UNDERSTANDING AND IMPROVING IN-CONTEXT LEARNING ON VISION-LANGUAGE MODELS

Shuo Chen[1,3] **Zhen Han**[1] **Bailan He**[1,3] **Mark Buckley**[3] **Philip Torr**[2]
Volker Tresp[1,4] **Jindong Gu**[2]*
[1]LMU Munich [2]University of Oxford [3]Siemens AG
[4]Munich Center for Machine Learning (MCML)
shuo.chen@campus.lmu.de, jindong.gu@eng.ox.ac.uk, hanzhen02111@163.com

ABSTRACT

In-context learning (ICL) on large language models (LLMs) has received great attention, and this technique can also be applied to vision-language models (VLMs) built upon LLMs. These VLMs can respond to queries by conditioning responses on a series of multimodal demonstrations, which comprise images, queries, and answers. Though ICL has been extensively studied on LLMs, its research on VLMs remains limited. The additional visual information in the demonstrations motivates the following research questions: which modality in the demonstration is more significant? How can we select effective multimodal demonstrations to enhance ICL performance? This study investigates the significance of both visual and language information. Our findings indicate that ICL in VLMs is predominantly driven by the textual information in the demonstrations whereas the visual information in the demonstrations barely affects the ICL performance. Motivated by our analysis, we propose a simple yet effective approach, termed Mixed Modality In-Context Example Selection (MMICES). MMICES considers both visual and language modalities when selecting demonstrations and shows better ICL performance. Extensive experiments are conducted to support our findings and improvement of the ICL performance of VLMs.

1 INTRODUCTION

The in-context learning (ICL) ability of large language models (LLMs) has received great attention and demonstrated impressive performance on various downstream tasks Brown et al. (2020); Touvron et al. (2023); Hoffmann et al. (2022); Chowdhery et al. (2022); Liang et al. (2022). The principal benefit of ICL is its ability to learn and adapt from the context by providing a set of question-and-answer pairs, referred to as demonstrations, without requiring any model parameter updates Brown et al. (2020); Bommasani et al. (2021). Recent vision-language models (VLMs) built upon LLMs have also displayed ICL ability Alayrac et al. (2022); Laurençon et al. (2023); Tsimpoukelli et al. (2021); Awadalla et al. (2023); Zhao et al. (2023); Peng et al. (2023). These pre-trained models can rapidly adapt to vision-language tasks using few-shot demonstrations, comprised of images, queries, and answers. For example, as shown in Fig. **??**, two images and the corresponding questions and answers are selected as demonstrations from an available support set. Then a pre-trained VLM generates answers for the query based on the demonstrations. While the ICL ability of LLMs has been intensively explored Min et al. (2022); Yoo et al. (2022); Wei et al. (2023); Lu et al. (2022); Liu et al. (2022); An et al. (2023), the understanding of such capability on VLMs remains largely underexplored. Unlike language models, the in-context demonstrations in VLMs integrate extra visual information. The question of whether visual or textual data in demonstrations contributes more significantly to ICL performance remains open. Furthermore, effective strategies for selecting ICL demonstrations in VLMs have yet to be established.

This study initially investigates the significance of both visual and language information within demonstrations via experiments across a diverse range of VLMs and vision-language tasks. Our

*corresponding author

experimental results indicate that textual information is crucial for successful ICL in VLMs. Surprisingly, omitting visual information barely affects the ICL performance. Specifically, when the images in the demonstrations are removed or replaced with blank images, ICL performance hardly drops. In comparison, altering text in the demonstrations degrades performance significantly.

Based on our analysis, we propose a simple yet effective strategy for choosing demonstrations for ICL on VLMs. This strategy, termed Mixed Modality In-Context Example Selection (MMICES), considers both visual and language modalities when selecting demonstrations. Firstly, the visual modality is used to filter potential demonstration candidates. Then MMICES ranks and selects demonstrations considering the language modality. By factoring in both visual and language information, demonstrations selected by MMICES are related to both the query image and text. Extensive experiments have verified the effectiveness of MMICES across multiple models and various datasets. To summarize, our main contributions are as follows:

- **Finding**: Our research examines the ICL ability of VLMs and reveals that textual information plays a more significant role than visual information in the demonstrations. Surprisingly, removing images from the demonstrations results in a negligible decline in the ICL performance whereas corruption of texts leads to a significant decrease.
- **Improvement**: Motivated by our analysis, we propose a simple yet effective method, dubbed MMICES, to enhance the in-context learning performance of pre-trained vision-language models. Extensive experiments show that MMICES outperforms existing demonstration selection methods in various settings.

2 Investigating the Importance of Visual and Textual Information in ICL

Experimental Setting for ICL Evaluation. Four popular vision-language datasets across three VL tasks are applied in this study to evaluate the significance of visual and textual information in ICL on VLMs, namely VQAv2 Goyal et al. (2017) and OK-VQA Marino et al. (2019) for visual question answering (VQA), GQA Hudson & Manning (2019) for visual reasoning, and MSCOCO Chen et al. (2015) for image captioning. Flamingo Alayrac et al. (2022) is used as an example in the following. Please refer to Appendix. C for more results.

Importance of Visual Information. ICL on VLMs 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***: demonstrations and queries have respective original image-question pairs.
- ***demo w/o images***: the visual information from the demo context C is removed by deleting all the images in C. 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***: the original images are replaced with blank images, *i.e.*, all the pixel values are set to 255. Although there are still images in the demonstrations, they do not provide any valuable information.
- ***demo w/o query images***: the image I_q presented in the query input Q is removed whereas the images in the demonstrations are retained.

Two common approaches to selecting demonstrations are investigated here, *i.e.*, random selection and Retrieval-based In-Context Examples Selection (RICES) Yang et al. (2022); Alayrac et al. (2022); Awadalla et al. (2023). The first approach randomly selects demonstrations from the support set, disregarding different queries. On the other hand, RICES retrieves demonstrations with similar images by comparing them to the query images.

Fig. 1 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* settings 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. 1 suggests that the visual information in the demonstrations has a minimal impact on the ICL performance.

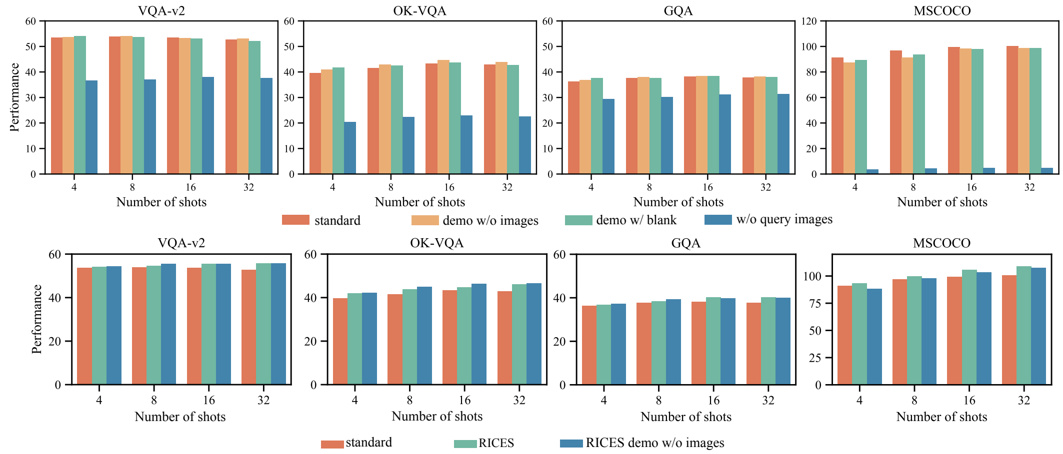


Figure 1: The top row shows that the ICL performance is almost the same when removing images in the demonstration. Compared to the *standard* scenario, exclusion and replacement of images in the demonstration hardly impact the performance (as shown in the first three bars of each sub-figure). Conversely, the removing query image results in substantial performance degradation (as indicated by the last bar in each sub-figure). The bottom row shows that the performance of RICES barely changes when removing the images in the demonstrations selected by visual similarity.

Compared to random selection, RICES has been proven useful to boost the ICL performance on various tasks Alayrac et al. (2022); Awadalla et al. (2023). If visual information has a minimal impact on ICL performance, 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 demonstrations selected by RICES are removed and all the other textual information remains unchanged. The results are presented in Fig. 1. Surprisingly, nearly all of the ICL performances in the *RICES demo w/o images* setting remain relatively unchanged. This suggests that the images in the selected demonstrations do not significantly contribute to the performance gain. Instead, the remaining textual information plays a more crucial role. 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 demonstrations that are also related to food and dinner. This relevant background knowledge aids the model in better comprehending the context and recalling the necessary information for generating an appropriate response to the query Liu et al. (2022); Yang et al. (2022); Dai et al. (2023); Olsson et al. (2022). Besides exploring the impact of visual information, we also assess the significance of textual information under several settings as shown in Appendix C.2. The experiments show that changes in texts can severely affect ICL performance. Hence, we conclude that in VLMs, in-context learning is primarily driven by textual information, which shows a more substantial influence than images. We also investigated the potential factors and suggested that the masked cross-attention layers could contribute to this phenomenon. More details are in Appendix D.

In summary, this section shows that language is more significant than visual information in the demonstrations. Excluding images from the demonstrations results in a negligible performance decline whereas text corruption and removing query images lead to a significant decrease.

