and by 4.52/2.13 compared to the second-best method. On Idefics2, MimIC was the only method to consistently outperform ICL, with average improvements of 1.31/10.36 on the VQA and captioning, respectively. Such comparisons validate the powerful of MimIC in improving the ICL performance. Also, MimIC achieves the best results on both LMMs with diverse architectures, indicating greater stability than other methods.

Training with fewer samples. We conduct further evaluations of trainable methods using a reduced number of train-

Method	VQAv2	OK-VQA	COCO
Head-sharing μ	57.89	50.86	111.98
Query-sharing μ	57.95	50.94	112.48
MimIC	**59.64**	**52.05**	**114.89**

Table 2. Performance comparisons with different settings.

ing samples. As illustrated in Fig. 4, on Idefics1, MimIC only requires 200 samples to exceed the performance of 32-shot ICL. In contrast, other trainable methods typically require a larger number of samples to achieve comparable ICL effectiveness. These comparisons suggest that using a more precise approximation of Eq. (2) requires fewer training samples to capture the general shift effect of ICL. For instance, compared to LIVE, we were able to exceed its highest performance using only about 1/8 of the data it required for training [35].

4.3. Ablations and More Analyses

We analyze the effects of various settings on Idefics1, including the necessity of employing a multi-head, query-dependent shift magnitude μ, and the impact of the diverse ICD shot numbers used as ICL guidance during training. Also, we compare the alignment distances between various methods and 32-shot ICL and the hallucinations generated by diverse methods.

Effect of multi-head and query-dependent shift magnitude μ. We compare two additional settings: (1) **Head-sharing μ**: This involves replacing the original linear layer $f : \mathbb{R}^{d_h} \rightarrow \mathbb{R}$ of each head with a new linear layer $g : \mathbb{R}^{d} \rightarrow \mathbb{R}$ that aggregates the information from N_h heads. As a result, all heads use the query-dependent shift magnitude μ produced by this linear layer. (2) **Query-sharing μ**: In this setting, the function f is removed for each head, and a learnable coefficient μ is introduced, leading to a fixed shift magnitude for each query token.

The results in Tab. 2 show that MimIC outperforms both head-sharing μ and query-sharing μ, suggesting that the multi-head and query-dependent shift magnitude not only capture features from different representation spaces but also that this dynamic behavior, which varies depending on the query, enhances generalization across diverse inputs.

Number of ICD shots. We examine the effect of varying the number of ICD shots on MimIC's performance during training. As shown in Fig. 5, there is a significant performance gap between 1-shot and 32-shot ICL. ICL performance is highly dependent on the number of demonstrations; when demonstrations are insufficient, the LMMs may misalign query representations. Although MimIC employs ICL as a guiding mechanism, its performance remains largely unaffected by different ICD configurations, demonstrating stability across varying training set sizes. This suggests that MimIC is able to learn the general shift of query representations from the demonstrations, thereby mitigating the negative impact of insufficient demonstrations during training and resulting in a more robust model capable of effectively extracting key task information.

	Zero-shot	LIVE	MimIC†	MimIC
VQAv2	42.97	33.79	32.13	**30.17**
OK-VQA	41.21	34.12	29.76	**28.25**

Table 3. L2 distance between 32-shot ICL and various methods.

	Zero-shot	32-shot ICL	TV	FV	LoRA	LIVE	MimIC
CHAIRs $\downarrow$	**5.93**	16.78	8.88	28.26	17.42	8.65	8.51
CHAIRi $\downarrow$	**5.58**	9.77	7.50	25.44	11.55	6.05	5.74
Recall $\uparrow$	30.72	42.59	36.22	27.69	42.93	42.84	**43.30**

Table 4. Caption hallucination metrics on various methods.

	4 shot		8 shot		16 shot		32 shot	
	ICL	MimIC	ICL	MimIC	ICL	MimIC	ICL	MimIC
CHAIRs $\downarrow$	5.57	**4.09**	5.41	**4.55**	5.56	**4.20**	9.77	**5.74**
CHAIRi $\downarrow$	7.20	**5.39**	7.00	**6.19**	7.70	**5.52**	16.78	**8.51**
Recall $\uparrow$	39.74	**39.84**	41.0	**41.62**	42.12	**42.69**	42.59	**43.30**

Table 5. The hallucination metrics on the image captioning task for MimIC trained with different shot numbers and the corresponding shot number of ICL.

