Can Multimodal Large Language Models Truly Perform Multimodal In-Context Learning?

Shuo Chen[1,3] Zhen Han[1] Bailan He[1,3] Jianzhe Liu [4] Mark Buckley[3]
Yao Qin [5] Philip Torr[2] Volker Tresp[1,6] Jindong Gu[2*]
[1]LMU Munich [2]University of Oxford [3]Siemens AG
[4]Technical University of Munich [5]University of California, Santa Barbara
[6]Munich Center for Machine Learning (MCML)
chenshuo.cs@outlook.com, jindong.gu@eng.ox.ac.uk

Abstract

Large Language Models (LLMs) with in-context learning (ICL) ability can quickly adapt to a specific context given a few demonstrations (demos). Recently, Multimodal Large Language Models (MLLMs) built upon LLMs have also shown multimodal ICL ability, i.e.*, responding to queries given a few multimodal demos, including images, queries, and answers. While ICL has been extensively studied on LLMs, its research on MLLMs remains limited. One essential question is whether these MLLMs can truly conduct multimodal ICL, or if only the textual modality is necessary. We investigate this question by examining two primary factors that influence ICL: 1) Demo content,* i.e.*, understanding the influences of demo content in different modalities. 2) Demo selection strategy,* i.e.*, how to select better multimodal demos for improved performance. Experiments revealed that multimodal ICL is predominantly driven by the textual content whereas the visual information in the demos has little influence. Interestingly, visual content is still necessary and useful for selecting demos to increase performance. Motivated by our analysis, we propose a simple yet effective approach, termed Mixed Modality In-Context Example Selection (MMICES), which considers both visual and language modalities when selecting demos. Extensive experiments are conducted to support our findings and verify the improvement brought by our method.*

1. Introduction

The in-context learning (ICL) ability of large language models (LLMs) has received great attention [6, 10, 15, 25, 37]. It enables the models to quickly adapt to a specific context given a set of question-and-answer pairs, referred to as demonstrations (demos), without any model parameter updates [5, 6, 12]. Recent Multimodal Large Language Models (MLLMs), which are built upon LLMs, have also displayed multimodal in-context learning (M-ICL) ability [1,3,20,38,46,48]. These pre-trained models can rapidly adapt to vision-language tasks using few-shot multimodal demos comprising images, queries, and answers. For example, as shown in Fig. 1, two images and the corresponding questions and answers are selected as demos from an available support set. Then a pre-trained MLLM generates answers for the query based on the demos. Although ICL on LLMs has been intensively explored [2,27,28,30,41,45], the understanding of this capability within MLLMs remains largely limited. An essential question is *whether these MLLMs can truly perform multimodal ICL, or if only the textual modality is necessary and such in-context learning is still unimodal*. This study aims to investigate this question by examining the two main factors influencing the ICL ability, *i.e.*, the demonstration content and the demonstration selection strategy [12, 42]. Specifically, we try to answer the following research questions: 1) How does the demo content in different modalities influence the ICL ability? Does the multimodal ICL rely heavily on the single textual modality? 2) How to select multimodal demos for better ICL performance? Should we rely on images, text, or both when selecting these demos?

For the first question, experiments on multiple MLLMs and vision-language (VL) tasks revealed that textual information is crucial for successful multimodal ICL. In comparison, omitting visual information barely affects the multimodal ICL. Specifically, when the images in the demos are removed or replaced with blank images, ICL performance hardly drops. In comparison, language information such as the correct label space and semantics is more crucial than images. Text corruption in the demos can degrade performance significantly. To further understand our findings, we analyze the information flow inside the model architecture. The analysis revealed that the core design of these

*corresponding author

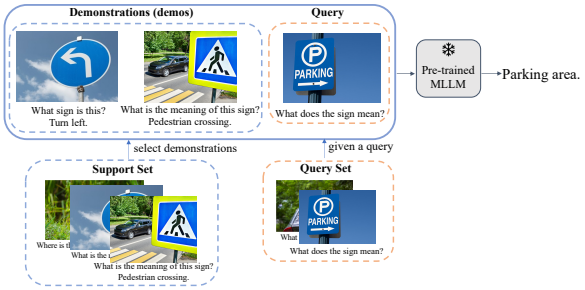

Figure 1. In-context learning (2-shot) on visual question answering. Pre-trained MLLMs can perform In-context Learning for a given query based on a few context demos (*i.e*., a list of images, questions, and answers) selected from a support set.

models, *i.e*., the masked-cross attention [1], can contribute to this phenomenon. The masked-cross attention enables the model to receive interleaved image-text sequences but can limit the influence of demonstration images on the ICL performance. Additional experiments on the model's inner states confirm that the demonstration's visual information hardly impacts the model's output. In contrast, text in the demos has a substantial and direct influence on the model's output. To conclude, *the content of demonstration images barely influences the ICL ability and the demonstration texts contribute more to the ICL performance*.