3 Improving ICL Performance on Vision-language models

Previous analysis has revealed the dominant role of textual information in the demonstrations. Therefore, the demonstration selection should also consider this textual information. Random selection neglects the necessary context for different queries. Demonstrations retrieved based on visual similarity outperform random selection, but this performance gain can be mostly attributed to the informative text within the selected examples, as discussed in Sec. 2. However, context demonstrations chosen solely on visual similarity may not always be informative for a specific query. This is because questions related to visually similar images are not necessarily interconnected. As demonstrated in the first row of Fig. 8 in the appendix, the query question addresses the store selling the

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

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

Table 1: The performances of random selection, RICES, and MMICES on `OF-9B` (left) and `IDEFICS-9B` (right). The highest performance in each shot scenario is highlighted in bold. The results are averaged over 5 evaluation seeds and are reported along with their standard deviations. The performance metric for the MSCOCO dataset is CIDEr, while for the remaining datasets, accuracy is reported in percentages. MMICES achieves the best ICL performance in almost all settings.

pizza, while the retrieved demonstrations ask about the type and shape of the pizza. To identify more informative demonstrations for queries, the retrieval process should not exclusively depend on visual information. It should also integrate the available textual information from both demonstrations and queries to find more useful demonstrations. Despite the importance, retrieval based exclusively on text presents its own challenges. For instance, general queries such as "What is in this picture?" often fail to provide sufficient information.

Addressing the above challenges, we utilize both modalities to select demonstrations 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 demonstrations 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 vision-language model, and K is a hyperparameter. Then MMICES considers textual information and selects N most similar ones from the pre-filtered K samples based on textual similarity calculated by a text encoder.

4 EXPERIMENTS

Experimental Setup. We investigate 7 different models from OpenFlamingo Awadalla et al. (2023) (OF) and IDEFICS Laurençon et al. (2023). Models used in this study vary in their model size (from 3B to 9B), pre-trained datasets, and whether fine-tuned by instruction tuning. Three representative VL tasks (visual question answering, visual reasoning, and image captioning) and 4 well-known VL datasets are applied in this work, including VQAv2 Goyal et al. (2017), OK-VQA Marino et al. (2019), GQA Hudson & Manning (2019) and MSCOCO Chen et al. (2015). Accuracy and CIDEr are used as metrics. More detailed information is in Appendix. B.

Results. MMICES outperforms random selection and RICE across almost all datasets on both models, as shown in Tab. 1. MMICES consistently boosts the ICL performance on OpenFlamingo across various vision-language 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 44.22%). 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. Overall, MMICES achieves better performance compared to random selection in all scenarios and RICES in most cases. More results and analysis are in Appendix. E.

5 Conclusion

This study explores the in-context learning capabilities of vision-language models. We find that the visual information in the demonstrations has a minimal impact on the ICL performance, while the text is more important. Based on our analysis, we propose selecting demonstrations based on both visual and text modalities and have designed the Mixed Modality In-Context Example Selection (MMICES) algorithm, which outperforms existing in-context example selection methods. We believe this study can help the community better understand the ICL ability of VLMs.

Acknowledgement

The authors acknowledge support by the German Federal Ministry for Education and Research (BMBF), funding project "Software Campus 2.0 (LMU Muenchen) (grant 01IS17048)". This paper is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. This work has been supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) as part of the project CoyPu under grant number 01MK21007K.

References

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.

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)*, pp. 11027–11052, Toronto, Canada, July 2023. Association for Computational Linguistics. URL `https://aclanthology.org/2023.acl-long.618`.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

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.

Hritik Bansal, Karthik Gopalakrishnan, Saket Dingliwal, Sravan Bodapati, Katrin Kirchhoff, and Dan Roth. Rethinking the role of scale for in-context learning: An interpretability-based case study at 66 billion scale. *arXiv preprint arXiv:2212.09095*, 2022.

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.

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.

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.

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*, pp. 15430–15442, 2023.

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.

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.

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*, pp. 6904–6913, 2017.

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 (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=iBBcRUlOAPR`.

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*, pp. 6700–6709, 2019.

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.

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*, 2023a.

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*, pp. 12888–12900. PMLR, 2022.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023b.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

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*, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL `https://aclanthology.org/2022.deelio-1.10`.

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)*, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL `https://aclanthology.org/2022.acl-long.556`.

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.

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*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.759`.

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.

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.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

MosaicML NLP Team et al. Introducing mpt-7b: A new standard for open-source, ly usable llms. `https://www.mosaicml.com/blog/mpt-7b`, 2023.

together.ai. Releasing 3b and 7b redpajama- incite family of models including base, instruction-tuned and chat models. `https://together.ai/blog/redpajama-models-v1`, 2023.

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.

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.

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*, pp. 4566–4575, 2015.

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.

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

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, pp. 3081–3089, 2022.

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*, pp. 2422–2437, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.155`.

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.

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*, 2023a.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Young-jae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023b.

A Related Work

In-Context Learning on Vision-Language Models. Frozen Tsimpoukelli et al. (2021) is the first attempt for ICL in multimodality by leveraging a frozen GPT-like LM. Flamingo Alayrac et al. (2022) 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 VLMs are therefore not suitable for ICL such as BLIP Li et al. (2022), MiniGPT Zhu et al. (2023a), *etc*. OpenFlamingo Awadalla et al. (2023) and IDEFICS Laurençon et al. (2023) are popular open-source reproductions of Flamingo with competitive ICL performance. Otter Li et al. (2023a) adopts instruction tuning to support more flexible tasks but its model architecture is the same as Flamingo's and still uses masked cross-attention to incorporate visual information. Besides, Yang et al. has explored better in-context configurations but this work has not studied the importance of visual and textual information and only conducted experiments on image captioning. Some other works aim to alleviate the dependency on large-scale pre-training, such as SINC Chen et al. (2023) and MetaVL Monajatipoor et al. (2023). However, their performances are not competitive compared to pre-trained VLMs such as Flamingo. In contrast to these studies, we focus on the understanding of the in-context learning ability of vision-language models and seek more effective demonstration selection strategies for diverse vision-language tasks.

Understanding In-Context Learning. LLMs have demonstrated impressive ICL ability (Brown et al., 2020; Touvron et al., 2023; Hoffmann et al., 2022; Chowdhery et al., 2022), *i.e.*, adapting to a new task conditioned on a few in-context demonstrations without any gradient update. A line of research focuses on understanding the importance of different aspects of the ICL demonstrations on LLMs Min et al. (2022); Yoo et al. (2022); Lu et al. (2022); Liu et al. (2022); An et al. (2023). Min et al. found that the correct input-label mapping is not as important as expected whereas label space exposure and demonstration distribution have much more influence on the ICL performance. Yoo et al. further found that correct demonstration labels can impact the ICL performance in certain specific scenarios. Moreover, Lu et al. demonstrated the influence of order sensitivity on the ICL performance. Liu et al. revealed that semantically similar examples to a test query can lead to better ICL performance. Besides, An et al. focused on how the diversity, similarity, and complexity of demonstrations affect ICL ability. Moreover, some works studied ICL on LLMs from the perspective of model architectures Olsson et al. (2022); Bansal et al. (2022) and revealed that model components are closely related to ICL performance. However, ICL on VLMs differs due to the additional visual information in the demonstrations and different model components. This study focuses on ICL on VLMs and aims to understand which aspect of information in the multimodal demonstrations holds greater significance.

B Experimental Setup

Model	Vision Encoder	Language Model
`OF-3B`	CLIP Vit-L/14	MPT-1B Team et al. (2023)
`OF-3B-I`	CLIP Vit-L/14	MPT-1B-I Team et al. (2023)
`OF-4B`	CLIP Vit-L/14	RedPajama-3B together.ai (2023)
`OF-4B-I`	CLIP Vit-L/14	RedPajama-3B-I together.ai (2023)
`OF-9B`	CLIP Vit-L/14	MPT-7B Team et al. (2023)
`IDEFICS-9B`	OpenCLIP Vit-H/14	LLaMA-7B Touvron et al. (2023)
`IDEFICS-9B-I`	OpenCLIP Vit-H/14	LLaMA-7B Touvron et al. (2023)

Table 2: Vision-language models studied in this work. `OF` stands for OpenFlamingo Awadalla et al. (2023) and `I` means instructed version.

Vision-language Models. We investigate different models from OpenFlamingo Awadalla et al. (2023) and IDEFICS Laurençon et al. (2023) with various model sizes as shown in Tab. 2. OpenFlamingo Awadalla et al. (2023) and IDEFICS Laurençon et al. (2023) are popular open-source reproductions of Flamingo with competitive ICL performance. The architecture of these models consists of a frozen large language model with decoder-only structure (*e.g.*, MPT Team et al. (2023) in OpenFlmaingo and LLaMA Touvron et al. (2023) in IDEFICS), a frozen visual encoder (*e.g.*, CLIP-ViT Radford et al. (2021)) followed by a trainable perceiver resampler. There are also trainable gated cross-attention layers interleaved between pre-trained LM layers to bridge the gap between visual and language information. Per-image attention masking is adopted in these cross-attention layers. This ensures that at any particular text token, the model focuses solely on the visual tokens from the immediately preceding image in the interleaved sequence, rather than on all preceding images. The 7 models used in this study vary in their model size (from 3B to 9B), pre-trained datasets, and whether fine-tuned by instruction tuning. OpenFlamigo is trained on 2B image-text pairs in LAION-2B Schuhmann et al. (2022) and 43M interleaved image-text sequences in Multimodal C4 Zhu et al. (2023b). IDEFICS is trained on OBELICS Laurençon et al. (2023) which contains 141M multimodal Engish web documents with 353M images and 115B tokens. Both models achieve competitive performance compared to Flamingo Alayrac et al. (2022). The instruction-finetuned versions are also used in this work. For instance, `IDEFICS-9B-I` starts from the base IDEFICS models and is fine-tuned by unfreezing all the parameters on various datasets, such as M3IT Li et al. (2023b) and LLaVA-Instruct Liu et al. (2023).