Alignment effect in latent space. Here, we quantitatively assess whether the MimIC attention heads and $\mathcal{L}_{\text{align}}$ proposed in Sec. 3.2 facilitate alignment with ICL. Using 200 samples from VQAv2 and OK-VQA, we computed the average L2 distance of the latent representations of the first answer token at each layer, compared to the 32-shot ICL. We also test a variant, MimIC†, in which $\mathcal{L}_{\text{align}}$ is replaced with KL divergence, as used in LIVE. The results in Tab. 3 demonstrate that MimIC exhibited the smallest distance to 32-shot ICL. Specifically, MimIC is closer to 32-shot ICL than MimIC†, suggesting that $\mathcal{L}_{\text{align}}$ more effectively enables the shift vectors to capture the characteristics of the ICL shift. Additionally, MimIC† showed a smaller distance to the LIVE, highlighting that the mimic attention heads are better able to mimic ICL more precisely.

Hallucinations. Fig. 6 presents cases where MimIC responds correctly while other methods fail. We also quantitatively analyze hallucinations in image captioning using CHAIRi and CHAIRs[38], which measure the proportion of hallucinated words. As shown in Tab. 4, MimIC generates fewer hallucinations than non-zero-shot methods while maintaining a high recall rate. Compared to zero-shot, MimIC has a slight increase in hallucinations, which is due to the limitations of ICL, as noted by [41], where more shots amplify hallucinations. Despite this, MimIC

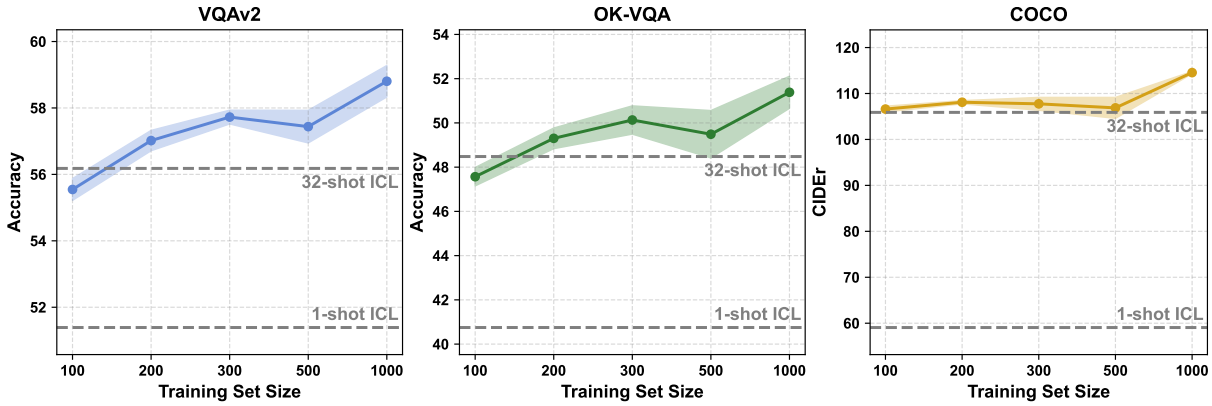

Figure 5. Performance of MimIC trained with varying ICD shots on Idefics-9b, with the shaded area indicating the standard deviation across 1, 4, 8, 16 and 32 shot settings.