To answer the second question, we investigate demo selection strategies using different modalities. Experiments show that visual information is still necessary and useful to boost multimodal ICL performance. Compared to random selection, using image similarity to select demos can improve ICL performance. This indicates that although visual content is less significant in ICL, it can influence the demo selection process and improve ICL performance further. A reason is that visually similar demos are more likely to bring informative textual information, which is helpful for enhanced ICL performance. However, we find that selecting solely on visual or language modality can fall short and a better selection strategy should consider both. Based on our analysis, we propose a simple yet effective demonstration selection strategy, termed Mixed Modality In-Context Example Selection (MMICES), which considers both visual and language modalities. Concretely, the visual modality is first used to filter potential demo candidates. Then MMICES ranks and selects demos considering the language modality. By factoring in both visual and language information, demos selected by MMICES are related to both the query image and text. In conclusion, *although the content of demo images does not influence the ICL ability, the similarity between demo images and the query image can greatly affect the demo selection and thus influence the ICL performance. Therefore, a better demonstration selection strategy should consider both modalities.*

To summarize, our main contributions are as follows:

- This research investigates whether MLLMs can truly perform multimodal ICL. By examining the influence of multimodal demo content, it is revealed that textual information is more essential than visual information in the demos. Surprisingly, removing images from the demos results in a negligible decline in the ICL performance, whereas corruption of texts leads to a significant performance decrease.
- Our study then explores the demo selection strategy and indicates that a single modality is inadequate for selecting better demonstrations. Both visual and textual modalities should be considered when selection demonstrations to enhance ICL performance.
- Motivated by our analysis, we propose a simple yet effective method, dubbed MMICES, to enhance the ICL performance of pre-trained MLLMs. Extensive experiments show that MMICES outperforms existing demonstration selection methods in various settings.

2. Related Work

Multimodal In-Context Learning. Frozen [38] is the first attempt for ICL in multimodality by leveraging a frozen GPT-like LM. Flamingo [1] demonstrated stronger ICL performance and can handle flexible interleaved text and visual sequences. It utilizes a masked cross-attention mechanism that integrates visual information into pre-trained LLMs and allows any number of visual inputs. This capability makes the ICL possible and many MLLMs are therefore not suitable for ICL such as BLIP [23], MiniGPT [49], *etc*. OpenFlamingo [3] and IDEFICS [20] are popular open-source reproductions of Flamingo with competitive ICL performance. Otter [22] adopts instruction tuning to support more flexible tasks but it is still based on OpenFlamingo. Some other works aim to alleviate the dependency on large-scale pre-training [8, 9, 19, 31]. However, their performances are not competitive compared to pre-trained MLLMs such as Flamingo. Some recent models also demonstrate in-context learning ability [16, 33, 35, 36, 48]. However, Kosmos [16, 33] is designed for grounding and referring. Emu [35, 36] focuses more on generating outputs varying in modalities, and MMICL [48] shows a limited in-context performance increase compared to the zero-shot setting. In contrast, we focus on understanding the in-context learning ability of these MLLMs and seek more effective demonstration selection strategies for diverse multimodal tasks.

Understanding In-Context Learning. LLMs have demonstrated impressive ICL ability [6, 10, 15, 37]. A line of research focuses on understanding the influence of demo content on ICL of LLMs [2, 27, 28, 30, 45]. [30] found that the correct input-label mapping is not as important as expected whereas label space exposure and demo distribution have much more influence on the ICL performance. [28]

demonstrated the influence of order sensitivity on the ICL performance. Besides, [2] focused on how the demo diversity, similarity, and complexity affect ICL ability. Additionally, some works studied how to select better demonstrations to increase ICL performance [42]. Various methods have been proposed, such as clustering retrieval [46, 47], iterative retrieval [34], demo retrievers [26, 40], *etc*. However, the understanding of ICL on MLLMs is still underexplored. [24, 43] have explored better in-context configurations but they have not studied the importance of visual and textual content, and only conducted experiments on a single task. This study aims to understand both the influence of multimodal demo content and the demo selection strategy on various tasks such as visual question answering, visual reasoning, and image captioning.

3. Understanding Influence of Demo Content in Multimodal ICL

Multimodal In-Context Learning Formulation on MLLMs. An input query q from a query set, *i.e*., an image I_q and a question/instruction T_q, coming after a context prompt C_q, is sent to a pre-trained MLLM f. The context prompt C_q consists of N task demonstrations from a support set S. Each demonstration includes image I_i, instruction T_i, and response R_i. Then f generates a response R_q to the input query q, *e.g*., the answer to T_q, based on image I_q and the demo context C_q. Specifically, the ICL can be written as: $R_q = f([C_q, q])$, where $q = \langle I_q, T_q \rangle$, $C_q = \{\langle I_i, T_i, R_i \rangle\}_N$.

Experimental Setting for M-ICL. Four popular VL datasets across three VL tasks are applied in this study to evaluate the modality significance, namely VQAv2 [14] and OK-VQA [29] for visual question answering (VQA), GQA [17] for visual reasoning, and MSCOCO [7] for image captioning. We take the Flamingo [1] as an example and draw similar conclusions from IDEFICS [20]. Please refer to the Supplementary Section 3 for more results.

3.1. Influence of Visual Information on M-ICL

Unlike ICL on LLMs, ICL in MLLMs incorporates visual information into the demonstration. This visual information can take the form of images used for tasks such as VQA. To evaluate the significance of images in the demonstrations, we have devised the following settings:

- ***standard*** setting refers to the scenario where both demonstrations and queries incorporate their respective original image-question pairs.
- ***demo w/o images*** describes the case where all the images in C are removed. This results in the context C with N text-only instructions such as the questions in VQA or the captions in the task of image captioning.
- ***demo w/ blank images*** refers to the scenario where the images and image position tokens in the demos are kept but the original images are replaced with blank ones, *i.e*., all pixel values are set to 255. These blank images do not provide any valuable information.
- ***demo w/o query images*** refers to the setting where the image I_q in the query Q is removed whereas the images in the demonstrations are retained.