Evaluation Datasets and Metrics. Three popular VL tasks (*i.e.*, visual question answering, visual reasoning, and image captioning) and 4 well-known VL datasets are applied in this work. For visual question answering, VQAv2 Goyal et al. (2017) and OK-VQA Marino et al. (2019) are adopted. Additionally, we incorporate GQA Hudson & Manning (2019) for visual reasoning and MSCOCO Chen et al. (2015) for image captioning. The statistics are in Tab. 3. 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 Vedantam et al. (2015) on the Karpathy-test split is used in MSCOCO.

Task	Dataset	# Images	# Image-text pairs
Visual Question Answering	VQAv2 Antol et al. (2015)	123.2K	658.1K
	OK-VQA Marino et al. (2019)	14K	14K
Visual Reasoning	GQA Hudson & Manning (2019)	82.3K	1087.7K
Image Captioning	MSCOCO Chen et al. (2015)	123.2K	576.8K

Table 3: Dataset Statistics. Four well-known datasets from three popular vision-language tasks are used in this study.

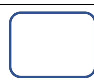
Setting	demo image	demo question	demo response	query image	query question
standard		What sign is this?	Turn left		What does the sign mean?
demo w/o images		What sign is this?	Turn left		What does the sign mean?
demo w/ blank images		What sign is this?	Turn left		What does the sign mean?
demo w/o query images		What sign is this?	Turn left		What does the sign mean?

Table 4: Examples for different visual demonstration settings with one demonstration and one query. *Demo w/o images* removes the images in the demonstration. *demo w/ blank images* replaces the images with blank ones. *demo w/o query images* removes the images in the query.

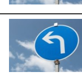
Setting	demo image	demo question	demo response	query image	query question
standard		What sign is this?	Turn left		What does the sign mean?
different answer for same question		What sign is this?	No entry		What does the sign mean?
random question		What kind of food is this?	Turn left		What does the sign mean?
random words as labels		What sign is this?	Hello		What does the sign mean?

Table 5: Examples for different textual demonstration settings with one demonstration and one query. The differences compared to the `standard` setting are highlighted in blue.

C Additional Results of Importance Investigation on Visual and Textual Information

In-Context Learning Formulation on VLMs. In vision-language in-context learning, 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 vision-language model 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 $q = \langle I_q, T_q \rangle, \;\; C_q = \{\langle I_i, T_i, R_i \rangle\}_N$.

C.1 Importance of Visual Information

To evaluate the importance of visual information, we have designed various demonstration settings as shown in Tab. 4.

- ***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 the visual information from the demo context is removed by deleting all the images in the context demonstration. The context then only includes 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 demonstrations are kept but the original images are replaced with blank images, *i.e.*, all the pixel values are set to 255. Although there are still images in the demonstrations, they do not provide any valuable information.
- ***demo w/o query images*** refers to the setting in which the image presented in the query input is removed whereas the images in the demonstrations are retained.

Performance of `OF-9B` and `IDEFICS-9B` across 4 datasets given random selected demonstrations are presented in Tab. 6 and Tab. 7. When compared to the *standard* setting, the *demo w/o images* and *demo w/ blank* settings largely maintain the ICL performance, with some aspects showing little change. In contrast, the *demo w/o query images* setting leads to a significant reduction in ICL performance, including up to a 50% decrease in VQA performance and nearly a 100% decrease in image captioning performance. We also conducted experiments using RICES, *i.e.*, Retrieval-based In-Context Examples Selection, in the *demo w/o images* setting and the results are in Tab. 8 and Tab. 9. The results also suggest that the images in the selected demonstrations do not significantly contribute to the performance gain. Instead, the remaining textual information is more important.

C.2 Importance of Textual Information

Importance of Textual information. Besides exploring the impact of visual information, we also assess the significance of textual information given randomly selected demonstrations using the following settings:

- ***standard*** refers to the case where demonstrations 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 T_j that has different content but the answer remains unchanged.
- ***random words as labels*** refers to the case where the original response R_i in the demonstration, such as answers in VQA and captions in image captioning, is replaced with random English words.

Dataset	Setting	4-shot	8-shot	16-shot	32-shot
VQAv2	standard	53.60	53.85	53.60	52.74
	demo w/o img	53.61	54.15	53.36	53.15
	demo w/ blank img	54.13	53.71	53.12	52.10
	demo w/o query img	36.72	37.11	37.95	37.67
OK-VQA	standard	39.62	41.56	43.40	42.97
	demo w/o img	40.98	42.86	44.61	43.91
	demo w/ blank img	41.77	42.57	43.64	42.82
	demo w/o query img	20.42	22.38	22.95	22.67
GQA	standard	36.32	37.74	38.28	37.85
	demo w/o img	36.86	38.13	38.40	38.23
	demo w/ blank img	37.63	37.73	38.36	38.03
	demo w/o query img	29.39	30.24	31.23	31.41
MSCOCO	standard	91.22	96.88	99.44	100.53
	demo w/o img	87.26	91.49	98.35	98.85
	demo w/ blank img	89.25	93.88	97.91	96.91
	demo w/o query img	3.57	4.30	4.90	4.85

Table 6: The performances of `OF-9B` on different visual demonstration settings given random selected demonstrations.

Dataset	Setting	4-shot	8-shot	16-shot	32-shot
VQAv2	standard	54.90	56.16	56.93	57.21
	demo w/o img	53.66	54.57	55.41	55.34
	demo w/ blank img	53.69	54.38	54.98	55.04
	demo w/o query img	38.64	39.27	39.71	39.99
OK-VQA	standard	49.24	49.54	51.47	51.86
	demo w/o img	47.63	48.28	48.74	48.99
	demo w/ blank img	47.66	48.55	49.83	50.24
	demo w/o query img	26.91	27.70	28.32	28.67
GQA	standard	39.35	40.54	41.38	41.87
	demo w/o img	38.64	39.45	40.27	40.85
	demo w/ blank img	38.36	39.94	40.71	41.36
	demo w/o query img	31.82	32.47	33.12	33.50
MSCOCO	standard	97.45	101.85	102.96	105.62
	demo w/o img	67.77	81.01	85.81	90.72
	demo w/ blank img	88.75	92.27	95.49	96.83
	demo w/o query img	2.86	3.14	3.05	3.02

Table 7: The performances of `IDEFICS-9B` on different visual demonstration settings given random selected demonstrations.

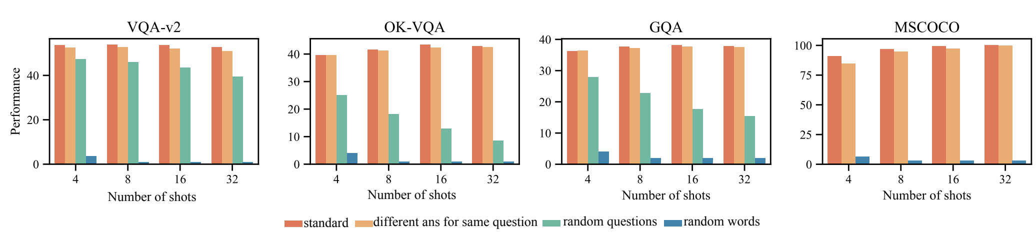


Figure 3: The ICL performance varies under different language demonstration settings. Performance in *different answer for same question* can still be maintained (the light orange bar in each sub-figure). However, performance significantly decreases in *random question* and *random words as labels*, as depicted by the green and blue bars respectively.

Dataset	**Method**	**4-shot**	**8-shot**	**16-shot**	**32-shot**
VQAv2	Random	53.60	53.85	53.60	52.74
	RICES	54.17	54.67	55.39	55.77
	RICES demo w/o img	54.38	55.46	55.56	55.71
OK-VQA	Random	39.62	41.56	43.40	42.97
	RICES	42.00	43.87	44.70	46.15
	RICES demo w/o img	42.23	44.94	46.20	46.65
GQA	Random	36.32	37.74	38.28	37.85
	RICES	36.92	38.54	40.17	40.35
	RICES demo w/o img	37.21	39.37	397.84	40.05
MSCOCO	Random	91.22	96.88	99.44	100.53
	RICES	93.45	99.74	105.76	109.12
	RICES demo w/o img	88.49	97.82	103.67	107.69

Table 8: The performances of `OF-9B` on different visual demonstration settings given demonstrations selected by RICES.

Dataset	**Method**	**4-shot**	**8-shot**	**16-shot**	**32-shot**
VQAv2	Random	54.90	56.16	56.93	57.21
	RICES	54.79	56.45	57.49	58.60
	RICES demo w/o img	54.94	56.20	57.19	57.67
OK-VQA	Random	49.24	49.54	51.47	51.86
	RICES	48.82	50.55	52.42	53.22
	RICES demo w/o img	48.02	50.24	51.60	51.76
GQA	Random	39.35	40.54	41.38	41.87
	RICES	39.86	41.27	43.01	43.67
	RICES demo w/o img	39.33	41.15	42.44	43.41
MSCOCO	Random	97.45	101.85	102.96	105.62
	RICES	91.20	102.58	108.93	111.03
	RICES demo w/o img	64.15	73.62	79.45	84.92

Table 9: The performances of `IDEFICS-9B` on different visual demonstration settings given demonstrations selected by RICES.