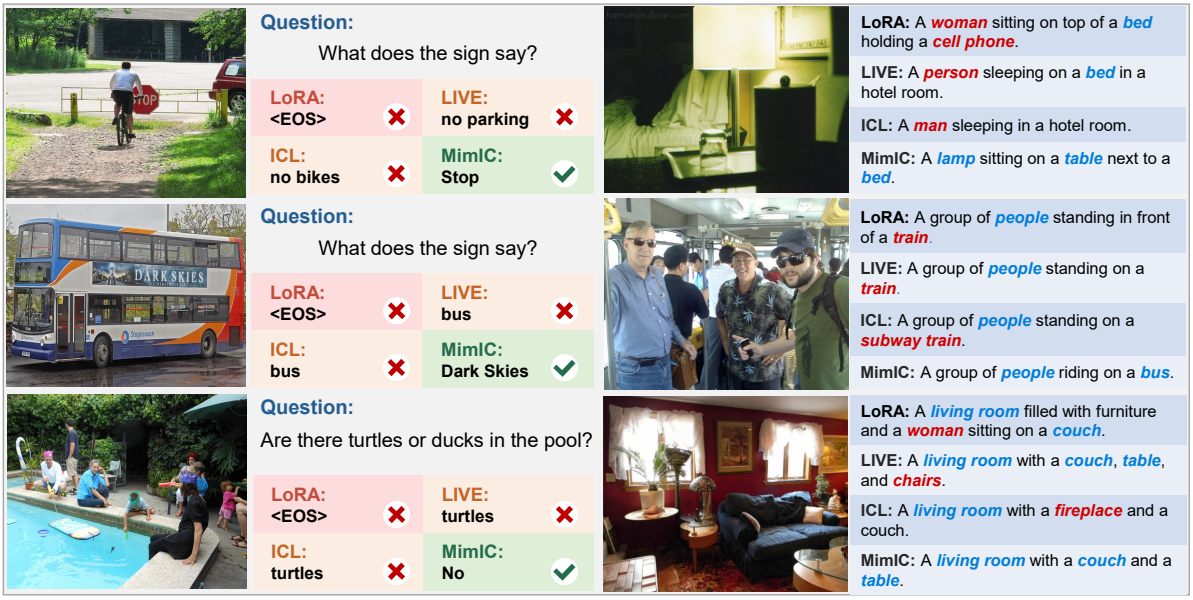

Figure 6. The visualizations of the cases where other methods appear hallucinations on visual question answering (left) and image captioning task (right). The *red* and *blue* words represent hallucination objects and correct objects, respectively.

still shows strong hallucination suppression. We also analyze hallucination levels in MimIC with varying shot counts and compare them to ICL models. The results in Tab. 5 show that hallucinations in MimIC increase with shot count but remain lower than ICL. Notably, MimIC with four shots has a lower hallucination rate than the zero-shot setting and improves recall significantly. This is due to a more precise approximation of the ICL mechanism, outperforming previous shift-based methods.

5. Conclusion

Motivated by the insight that in-context demonstrations function as shift vectors applied to the hidden states of query tokens, we propose **Mim**ic **I**n-**C**ontext Learning (**MimIC**) to learn this effect in Large Multimodal Models (LMMs). MimIC operates by inserting distinct shift vectors into different heads of attention layers, employing a linear layer to generate query-dependent magnitudes for these shift vectors, and leveraging a layer-wise alignment loss to align with ICL. Empirical evaluations across three diverse tasks using two LMMs demonstrate MimIC achieves competitive few-shot in-context learning performance with significantly reduced inference latency compared to traditional ICL methods, requires fewer training samples relative to LIVE, consistently surpasses other shift vector-based approaches, utilizes fewer parameters than LoRA yet yields superior results, and substantially reduces hallucinations compared to existing techniques.

6. Acknowledgment

This work is supported by the National Science Foundation of China (62206048), the Natural Science Foundation of Jiangsu Province (BK20220819), and the Fundamental Research Funds for the Central Universities (2242024k30035). This research work is also supported by the Big Data Computing Center of Southeast University

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1539–1550, 2024. 3

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901. Curran Associates, Inc., 2020. 1, 2

[4] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022. 3

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5

[6] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. *arXiv preprint arXiv:2212.10559*, 2022. 3

[7] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 1, 2

[8] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023. 2

[9] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online, 2021. Association for Computational Linguistics. 3

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5

[11] Wei Guo, Hao Wang, Luankang Zhang, Jin Yao Chin, Zhongzhou Liu, Kai Cheng, Qiushi Pan, Yi Quan Lee, Wanqi Xue, Tingjia Shen, et al. Scaling new frontiers: Insights into large recommendation models. *arXiv preprint arXiv:2412.00714*, 2024. 1

[12] Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, and Shuo Li. Geometric visual similarity learning in 3d medical image self-supervised pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9538–9547, 2023. 3

[13] Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*, 2023. 2, 3, 6

[14] 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. *arXiv preprint arXiv:2106.09685*, 2021. 6

[15] Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. Multimodal task vectors enable many-shot multimodal in-context learning. *arXiv preprint arXiv:2406.15334*, 2024. 3