Fig. 2 presents the ICL performance in different visual demonstration settings, given randomly selected demonstrations. Compared with the *standard* setting, both the *demo w/o images* and *demo w/ blank* retain most of the ICL performance and some performances remain relatively unchanged. Conversely, the *demo w/o query images* setting results in a substantial decline in the ICL performance, with up to a 50% performance drop on VQA and nearly a 100% performance decrease on image captioning. Fig. 2 suggests that the visual information in the demonstrations has a minimal impact on the ICL performance. But the images in the query are still important for the inference.

3.2. Influence of Textual Information on M-ICL

Besides exploring the impact of visual information in the demonstrations, we also assess the significance of textual content using the following settings:

- ***standard*** refers to the case where demos incorporate their respective original image-question pairs.
- ***different answer for same question*** corresponds to the case where the original answer is replaced with another one from the same question. Despite the question remains the same, the replacement answer can vary due to the differences in the image content.
- ***random question*** describes the case where the original question T_i is replaced with another different T_j but the answer remains unchanged.
- ***random words as labels*** refers to the case where the original R_i in the demo, such as answers in VQA and captions in image captioning, is replaced with random English words. The demo text is hence meaningless.

ICL performance across these settings is displayed in Fig. 3. Compared to *standard* setting, *random question* leads to a significant drop in performance, and altering labels to random words drastically reduces the performance to nearly zero, as seen in the last bar of each sub-figure. When compared to results in Fig. 6, changes in texts can severely affect ICL performance. However, *different answer for same question* only marginally impacts performance, regardless of incorrect labels related to the provided query image. This can be because the correct label space and semantics are more crucial than images. Even when answers are inconsistent with images, they can still offer the correct label space and semantics, because they respond to the same question. For example, given an image showing

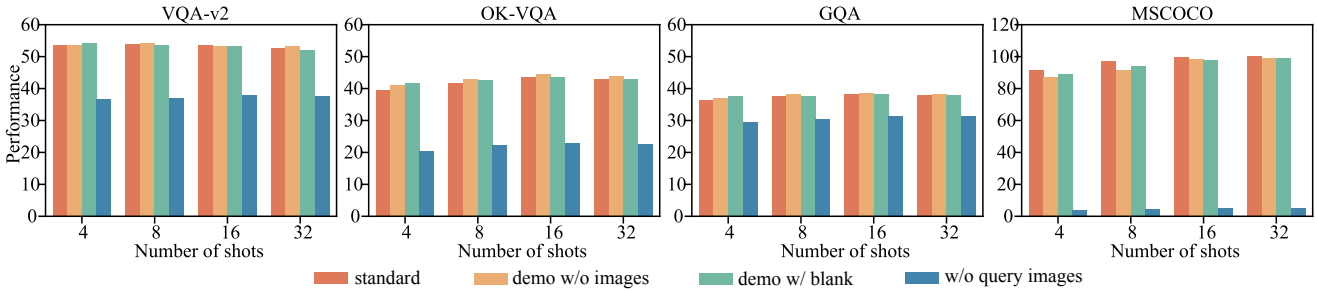

Figure 2. The ICL performance is almost the same when removing the visual information in the demonstration. Compared to the *standard* scenario, exclusion and replacement of images in the demonstration hardly impact the ICL performance (as shown in the first three bars of each sub-figure). Conversely, the removal of the query image results in substantial performance degradation (as indicated by the last bar in each sub-figure).

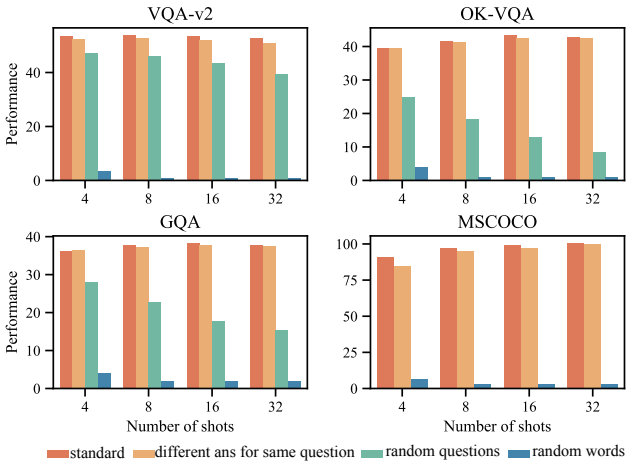

Figure 3. The ICL performance varies under different text demo settings. Performance in *different answer for same question* can still be maintained (the light orange bars). However, performance significantly decreases in *random question* and *random words as labels* (the green and blue bars).

3 books and the question "How many books are there?", other answers (*e.g.*, 5) to this question still show the correct label space, *i.e.* a number. This finding is consistent with the conclusion from previous experiments, indicating that images have minimal influence on the outcomes. This phenomenon again proves the critical role of texts in the demonstrations. Hence, we conclude that in MLLMs, the in-context learning ability is primarily driven by textual information, which shows a more substantial influence than visual information in the demonstrations.