Dataset	Setting	4-shot	8-shot	16-shot	32-shot
VQAv2	standard	53.60	53.85	53.46	52.74
	diff ans for same question	52.49	52.70	52.06	50.92
	random question	41.48	33.94	27.93	20.03
	random words as labels	3.59	0.03	0.00	0.00
OK-VQA	standard	39.62	41.56	43.40	42.97
	diff ans for same question	39.63	41.23	42.41	42.44
	random question	25.03	18.23	13.00	8.59
	random words as labels	3.95	0.10	0.01	0.00
GQA	standard	36.23	35.92	37.29	34.38
	diff ans for same question	36.38	37.25	37.75	37.58
	random question	28.01	22.83	17.71	15.44
	random words as labels	2.06	0.05	0.00	0.00
MSCOCO	standard	91.23	96.88	99.44	100.53
	diff ans for same question	84.96	94.95	97.44	99.71
	random words as labels	1.60	0.62	0.17	0.00

Table 10: The performances of `OF-9B` on different textual demonstration settings given random selected demonstrations.

Dataset	Setting	4-shot	8-shot	16-shot	32-shot
VQAv2	standard	54.90	56.16	56.93	57.21
	diff ans for same question	54.10	55.21	56.15	57.01
	random question	47.25	45.94	43.53	39.48
	random words as labels	5.91	0.34	0.03	0.00
OK-VQA	standard	49.24	49.54	51.47	51.86
	diff ans for same question	49.25	50.18	51.11	50.95
	random question	38.41	34.04	30.08	29.53
	random words as labels	7.38	1.33	0.30	0.11
GQA	standard	39.35	40.54	41.38	41.87
	diff ans for same question	38.80	40.07	41.49	41.92
	random question	33.65	33.61	32.13	30.04
	random words as labels	3.14	0.27	0.02	0.03
MSCOCO	standard	97.45	101.85	102.96	105.62
	diff ans for same question	84.12	64.83	52.70	53.38
	random words as labels	0.00	0.00	0.00	0.00

Table 11: The performances of `IDEFICS-9B` on different textual demonstration settings given random selected demonstrations.

ICL performance across these settings is displayed in Fig. 3. Compared to *standard* setting, *different answer for same question* only marginally impacts performance, regardless of incorrect labels related to the provided query image. This finding is consistent with the conclusion from previous experiments, indicating that images have minimal influence on the outcomes. However, *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. Performance of `OF-9B` and `IDEFICS-9B` across 4 datasets given randomly selected demonstrations are presented in Tab. 10 and Tab. 11. When compared to results in Fig. **??**, changes in texts can severely affect ICL performance. Hence, we conclude that in VLMs, in-context learning is primarily driven by textual information, which shows a more substantial influence than images.

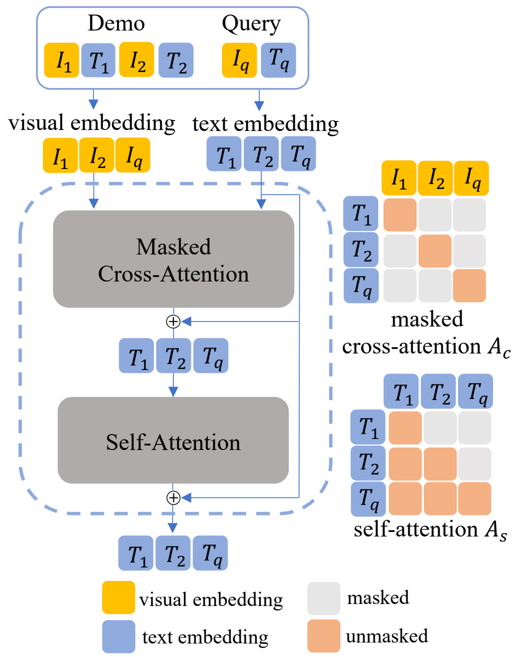


Figure 4: Model block supporting interleaved image-text inputs. Visual and language information, *i.e.*, I and T, are first fused using a masked cross-attention layer, where each text token is only conditioned on the last preceding image. Visual embeddings $\mathbf{I_1}$ and $\mathbf{I_2}$ from demonstration 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 $\mathbf{A_c}$.

D Understanding Multimodal Information Flow inside Model

The empirical examination outlined in Sec. 2 highlights the dominant role of textual information in ICL for VLMs, yet leaves several questions unanswered: 1) Why do the images in context demonstrations barely affect the ICL performance? 2) Why is the query image still useful? 3) Why does the textual information dominate the ICL ability? To fully investigate the underlying reasons, this section delves into the model details to analyze the influence of both visual and language information in the context demonstrations.

VLMs with ICL ability can handle interleaved text and visual sequences, making in-context few-shot learning possible Alayrac et al. (2022). An illustration is presented in Fig. 4, with two demonstrations and a query, each of which contains an image and corresponding text such as I_1 and T_1 in the first demonstration. The masked cross-attention layer enables the language models to incorporate visual information for the next-token prediction. This layer also limits which visual tokens the model sees at each text token. Specifically, at a given text token, the model only attends to the 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 representation $\mathbf{I_q}$ 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 demonstration 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 Alayrac et al. (2022) complicates the realization of 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, in the ICL settings, it is observed that the generated output tokens primarily focus on the latest image, *i.e.*, the query image. However, it largely disregards the visual information of the previous demonstration images.

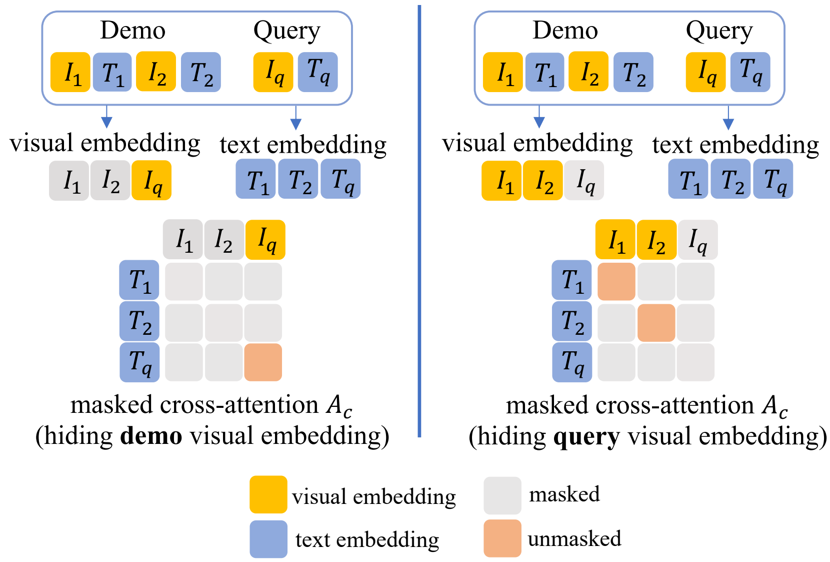


Figure 5: Compared with the standard setting, we hide demo visual embedding and query visual embedding respectively to explore the influence of different visual embeddings.

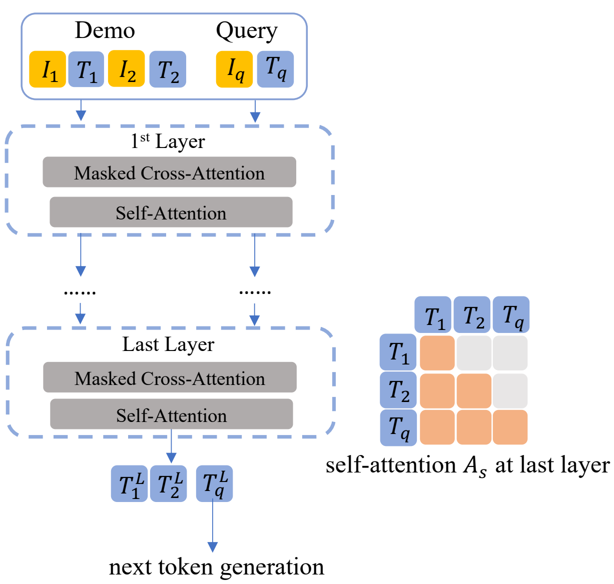


Figure 6: We compute the cosine similarity on the last row of hidden states,*i.e.*, $\mathbf{T}_q^L$ in this figure, and attention weights, *i.e.*, $\mathbf{A}_s$ in this figure, in the last decoder layer for each generation forward and then average the results over the whole dataset.