[16] Jincen Jiang, Qianyu Zhou, Yuhang Li, Xuequan Lu, Meili Wang, Lizhuang Ma, Jian Chang, and Jian Jun Zhang. Dg-pic: Domain generalized point-in-context learning for point cloud understanding. In *European Conference on Computer Vision (ECCV)*, pages 455–474. Springer, 2024. 1

[17] Sawan Kumar and Partha Talukdar. Reordering examples helps during priming-based few-shot learning. *arXiv preprint arXiv:2106.01751*, 2021. 3

[18] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 3, 5

[19] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 2, 3, 5

[20] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308, 2024. 2

[21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[22] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 3

[23] Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26710–26720, 2024. 1, 3, 5

[24] Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. Unified demonstration retriever for in-context learning. *arXiv preprint arXiv:2305.04320*, 2023. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[26] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 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, Dublin, Ireland and Online, 2022. Association for Computational Linguistics. 1, 3

[27] Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*, 2023. 2

[28] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, 2022. Association for Computational Linguistics. 1, 3

[29] Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*, 2024. 1

[30] Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. How does the textual information affect the retrieval of multimodal in-context learning? *arXiv preprint arXiv:2404.12866*, 2024. 3

[31] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 5

[32] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022. 3

[33] Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. *arXiv preprint arXiv:2305.16938*, 2023. 3

[34] Yingzhe Peng, Xu Yang, Haoxuan Ma, Shuo Xu, Chi Zhang, Yucheng Han, and Hanwang Zhang. Icd-lm: Configuring vision-language in-context demonstrations by language modeling. *arXiv preprint arXiv:2312.10104*, 2023. 3

[35] Yingzhe Peng, Chenduo Hao, Xu Yang, Jiawei Peng, Xinting Hu, and Xin Geng. Learnable in-context vector for visual question answering. *arXiv preprint arXiv:2406.13185*, 2024. 2, 3, 5, 6, 7, 1

[36] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025. 1

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. 2019. 1

[38] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 7

[39] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 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, Seattle, United States, 2022. Association for Computational Linguistics. 3

[40] Tingjia Shen, Hao Wang, Chuhan Wu, Jin Yao Chin, Wei Guo, Yong Liu, Huifeng Guo, Defu Lian, Ruiming Tang, and Enhong Chen. Optimizing sequential recommendation models with scaling laws and approximate entropy. *arXiv preprint arXiv:2412.00430*, 2024. 1

[41] Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. Beyond task performance: Evaluating and reducing the flaws of large multimodal models with in-context learning. *arXiv preprint arXiv:2310.00647*, 2023. 7

[42] Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. Parrot: Multilingual visual instruction tuning. *arXiv preprint arXiv:2406.02539*, 2024. 3

[43] Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6292–6307, Toronto, Canada, 2023. Association for Computational Linguistics. 1, 3

[44] Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023. 2, 3, 6

[45] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Pro-*

cessing Systems, pages 24824–24837. Curran Associates, Inc., 2022. 3

[47] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 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, Toronto, Canada, 2023. Association for Computational Linguistics. 3

[48] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 3

[49] Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. *arXiv preprint arXiv:2407.00902*, 2024. 3

[50] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[51] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3081–3089, 2022. 3, 6

[52] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *International Conference on Machine Learning*, pages 39818–39833. PMLR, 2023. 3

[53] Chao Yi, Yuhang He, De-Chuan Zhan, and Han-Jia Ye. Bridge the modality and capability gaps in vision-language model selection. *Advances in Neural Information Processing Systems*, 37:34429–34452, 2024. 1

[54] Mingjia Yin, Hao Wang, Wei Guo, Yong Liu, Suojuan Zhang, Sirui Zhao, Defu Lian, and Enhong Chen. Dataset regeneration for sequential recommendation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3954–3965, 2024. 1

[55] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023. 1

[56] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2

[57] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *ArXiv*, abs/2409.02813, 2024. 3

[58] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021. 1, 3

[59] Zhen Zhao, Jingqun Tang, Chunhui Lin, Binghong Wu, Can Huang, Hao Liu, Xin Tan, Zhizhong Zhang, and Yuan Xie. Multi-modal in-context learning makes an ego-evolving scene text recognizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15567–15576, 2024. 3