3.3. Further Investigating Content Influence

Previous subsections highlight the dominant role of textual information in ICL for MLLMs, yet leave several questions unanswered: 1) Why do the images in demos barely affect the ICL performance? 2) Why is the query image still useful? 3) Why does the textual information dominate the ICL ability? Existing literature [13, 21, 22, 48] shows that the pre-training datasets cannot provide sophisticated con-

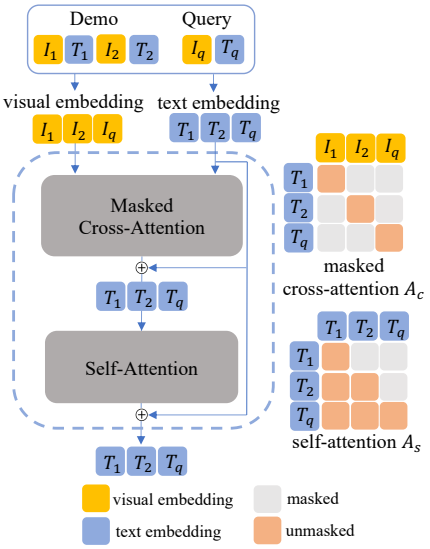

Figure 4. Model block supporting interleaved image-text inputs. Visual and language information, *i.e.*, I and T, are first fused using a masked cross-attention layer, where each text token is only conditioned on the last preceding image. Visual embeddings $\mathbf{I_1}$ and $\mathbf{I_2}$ from demonstration images cannot directly influence query text embedding $\mathbf{T_q}$, and $\mathbf{T_q}$ only sees $\mathbf{I_q}$ in the masked cross-attention, as shown in the last row of $\mathbf{A_c}$.

text information and may limit the ICL ability of the trained MLLMs. To further investigate the reasons, this work studies the popular model structures used in these models, *i.e.*, the masked cross-attention architecture [1], and analyzes the influence of multimodal demos.

The ability to handle interleaved text and image sequences makes ICL possible [1]. An illustration is presented in Fig. 4, with two demos and a query, each of which contains an image and corresponding text such as I_1 and T_1 in the first demo. The masked cross-attention layer enables the language models to incorporate visual information for the next-token prediction. This layer also limits the visual tokens the model can see at each text token. Specifically, at a given text token, the model only attends to the

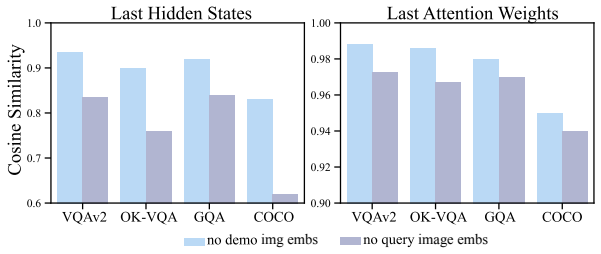

Figure 5. The left figure shows the cosine similarity between hidden states in the standard setting and removing images in the demos (blue bars). Grey bars are cosine similarity between standard setting and removing query images. The right figure shows the similarity of the corresponding attention weights in the last decoder layer. Omitting demonstration visual embeddings leads to similar hidden states, but excluding query images increases their dissimilarity, indicating the minimal influence of the demo images.

visual tokens of the last preceding image, rather than to all previous images in the interleaved sequence. For example, text embedding $\mathbf{T_q}$ can only attend to the query image $\mathbf{I_q}$ in the masked cross-attention layer, as shown in the last row of A_c in Fig. 4. Therefore, demonstration images I_1 and I_2 cannot directly pass their visual information to the query text embedding $\mathbf{T_q}$, as $\mathbf{T_q}$ is limited to interacting with the query image representation $\mathbf{I_q}$ in the masked cross-attention layer. Only in the subsequent self-attention layer can $\mathbf{T_q}$ indirectly access the information from I_1 and I_2 through the demo text embeddings $\mathbf{T_1}$ and $\mathbf{T_2}$. Because they have already processed the visual information from I_1 and I_2 in the masked cross-attention layer. We argue that the masked cross-attention mechanism with such per-image attention masking [1] diminishes text tokens' dependency on all previous images. In other words, relying solely on the self-attention layer for transferring visual information to text tokens is difficult. Thus, it is observed that the generated output tokens primarily focus on the latest image, *i.e.*, the query image, and largely disregard the visual information of the previous images, *i.e.*, the demo images.

To verify our assumptions, we compare the self-attention weights and self-attention outputs of the language decoder block in the standard setting with two scenarios, *i.e.*, with and without providing visual information in demos. If the combination of masked cross-attention across modalities and the self-attention on text tokens maintains the dependency on previous images, excluding visual information from previous demonstration images will lead to different attention behaviors, *e.g.*, different attention weights and hidden states. Otherwise, if the weights and hidden states remain almost the same after removing visual information in the demos, the model does not attend much to previous demonstration images. Specifically, we compute the cosine similarity on the last row of hidden states and attention weights in the last decoder layer for each generation forward and then average the results over the whole dataset. To remove the visual information in the demos, we mask the visual embeddings of the demo images, such as $\mathbf{I_1}$ and $\mathbf{I_2}$, by setting the weights to 0 and keeping the query image embedding $\mathbf{I_q}$. Fig. 5 presents the results. Removing demonstration visual embeddings leads to around 90% similar hidden states whereas excluding query images makes the hidden states much more dissimilar. These differences in similarity confirm our assumption and analysis above, highlighting the insignificance of demo images for existing MLLMs in ICL. This popular architecture's inherent limitation could hinder models like Flamingo [1, 3] and IDEFICS' [20] ICL ability. For some newer models that do not use this structure, such as Qwen-VL [4], similar patterns are also observed as shown in Supplementary Tab. 9. However, as different factors like training scheme [13] and dataset design [48] may come into play. Further investigation into these models is part of our future work.

4. Understanding Demo Selection in M-ICL

4.1. Demo Selection using Single Modality

Sec. 3 highlights the influence of text in the demos and this section further investigates how to select effective demos for better multimodal ICL performance. Two common approaches to selecting demos are investigated here, *i.e.*, random selection and Retrieval-based In-Context Examples Selection (RICES) [1, 3, 44]. The first randomly selects demos from the support set, disregarding different queries, whereas RICES retrieves demos with similar images by comparing them to the query images.