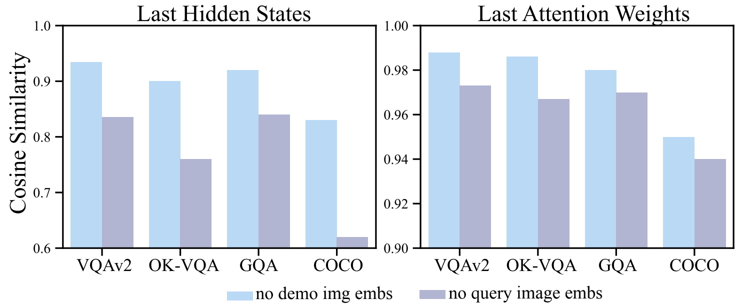


Figure 7: The left figure shows the cosine similarity between hidden states in the standard setting and removing images in the demonstrations (blue bars). Grey bars are cosine similarity between standard setting and removing query images. The right figure shows the similarity of the corresponding attention weights in the last decoder layer. Omitting demonstration visual embeddings leads to similar hidden states, but excluding query images increases their dissimilarity.

To verify our assumptions, we design experiments to 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 demonstrations. 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 demonstration examples, it indicates that the model does not much attend to previous demonstration images. Specifically, we have devised three settings.

- ***standard*** refers to the original ICL setting where visual embeddings in demonstrations and queries are retained.
- ***hide demo visual embedding*** describes the case where the visual embeddings from demonstration images are masked and the model can only see the images from the query, as shown in the left side of Fig. 5.
- ***hide query visual embedding*** refers the case where the visual embeddings from query images are masked, as shown in the right side of Fig. 5.

To examine the varying effects of visual embeddings in demonstrations and queries, we can compare the hidden states and attention weights in the last layer. In particular, we extract the last row of the hidden states (referred to as $\mathbf{T}_q^L$ in Fig. 6) and the attention weights in the last layer. We then compute the cosine similarity between these extracted values and their counterparts in the *standard* setting. 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 demonstration, we mask the visual embeddings of the demonstration images, such as $\mathbf{I_1}$ and $\mathbf{I_2}$, in demonstrations by setting the corresponding weights to 0 and keeping the query image embedding $\mathbf{I_q}$.

Fig. 7 presents the results. The removal of 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. They further highlight the insignificance of demonstration images when applying in-context learning to existing vision-language models.

In summary, this section explores the information flow in the masked cross-attention layers in VLMs and assesses the impact of the visual information in the demonstrations. The analysis reveals that: 1) Image embeddings from demonstrations do not directly contribute to the attention computation for the answer generation, thus, they have a minimal effect on the ICL performance. 2) However, query image embeddings directly connect with the answer token embeddings, making these images valuable. 3) Textual information from demonstrations can directly influence the generated answer embeddings during the self-attention process, exhibiting a significant influence on model generation.

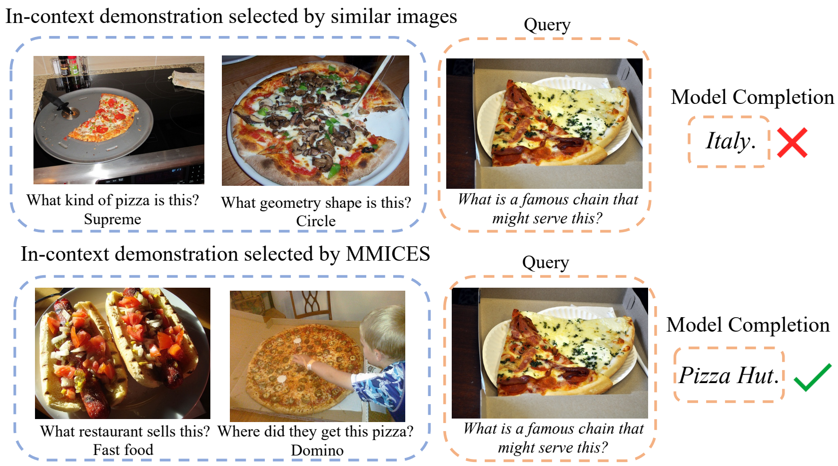


Figure 8: Demonstrations selected by similar images (the first row) and selected by MMICES (the second row) given the same query. Demonstrations containing similar images do not necessarily include related textual information to the given query. MMICES considers both visual and language modalities during retrieval and can provide more informative demonstrations for ICL.

E More Results on the ICL Performance Improvement

E.1 More Results

Algorithm 1: MMICES

Input: query dataset Q, support dataset S, vision encoder E_v, text encoder E_t, K, N
Output: context demonstrations C chosen from S for Q

```
Initialization: C ← [];
for query q ∈ Q do
    v_q ← E_v(q) ;
    t_q ← E_t(q) ;
    visual similar samples ← choose K most similar samples
      from S based on v_q ;
    demos ← choose N most similar demos from visual
      similar samples based on t_q ;
    C += demos ;
end
```

We have conducted experiments using various models and VL datasets, which are listed in Table 2 and Table 3. The results, based on all models, are obtained from demonstrations selected using random selection, RICES, and MMICES, and are presented in Table 12 to Table 18. Overall, MMICES outperforms the other two methods and achieves the best results in most cases. Tab. 22 presents examples selected by MMICES and RICES.

E.2 Ablation Study

The choices of K. The number of pre-filtered samples, denoted as K, selected by visual similarity is a hyperparameter in MMICES. A larger value of K allows for a broader selection space for the second filtering stage, while a smaller value of K is more efficient. The performance comparison for different values of K ($k \in \{50, 100, 200, 300\}$) is presented in Table 19. A larger K results in a greater number of candidate demonstrations filtered by visual similarity, which is particularly useful when the number of shots is small. However, a larger K may also include visual-unrelated demonstrations despite having similar text, potentially leading to a negative impact on performance.

Textual information on image captioning. MMICES considers both visual and textual information when selecting demonstrations. It chooses demonstrations that have both similar images and similar texts. However, in the task of image captioning, the textual information in the queries cannot be directly used as the desired response. To obtain the desired textual information, MMICES first uses the generated captions from the in-context learning setting with randomly selected demonstrations. It then further selects similar demonstrations. The performance comparison for different shot num-

Dataset	Method	0-shot	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	43.45 (0.16)	44.79 (0.12)	45.05 (0.05)	45.30 (0.17)	45.64 (0.20)
	RICES	43.45 (0.16)	44.64 (0.09)	45.71 (0.12)	46.30 (0.03)	47.48 (0.05)
	MMICES	43.45 (0.16)	**47.00 (0.06)**	**48.46 (0.07)**	**49.50 (0.06)**	**49.68 (0.03)**
OK-VQA	Random	28.18 (0.25)	30.46 (0.29)	30.29 (0.50)	31.40 (0.25)	31.40 (0.44)
	RICES	28.18 (0.25)	30.89 (0.09)	32.47 (0.04)	33.97 (0.12)	34.85 (0.04)
	MMICES	28.18 (0.25)	**35.34 (0.19)**	**37.41 (0.01)**	**38.00 (0.13)**	**38.23 (0.09)**
GQA	Random	28.70 (0.22)	30.57 (0.09)	32.31 (0.19)	33.49 (0.30)	33.33 (0.10)
	RICES	28.70 (0.22)	30.96 (0.06)	32.69 (0.20)	34.08 (0.11)	35.02 (0.04)
	MMICES	28.70 (0.22)	**37.70 (0.06)**	**38.49 (0.10)**	**38.85 (0.17)**	**38.37 (0.16)**
MSCOCO	Random	75.14 (0.69)	76.48 (0.50)	82.01 (0.35)	86.52 (1.00)	90.53 (0.42)
	RICES	75.14 (0.69)	90.30 (0.09)	97.38 (0.36)	102.91 (0.26)	105.62 (0.10)
	MMICES	75.14 (0.69)	**99.21 (0.23)**	**103.42 (0.35)**	**106.94 (0.21)**	**109.19 (0.31)**

Table 12: The performances of random selection, RICES, and MMICES on `OF-3B`. The highest performance in each shot scenario is highlighted in bold. The results are averaged over 5 evaluation seeds and are reported along with their standard deviations. The performance metric for the MSCOCO dataset is CIDEr, while for the remaining datasets, accuracy is reported in percentages. MMICES achieves the best performance in all settings on all datasets.

Dataset	Method	0-shot	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	43.55 (0.18)	45.54 (0.12)	45.77 (0.19)	45.71 (0.15)	45.05 (0.19)
	RICES	43.55 (0.18)	45.06 (0.09)	45.41 (0.07)	45.65 (0.04)	46.11 (0.12)
	MMICES	43.55 (0.18)	**48.41 (0.01)**	**48.38 (0.05)**	**48.96 (0.05)**	**48.86 (0.04)**
OK-VQA	Random	29.07 (0.17)	31.26 (0.44)	31.85 (0.10)	32.08 (0.20)	31.37 (0.12)
	RICES	29.07 (0.17)	32.30 (0.11)	33.76 (0.14)	34.52 (0.07)	35.51 (0.03)
	MMICES	29.07 (0.17)	**37.10 (0.13)**	**38.65 (0.09)**	**39.04 (0.10)**	**38.24 (0.03)**
GQA	Random	29.68 (0.17)	32.07 (0.06)	33.43 (0.30)	33.75 (0.24)	33.18 (0.28)
	RICES	29.68 (0.17)	30.96 (0.06)	33.27 (0.26)	34.17 (0.15)	34.36 (0.08)
	MMICES	29.68 (0.17)	**37.72 (0.11)**	**38.64 (0.06)**	**38.58 (0.03)**	**38.25 (0.15)**
MSCOCO	Random	75.10 (0.24)	82.11 (0.68)	86.14 (0.39)	90.17 (0.46)	92.86 (0.44)
	RICES	75.10 (0.24)	92.43 (0.23)	99.36 (0.23)	104.48 (0.33)	106.88 (0.21)
	MMICES	75.10 (0.24)	**100.43 (0.14)**	**104.82 (0.13)**	**107.61 (0.18)**	**109.44 (0.25)**

Table 13: The performances of random selection, RICES, and MMICES on `OF-3BI`. MMICES achieves the best performance in all settings on all datasets.

Dataset	Method	0-shot	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	44.05 (0.20)	47.74 (0.24)	47.10 (0.04)	44.32 (0.12)	41.88 (0.25)
	RICES	44.05 (0.20)	47.70 (0.04)	46.68 (0.18)	44.91 (0.07)	42.86 (0.08)
	MMICES	44.05 (0.20)	**48.89 (0.04)**	**48.61 (0.09)**	**46.45 (0.07)**	**43.73 (0.06)**
OK-VQA	Random	31.31 (0.32)	35.01 (0.25)	33.87 (0.20)	29.04 (0.16)	27.09 (0.29)
	RICES	31.31 (0.32)	34.97 (0.16)	33.41 (0.07)	29.47 (0.09)	28.79 (0.08)
	MMICES	31.31 (0.32)	**37.46 (0.09)**	**37.20 (0.10)**	**33.99 (0.12)**	**30.23 (0.05)**
GQA	Random	27.16 (0.01)	31.45 (0.35)	33.07 (0.25)	33.17 (0.33)	32.64 (0.13)
	RICES	27.16 (0.01)	31.38 (0.24)	33.68 (0.18)	34.58 (0.25)	34.42 (0.19)
	MMICES	27.16 (0.01)	**38.54 (0.16)**	**39.53 (0.13)**	**39.31 (0.12)**	**37.22 (0.11)**
MSCOCO	Random	76.45 (0.65)	81.41 (0.19)	90.48 (0.35)	92.83 (0.66)	93.72 (0.61)
	RICES	76.45 (0.65)	89.25 (0.17)	96.60 (0.24)	102.70 (0.20)	105.14 (0.05)
	MMICES	76.45 (0.65)	**98.61 (0.17)**	**102.56 (0.13)**	**105.66 (0.04)**	**105.89 (0.21)**

Table 14: The performances of random selection, RICES, and MMICES on `OF-4B`. MMICES achieves the best performance in all settings on all datasets.

Dataset	Method	0-shot	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	45.55 (0.29)	47.74 (0.11)	46.20 (0.15)	44.01 (0.23)	46.33 (0.14)
	RICES	45.55 (0.29)	48.24 (0.08)	46.27 (0.12)	44.32 (0.13)	47.55 (0.12)
	MMICES	45.55 (0.29)	**49.03 (0.04)**	**48.22 (0.07)**	**47.42 (0.03)**	**48.85 (0.05)**
OK-VQA	Random	32.15 (0.21)	34.56 (0.31)	33.73 (0.27)	31.61 (0.15)	34.29 (0.62)
	RICES	32.15 (0.21)	34.86 (0.05)	34.40 (0.09)	32.52 (0.13)	36.73 (0.06)
	MMICES	32.15 (0.21)	**38.14 (0.07)**	**38.23 (0.16)**	**36.08 (0.09)**	**37.32 (0.14)**
GQA	Random	28.42 (0.07)	32.10 (0.23)	33.53 (0.32)	34.32 (0.25)	35.53 (0.29)
	RICES	28.42 (0.07)	32.59 (0.08)	34.51 (0.25)	35.19 (0.15)	37.07 (0.10)
	MMICES	28.42 (0.07)	**38.61 (0.09)**	**39.48 (0.16)**	**39.73 (0.13)**	**39.56 (0.06)**
MSCOCO	Random	80.30 (0.15)	85.97 (0.46)	91.71 (0.12)	96.70 (0.19)	98.06 (0.31)
	RICES	80.30 (0.15)	92.67 (0.08)	101.38 (0.15)	105.75 (0.13)	**108.22 (0.05)**
	MMICES	80.30 (0.15)	**100.59 (0.07)**	**105.16 (0.22)**	**108.08 (0.10)**	107.96 (0.20)

Table 15: The performances of random selection, RICES, and MMICES on `OF-4BI`. MMICES achieves the best performance in most cases.

Dataset	Method	0-shot	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	51.38 (0.17)	53.52 (0.11)	53.74 (0.19)	53.33 (0.26)	52.38 (0.10)
	RICES	51.38 (0.17)	**54.03 (0.13)**	**54.67 (0.06)**	**55.39 (0.12)**	**55.77 (0.08)**
	MMICES	51.38 (0.17)	53.11 (0.03)	53.56 (0.05)	54.04 (0.04)	55.14 (0.02)
OK-VQA	Random	37.62 (0.39)	39.62 (0.29)	41.56 (0.20)	43.40 (0.39)	42.97 (0.11)
	RICES	37.62 (0.39)	42.13 (0.13)	43.87 (0.15)	44.90 (0.10)	46.15 (0.06)
	MMICES	37.62 (0.39)	**44.18 (0.11)**	**45.61 (0.08)**	**46.93 (0.08)**	**46.79 (0.10)**
GQA	Random	34.04 (0.19)	36.32 (0.29)	37.74 (0.32)	38.28 (0.10)	37.85 (0.11)
	RICES	34.04 (0.19)	36.92 (0.33)	38.54 (0.14)	40.16 (0.14)	40.21 (0.32)
	MMICES	34.04 (0.19)	**40.73 (0.09)**	**41.85 (0.10)**	**42.21 (0.12)**	**42.07 (0.08)**
MSCOCO	Random	79.52 (0.31)	89.82 (0.23)	96.81 (0.10)	99.44 (0.19)	100.53 (0.26)
	RICES	79.52 (0.31)	93.45 (0.07)	99.74 (0.27)	105.76 (0.03)	109.12 (0.20)
	MMICES	79.52 (0.31)	**100.24 (0.20)**	**104.90 (0.3)**	**108.66 (0.17)**	**109.64 (0.24)**

Table 16: The performances of random selection, RICES, and MMICES on `OF-9B`. MMICES achieves the best performance in most cases.

Dataset	Method	0-shot	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	52.59 (0.30)	54.90 (0.05)	56.16 (0.02)	56.93 (0.18)	57.21 (0.17)
	RICES	52.59 (0.30)	54.79 (0.09)	56.45 (0.05)	57.49 (0.06)	58.52 (0.02)
	MMICES	52.59 (0.30)	**56.15 (0.01)**	**58.17 (0.03)**	**59.23 (0.01)**	**59.69 (0.02)**
OK-VQA	Random	44.77 (0.22)	49.24 (0.22)	49.54 (0.12)	50.89 (0.12)	51.86 (0.12)
	RICES	44.77 (0.22)	48.82 (0.02)	50.55 (0.05)	52.42 (0.03)	53.22 (0.04)
	MMICES	44.77 (0.22)	**49.63 (0.02)**	**52.16 (0.03)**	**53.65 (0.07)**	**54.16 (0.05)**
GQA	Random	36.45 (0.22)	39.35 (0.26)	40.54 (0.17)	41.38 (0.18)	41.87 (0.13)
	RICES	36.45 (0.22)	39.86 (0.13)	41.27 (0.29)	42.65 (0.21)	43.67 (0.19)
	MMICES	36.45 (0.22)	**42.66 (0.05)**	**44.22 (0.08)**	**45.19 (0.05)**	**45.36 (0.09)**
MSCOCO	Random	48.61 (0.52)	96.45 (0.36)	100.85 (0.36)	103.96 (0.38)	105.02 (0.43)
	RICES	48.61 (0.52)	91.20 (0.10)	102.58 (0.15)	108.93 (0.10	111.02 (0.08)
	MMICES	48.61 (0.52)	**101.13 (0.12)**	**109.31 (0.09)**	**112.72 (0.05)**	**113.37 (0.09)**

Table 17: The performances of random selection, RICES, and MMICES on `IDEFICS-9B`. MMICES achieves the best performance in all cases.

Dataset	Method	0-shot	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	62.99 (0.03)	63.94 (0.13)	64.43 (0.14)	64.64 (0.10)	64.87 (0.09)
	RICES	62.99 (0.03)	**64.13 (0.08)**	**64.69 (0.03)**	65.11 (0.05)	65.22 (0.03)
	MMICES	62.99 (0.03)	63.51 (0.13)	64.46 (0.04)	**65.26 (0.04)**	**65.50 (0.02)**
OK-VQA	Random	46.18 (0.17)	48.78 (0.48)	49.92 (0.16)	51.18 (0.20)	51.41 (0.12)
	RICES	46.18 (0.17)	49.80 (0.03)	51.32 (0.02)	52.42 (0.05)	53.35 (0.03)
	MMICES	46.18 (0.17)	**51.65 (0.08)**	**53.21 (0.03)**	**53.89 (0.03)**	**54.14 (0.01)**
GQA	Random	41.83 (0.21)	43.99 (0.20)	45.70 (0.16)	46.39 (0.08)	46.89 (0.17)
	RICES	41.83 (0.21)	44.79 (0.18)	45.63 (0.07)	46.57 (0.16)	46.82 (0.06)
	MMICES	41.83 (0.21)	**46.33 (0.12)**	**47.51 (0.09)**	**47.87 (0.13)**	**48.47 (0.11)**
MSCOCO	Random	124.15 (0.63)	**132.80 (0.63)**	**133.02 (0.39)**	**132.23 (0.37)**	**132.93 (0.32)**
	RICES	124.15 (0.63)	124.97 (0.11)	126.84 (0.10)	127.85 (0.10)	128.76 (0.08)
	MMICES	124.15 (0.63)	125.42 (0.12)	128.50 (0.09)	129.71 (0.06)	130.55 (0.09)

Table 18: The performances of random selection, RICES, and MMICES on `IDEFICS-9BI`. MMICES achieves the best performance in most cases.