Compared to random selection, RICES has been proven useful to boost the ICL performance on various tasks [1, 3]. If visual information has a minimal impact on ICL performance, as shown in Sec. 3, why does RICES yield better results? To answer this question, we applied the *demo w/o images* setting to RICES, referred to as *RICES demo w/o images*. It means that the images in the context demos selected by RICES are removed and all the other textual information remains unchanged. The results are presented in Fig. 6. Surprisingly, nearly all of the ICL performances in the *RICES demo w/o images* setting remain relatively unchanged. This suggests that images in the selected demos do not significantly contribute to the performance gain. Instead, the remaining textual information plays a more crucial role and this also aligns with our findings in Sec. 3. Nevertheless, images can still serve as a good criterion for selecting demos. This is because the demonstration texts retrieved by RICES contain query-related background information, which is a crucial factor in achieving such performance gain. For instance, given a query image depicting a dinner table laden with food, RICES selects demos that are also related to food and dinner. This relevant background knowledge aids the model in better comprehending the con-

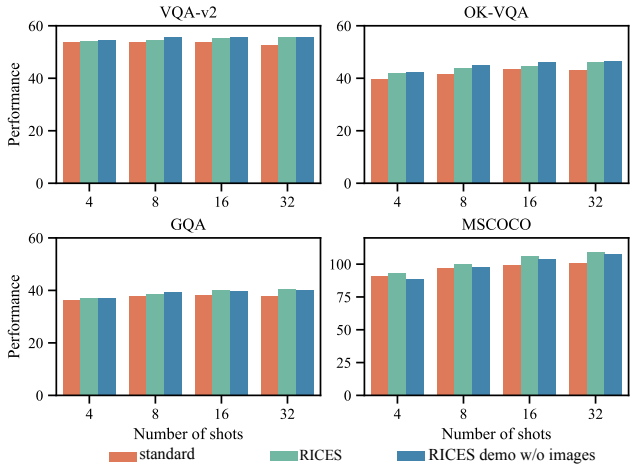

Figure 6. The performance of RICES barely changes when removing the images in the demos. RICES leads to better performance (middle bar in each sub-figure) compared to the standard random selection (the first bar). However, disregarding the images in the demos chosen by RICES has a minimal impact on the performance (the last bar) compared to the original RICES.

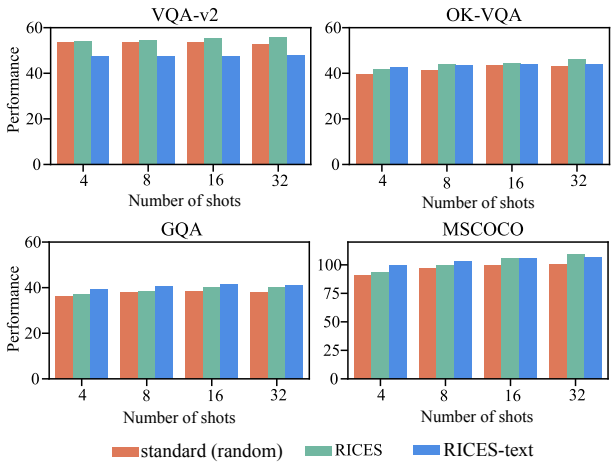

Figure 7. Although textual content in the demos is essential, retrieval based exclusively on text (the last bar in each sub-fig) cannot guarantee a better ICL performance.

text and recalling the necessary information for generating an appropriate response to the query [11, 27, 32, 44].

However, context demos chosen solely on visual similarity may not always be informative. This limitation arises because questions related to visually similar images are not necessarily interconnected. In other words, questions in the demos can differ from those in the query despite the images' similarity. As demonstrated in the first row of Fig. 8, the query question addresses the store selling the pizza, while the retrieved demos ask about the type and shape of the pizza. To identify more informative and relevant demos for queries, the retrieval process should not exclusively depend

Algorithm 1 Mixed Modality In-Context Example Selection (MMICES)

Require: query dataset Q, support dataset S, vision encoder E_v, text encoder E_t, number of pre-filtered samples K, number of demos N

$C \leftarrow [\,]$ ▷ initialize selected demos
for all query $q \in Q$ **do** ▷ select demos for each query
  $v_q \leftarrow E_v(q)$ ▷ obtain visual embedding
  $t_q \leftarrow E_t(q)$ ▷ obtain text embedding
  $s \leftarrow select(K, S, v_q)$ ▷ choose K most similar samples from S given v_q
  $c \leftarrow select(N, s, t_q)$ ▷ choose N most similar demos from s given t_q
  $C \mathrel{+}= c$ ▷ add selected demos to C
end for
return context demos C chosen from S for Q

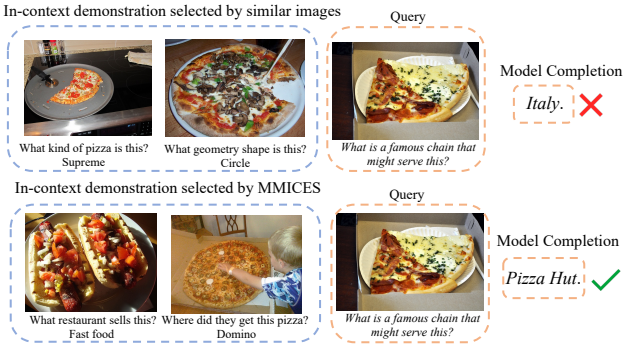

Figure 8. Demos selected by similar images (the top column) and selected by MMICES (the bottom column) given the same query. Demos containing similar images do not necessarily include related textual information to the given query. MMICES considers both visual and language modalities. It provides more informative demos and boosts the ICL performance. More empirical results can be found in Supplementary, Tab. 23.