Dataset	K	4-shot	8-shot	16-shot	32-shot
GQA	50	39.43	40.50	40.99	40.48
	100	40.72	41.15	41.89	41.09
	200	40.73	41.85	42.21	42.07
	300	40.76	41.63	42.28	42.20
OK-VQA	50	43.46	45.79	47.48	47.21
	100	43.40	45.72	46.50	47.17
	200	44.18	45.61	46.93	46.79
	300	44.21	45.66	46.00	46.79

Table 19: Performance of MMICES given different K.

bers is shown in Tab. 20. MMICES achieves the best performance when using generated captions based on the 4-shot setting.

Different Choice of Modality Mixture. Compared to RICES, which only compares image similarity, MMICES considers both visual and language modalities. We also investigate the performance of ICL when examples are retrieved using only text similarity (referred to as *text*), and when retrieved by first comparing language and then selecting based on image similarity (referred to as *text-image*). Full results are presented in Table 21. Factoring in both modalities consistently improves ICL performance compared to selecting based solely on one modality.

ICL Setting	4-shot	8-shot	16-shot	32-shot
Random	89.82	96.81	99.44	100.53
RICES	93.45	99.74	105.76	109.12
MMICES given Random				
0-shot	95.31	100.53	105.06	107.90
4-shot	97.72	102.81	107.37	110.15
8-shot	99.90	104.95	108.20	110.31
16-shot	100.08	104.82	109.11	110.26
32-shot	100.24	104.90	108.66	109.64

Table 20: MMICES on MSCOCO with generated captions from ICL with randomly selected demonstrations. Based on results with 0-shot, MMICES obtain better results in r-shot and 8-shot settings. Given generated captions with 4-shot, MMICES achieves the best results in all settings.

Data	Method	4-shot	8-shot	16-shot	32-shot
VQAv2	Random	53.52	53.74	53.33	52.38
	RICES	54.03	54.67	55.39	55.77
	text	47.71	47.46	47.49	47.83
	text-image	50.27	50.37	49.84	50.56
	MMICES	53.11	53.56	54.04	55.14
OK-VQA	Random	39.62	41.56	43.40	42.97
	RICES	42.13	43.87	44.90	46.15
	text	42.80	43.54	44.01	44.07
	text-image	43.61	45.53	45.01	45.50
	MMICES	44.18	45.61	46.93	46.79
GQA	Random	36.32	37.74	38.28	37.85
	RICES	36.92	38.54	40.16	40.21
	text	39.18	40.68	41.59	41.58
	text-image	40.93	42.12	42.70	42.63
	MMICES	40.73	41.85	42.21	42.07
COCO	Random	89.82	96.81	99.44	100.53
	RICES	93.45	99.74	105.76	109.12
	text	99.84	102.88	105.57	106.52
	text-image	100.72	104.93	106.97	108.56
	MMICES	100.24	104.90	108.66	109.64

Table 21: Performance with different modality mixture. RICES compares image similarity. *text* only considers text similarity. *text-image* selects demonstrations by first comparing language similarity and then comparing image similarity.

F Additional Experimental Analysis

This study has conducted extensive experiments on various vision-language models, using different sizes, backbone language models, and pre-training datasets (as shown in Tab. 2). This section further discusses our observations and findings for these different models.

Experiments across models with different sizes. The ICL performance of different sizes of OpenFlamingo models is presented in Fig. 9 to Fig. 11. MMICES consistently improves the ICL performance on these datasets across various model sizes. Larger models, such as `OF-9B`, demonstrate better performance compared to smaller models, particularly in visual question answering (Fig. 9) and visual reasoning (Fig. 10). It is worth noting that MMICES achieves better performance on smaller-size models compared to larger-size models using RICES and random selection, especially in the 4 and 8-shot settings.

Experiments across different models. The performance gained from MMICES is consistent across different models, as shown in Fig. 12 to Fig. 14. IDEFICS achieves better performance compared to OpenFlamingo, and this difference can be attributed to the use of different pre-training datasets and language models in these two models Laurençon et al. (2023).

Impact of Underlying Vision-language Models. This study has also conducted extensive experiments on different vision-language models with varying sizes. Fig. 16 presents a performance comparison on the GQA dataset across models from `OF-3B` to `OF-9B`. MMICES consistently outperforms random selection and 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. Moreover, the performance gained from MMICES is consistent across different models as shown in Fig. 17.

Ablation Study. The number of pre-filtered samples, *i.e.*, K defined in Alg. 1, selected by visual similarity is a hyperparameter in MMICES. Besides different K, as MMICES considers both visual and language modalities, we also investigate the ICL performance when the examples are retrieved

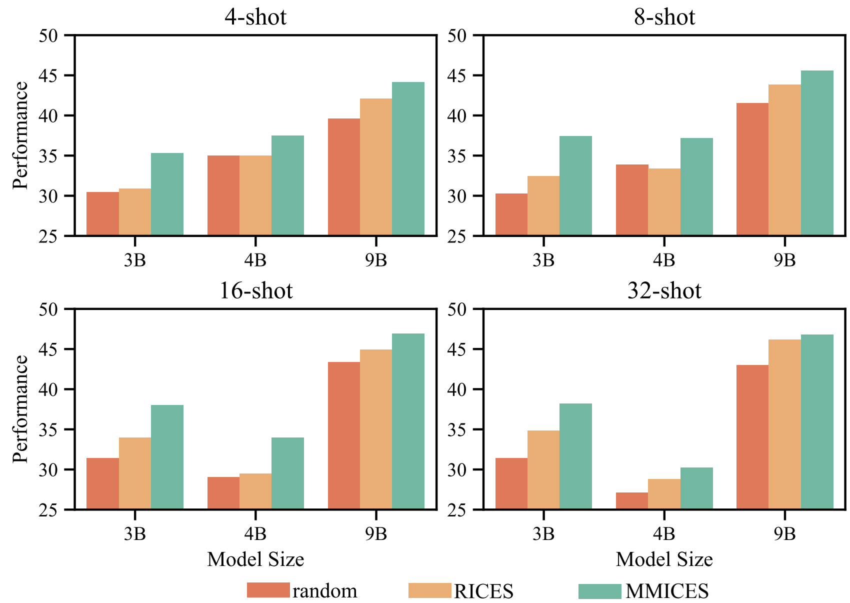


Figure 9: The performance of ICL (on OK-VQA) is consistently enhanced by MMICES on Open-Flamingo with different sizes.

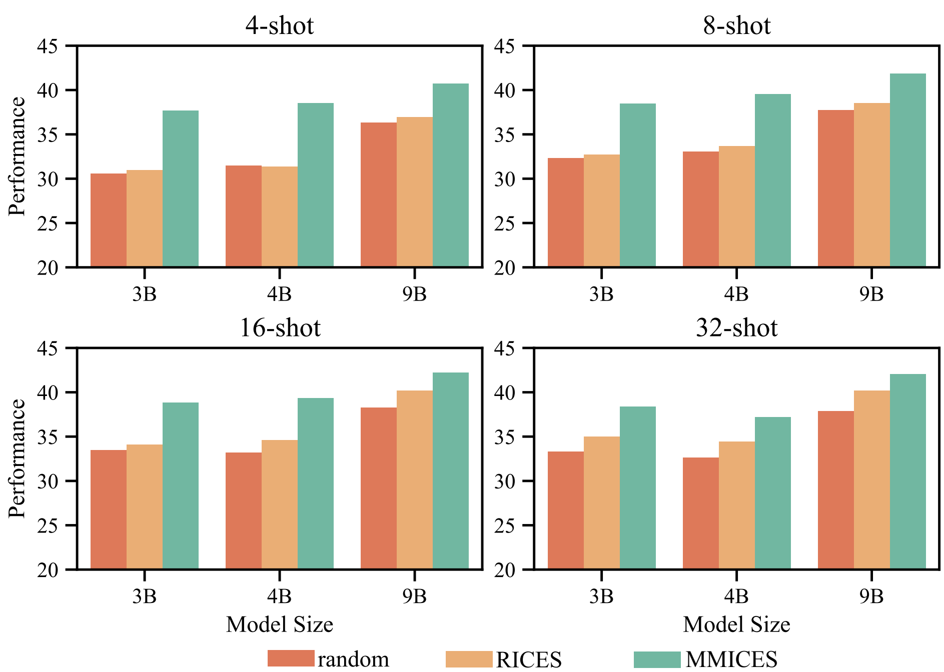


Figure 10: The performance of ICL (on GQA) is consistently enhanced by MMICES on Open-Flamingo with different sizes.