on visual information. It should also integrate the available textual information from both demos and queries to find more informative demos. Despite the importance of textual information, retrieval based exclusively on text presents its own challenges as shown in Fig. 7. The *RICES-text* setting describes the case where demos are selected only using text similarities. The ICL performance of *RICES-text* is lower than RICES on OK-VQA and even worse compared to the random selection on VQAv2. This can be attributed to the limitations of general text queries, such as "What is in this picture?". Such queries lack specificity, providing insufficient information and potentially misleading the model generation. In summary, selecting demos based on a single modality is inadequate to provide informative and suitable demonstrations for the model to perform ICL.

4.2. Mixed Modality in-Context Example Selection

Addressing the above challenges of using a single modality to select demos, we propose to explicitly utilize multimodal information to select demos and design a simple yet effective method, named Mixed Modality In-Context Example Selection (MMICES). It initially selects candidates based on image similarity, followed by a reranking based on text similarity, as shown in Alg. 1. The objective is to select N context demos for each query q in the query dataset Q (*e.g.*, the test dataset of VQAv2) from the support dataset S (*e.g.*, the training dataset of VQAv2). First, K pre-filtered samples from S are selected based on visual feature similarity. The visual features are extracted from the vision encoder of the MLLM, and K is a hyperparameter. Then MMICES also considers textual information and selects N most similar ones from the pre-filtered K samples based on textual similarity calculated by a text encoder. **Why is the visual modality in MMICES helpful given its marginal impact in context?** While the impact of demonstration images has been found to be minimal, they are useful in preliminary demonstration retrieval. *This is mainly because similar images are more likely to bring texts containing related information.* Therefore, MMICES employs a mixed-modality approach to ensure the provision of high-quality context demos. As illustrated in the second row of Fig. 8, the demonstrations selected by MMICES are more closely related to the query and consequently assist the model in generating the correct response.

5. Experiments

5.1. Experimental Setup

Multimodal Large Language Models. Seven different models from OpenFlamingo [3] (OF) and IDEFICS [20] are used in this study. The architecture of OF and IDEFICS consists of a frozen large language model with decoder-only structure ,a frozen visual encoder followed by a trainable perceiver resampler [1, 18]. Trainable cross-attention layers are used to bridge the gap between visual and language information. Models used in this study vary in their model size (from 3B to 9B), pre-trained datasets, and whether fine-tuned by instruction tuning. The instruction-tuned versions are also used in this work, such as `IDEFICS-9B-I`. Besides, experiments are also conducted on some recent models such as MMICL [48], as described in Supplementary.

Evaluation Datasets and Metrics. Three popular tasks (visual question answering, visual reasoning, and image captioning) and 4 well-known datasets are used. For visual question answering, VQAv2 [14] and OK-VQA [29] are adopted. GQA [17] is used for visual reasoning and MSCOCO [7] for image captioning. Accuracy on the Karpathy-test split is evaluated for VQAv2. For OK-VQA, accuracy on the validation split is evaluated, and accuracy on the test-dev split is used for GQA. CIDEr [39] on the Karpathy-test split is used in MSCOCO.

Table 1. The performances of random selection (Random), RICES, and MMICES on `OF-9B`. The highest performance in each scenario is in bold. The results are averaged over 5 evaluation seeds and are reported with their standard deviations. The metric for the MSCOCO is CIDEr. For the other three datasets, top-1 accuracy is reported in percentages.

Dataset	Method	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	53.52 ±0.11	53.74 ±0.19	53.33 ±0.26	52.38 ±0.10
	RICES	**54.03** ±0.13	**54.67** ±0.06	**55.39** ±0.12	**55.77** ±0.08
	MMICES	53.11 ±0.03	53.56 ±0.05	54.04 ±0.04	55.14 ±0.02
OK-VQA	Random	39.62 ±0.29	41.56 ±0.20	43.40 ±0.39	42.97 ±0.11
	RICES	42.13 ±0.13	43.87 ±0.15	44.90 ±0.10	46.15 ±0.06
	MMICES	**44.18** ±0.11	**45.61** ±0.08	**46.93** ±0.08	**46.79** ±0.10
GQA	Random	36.32 ±0.29	37.74 ±0.32	38.28 ±0.10	37.85 ±0.11
	RICES	36.92 ±0.33	38.54 ±0.14	40.16 ±0.14	40.21 ±0.32
	MMICES	**40.73** ±0.09	**41.85** ±0.10	**42.21** ±0.12	**42.07** ±0.08
MSCOCO	Random	89.82 ±0.23	96.81 ±0.10	99.44 ±0.19	100.53 ±0.26
	RICES	93.45 ±0.07	99.74 ±0.27	105.76 ±0.03	109.12 ±0.20
	MMICES	**100.24** ±0.20	**104.90** ±0.30	**108.66** ±0.17	**109.64** ±0.24

Table 2. The performances of random selection (Random), RICES, and MMICES on `IDEFICS-9B`. MMICES achieves the best ICL performance in all settings.

Dataset	Method	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	54.90 ±0.05	56.16 ±0.02	56.93 ±0.18	57.21 ±0.17
	RICES	54.79 ±0.09	56.45 ±0.05	57.49 ±0.06	58.52 ±0.02
	MMICES	**56.15** ±0.01	**58.17** ±0.03	**59.23** ±0.01	**59.69** ±0.02
OK-VQA	Random	49.24 ±0.22	49.54 ±0.12	50.89 ±0.12	51.86 ±0.12
	RICES	48.82 ±0.02	50.55 ±0.05	52.42 ±0.03	53.22 ±0.04
	MMICES	**49.63** ±0.02	**52.16** ±0.03	**53.65** ±0.07	**54.16** ±0.05
GQA	Random	39.35 ±0.26	40.54 ±0.17	41.38 ±0.18	41.86 ±0.13
	RICES	39.86 ±0.13	41.27 ±0.29	42.65 ±0.21	43.67 ±0.19
	MMICES	**42.66** ±0.05	**44.22** ±0.08	**45.19** ±0.05	**45.36** ±0.09
MSCOCO	Random	96.45 ±0.36	100.85 ±0.36	103.96 ±0.38	105.02 ±0.43
	RICES	91.20 ±0.10	102.58 ±0.15	108.93 ±0.10	111.02 ±0.08
	MMICES	**101.13**±0.12	**109.31** ±0.09	**112.72** ±0.05	**113.37** ±0.09

5.2. Results

MMICES outperforms random selection and RICE across almost all datasets on both OpenFlamingo and IDEFICS, as shown in Tab. 1 and Tab. 2. MMICES consistently boosts the ICL performance on OpenFlamingo across various tasks. On GQA, MMICES with only 4 shots (40.73%) is better than the 32-shot random selection (37.85%) and 32-shot RICES (40.35%). MMICES is also consistently better on OK-VQA where given only 8 context examples, the performance (*i.e.*, 45.5%) is better than random 32 shots (42.97%) and RICES's 16 shots (44.70%). The performance gain is also evident on `IDEFICS-9B` across all datasets. For instance, MMICES increases the accuracy on GQA by around 10% given 8 context examples (from 40.54% to 44.22%) compared to random selection and by around 7% compared to RICES (from 41.27% to

Table 3. The performances of random selection, RICES, and MMICES on VQAv2 on OpenFlamingo and IDEFICS. MMICES achieves the best performance in most cases

Model	Method	4-shot	8-shot	16-shot	32-shot
OF-3B	Random	44.79±0.12	45.05±0.05	45.30±0.17	45.64±0.20
	RICES	44.64±0.09	45.71±0.12	46.30±0.03	47.48±0.05
	MMICES	**47.00**±0.06	**48.46**±0.07	**49.50**±0.06	**49.68**±0.03
OF-4B	Random	47.74±0.24	47.10±0.04	44.32±0.12	41.88±0.25
	RICES	47.70±0.04	46.68±0.18	44.91±0.07	42.86±0.08
	MMICES	**48.89**±0.04	**48.61**±0.09	**46.45**±0.07	**43.73**±0.06
OF-9B	Random	53.52±0.11	53.74±0.19	53.33±0.26	52.38±0.10
	RICES	**54.03**±0.13	**54.67**±0.06	**55.39**±0.12	**55.77**±0.08
	MMICES	53.11±0.03	53.56±0.05	54.04±0.04	55.14±0.02
IDEFICS-9B	Random	54.90±0.05	56.16±0.02	56.93±0.18	57.21±0.17
	RICES	54.79±0.09	56.45±0.05	57.49±0.06	58.52±0.02
	MMICES	**56.15**±0.01	**58.17**±0.03	**59.23**±0.01	**59.69**±0.02

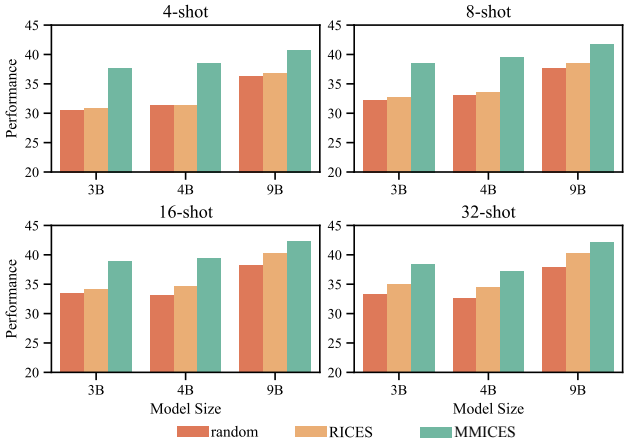

Figure 9. MMICES consistently enhances the ICL performance across models of varying sizes. MMICES on smaller models can even outperform RICES on larger models. Results here are from GQA and more results are in Supplementary Section 4.

44.22%). Besides, MMICES achieves comparable accuracy on GQA given only 4 demonstrations compared to RICES with 16 demonstrations. All the 16-shot performances from MMICES are higher compared to 32-shot random selection and 32-shot RICES, which indicates that with only half of the context examples, MMICES achieves even better results. On VQAv2, although the performances on OF-9B do not outperform the RICES, MMICES on all other models still achieves better results (Tab. 3). Overall, MMICES achieves consistently better performance compared to random selection and RICES across models and datasets.

We have also conducted extensive experiments on different MLLMs with varying sizes. Fig. 9 presents a performance comparison on the GQA dataset across models from OF-3B to OF-9B. MMICES consistently outperforms RICES by a notable margin. It is worth mentioning that MMICES on smaller-size models can achieve better performance, compared to larger-size models using RICES and random selection, especially in 4 and 8-shot settings.

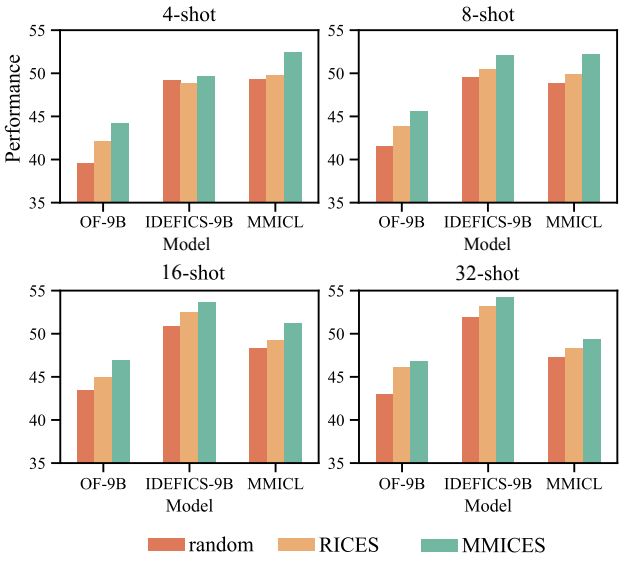

Figure 10. The performance of ICL (on OK-VQA) is consistently enhanced by MMICES across different models, including OpenFlamingo [3], IDEFICS [20], and MMICL [48].

Moreover, the performance gained is consistent across different MLLMs as shown in Fig. 10.

6. Conclusion

This study explores the multimodal in-context learning capability of multimodal large language models and asks whether these models can truly perform multimodal in-context learning. We first investigate the influence of multimodal demo content on ICL performance and find that *the visual information in the demonstrations has a minimal impact on the ICL performance, while the text is much more important for in-context learning*. This work further explores the demo selection strategy and shows that *visual information is still useful for demonstration selection. Besides, both visual and language modalities are necessary to select demonstrations*. Based on our analysis, we propose selecting demos based on both visual and text modalities and have designed the Mixed Modality In-Context Example Selection (MMICES) algorithm. Despite its simplicity, MMICES outperforms existing in-context example selection methods across various models and datasets.

Acknowledgment

This paper is supported by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. This work is partially supported by the UKRI grant: Turing AI Fellowship EP/W002981/1 and EPSRC/MURI grant: EP/N019474/1. We would also like to thank the Royal Academy of Engineering.

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. 1, 2, 3, 4, 5, 7

[2] Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11027–11052, Toronto, Canada, July 2023. Association for Computational Linguistics. 1, 2, 3

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1, 2, 5, 7, 8

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 5

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 1

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1, 2

[7] 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. 3, 7

[8] Yixin Chen, Shuai Zhang, Boran Han, and Jiaya Jia. Lightweight in-context tuning for multimodal unified models. *arXiv preprint arXiv:2310.05109*, 2023. 2

[9] Yi-Syuan Chen, Yun-Zhu Song, Cheng Yu Yeo, Bei Liu, Jianlong Fu, and Hong-Han Shuai. Sinc: Self-supervised in-context learning for vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15430–15442, 2023. 2

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 1, 2

[11] 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. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023. 6

[12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022. 1

[13] Sivan Doveh, Shaked Perek, M Jehanzeb Mirza, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision & language models. *arXiv preprint arXiv:2403.12736*, 2024. 4, 5

[14] 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. 3, 7

[15] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 2

[16] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3, 7

[18] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 7

[19] Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters: Low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021. 2

[20] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023. 1, 2, 3, 5, 7, 8

[21] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 4

[22] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal

model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 2, 4

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[24] Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. *arXiv preprint arXiv:2312.01571*, 2023. 3

[25] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022. 1

[26] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*, 2021. 3

[27] 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, May 2022. Association for Computational Linguistics. 1, 2, 6

[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, May 2022. Association for Computational Linguistics. 1, 2

[29] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 7

[30] 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? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 1, 2

[31] Masoud Monajatipoor, Liunian Harold Li, Mozhdeh Rouhsedaghat, Lin F Yang, and Kai-Wei Chang. Metavl: Transferring in-context learning ability from language models to vision-language models. *arXiv preprint arXiv:2306.01311*, 2023. 2

[32] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022. 6

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

[34] Alexander Scarlatos and Andrew Lan. Reticl: Sequential retrieval of in-context examples with reinforcement learning. *arXiv preprint arXiv:2305.14502*, 2023. 3

[35] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286*, 2023. 2

[36] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*, 2023. 2

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 2

[38] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 1, 2

[39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 7

[40] Liang Wang, Nan Yang, and Furu Wei. Learning to retrieve in-context examples for large language models. *arXiv preprint arXiv:2307.07164*, 2023. 3

[41] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023. 1

[42] Xin Xu, Yue Liu, Panupong Pasupat, Mehran Kazemi, et al. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*, 2024. 1, 3

[43] Xu Yang, Yongliang Wu, Mingzhuo Yang, and Haokun Chen. Exploring diverse in-context configurations for image captioning. *arXiv preprint arXiv:2305.14800*, 2023. 3

[44] 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*, volume 36, pages 3081–3089, 2022. 5, 6

[45] Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. Ground-truth labels matter: A deeper look into input-label demonstrations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2437, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics. 1, 2

[46] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng,

and Meng Jiang. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*, 2022. 1, 3

[47] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 3

[48] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 1, 2, 4, 5, 7, 8

[49] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2