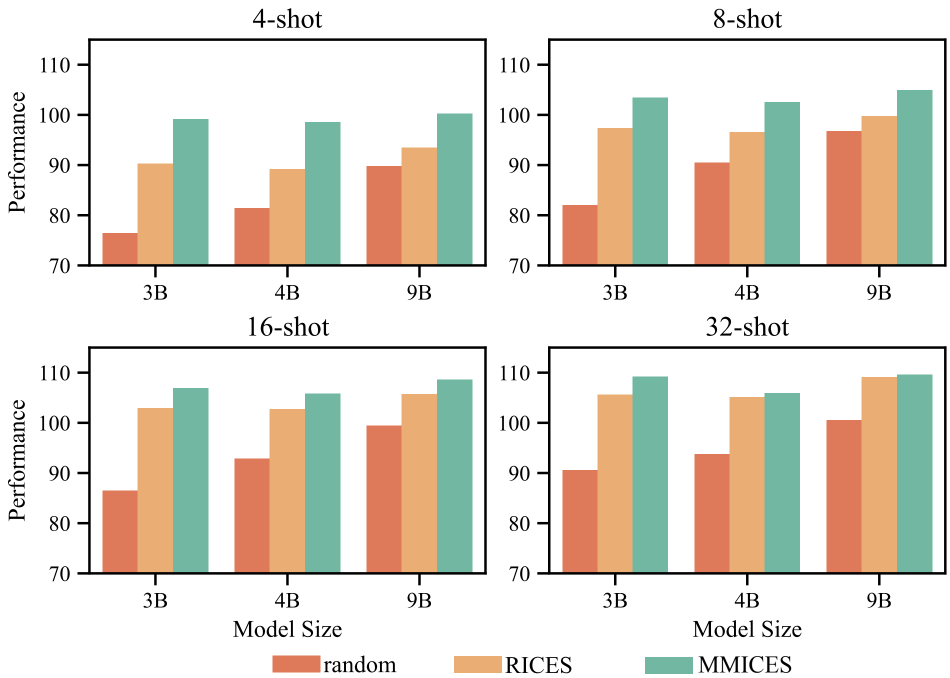


Figure 11: The performance of ICL (on COCO) is consistently enhanced by MMICES on Open-Flamingo with different sizes.

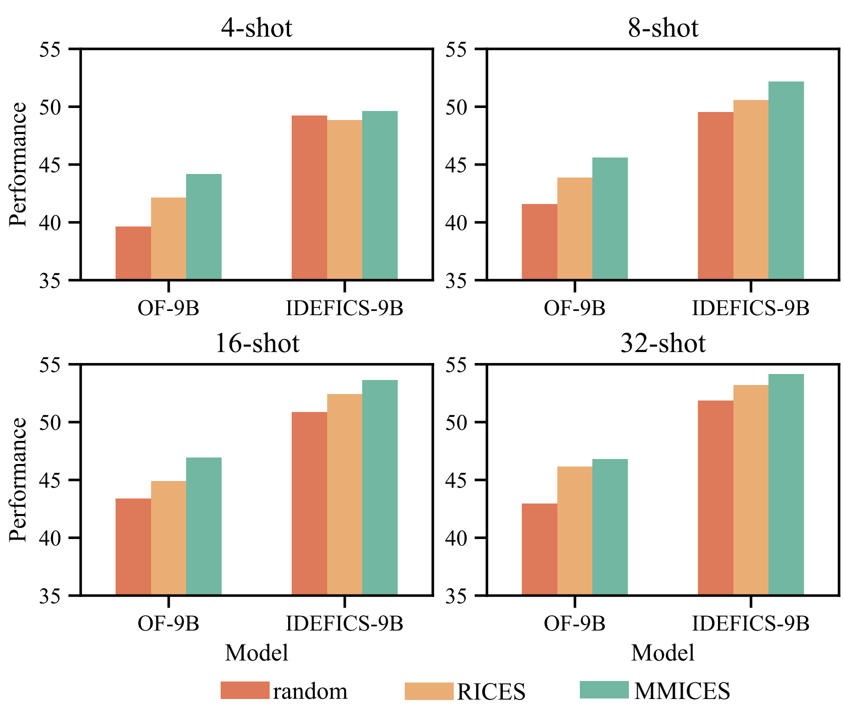


Figure 12: The performance of ICL (on OK-VQA) is consistently enhanced by MMICES across different models.

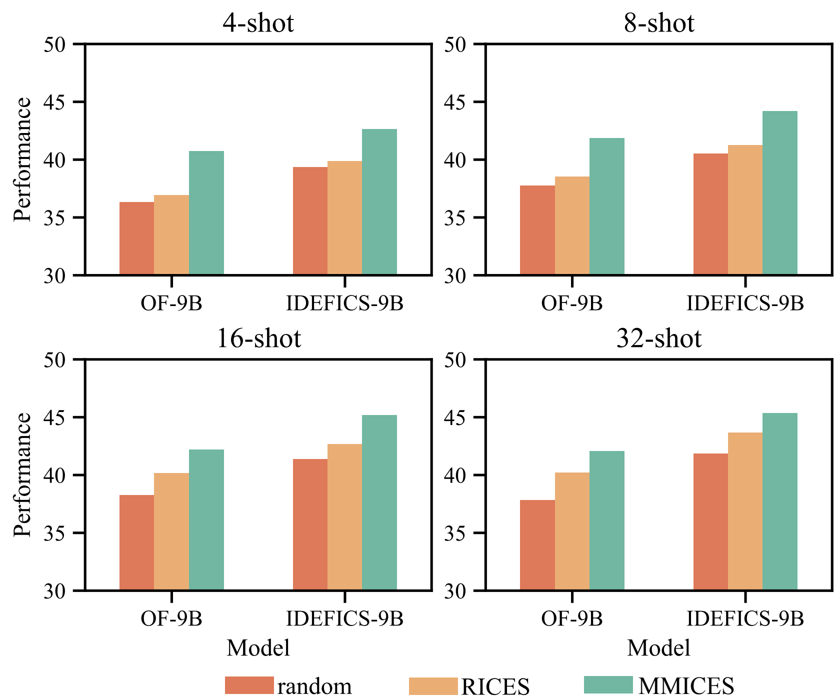


Figure 13: The performance of ICL (on GQA) is consistently enhanced by MMICES across different models.

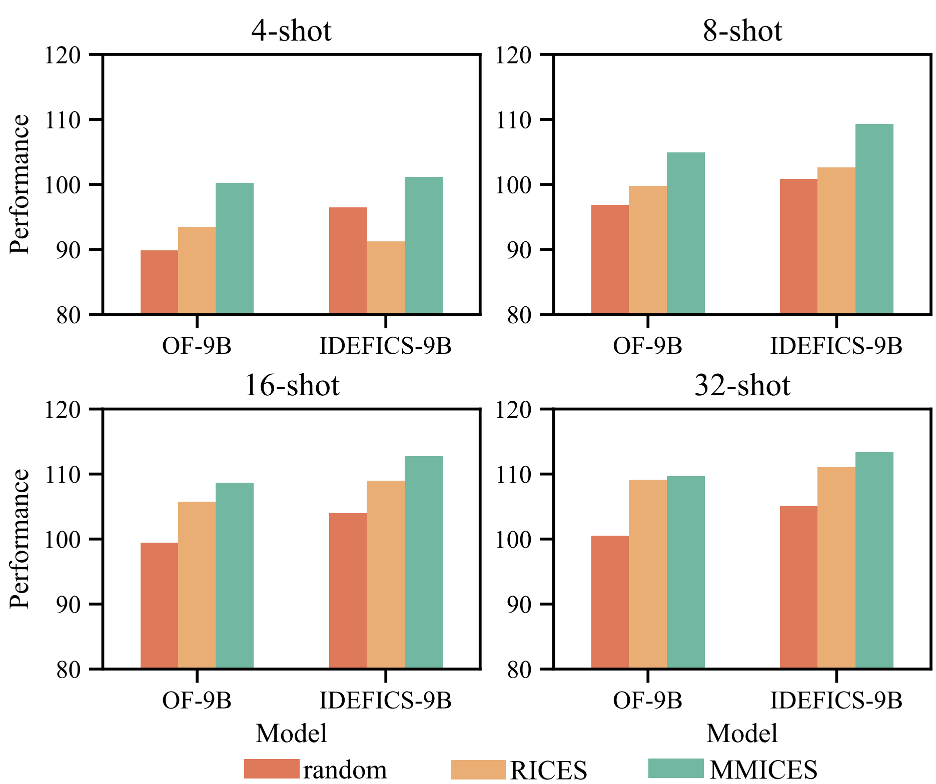


Figure 14: The performance of ICL (on COCO) is consistently enhanced by MMICES across different models.

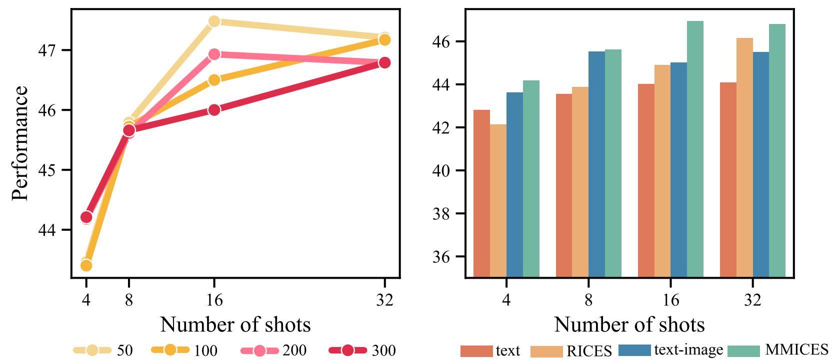


Figure 15: Comparison of performance on OK-VQA given different K (left) and different mixture of modality (right).

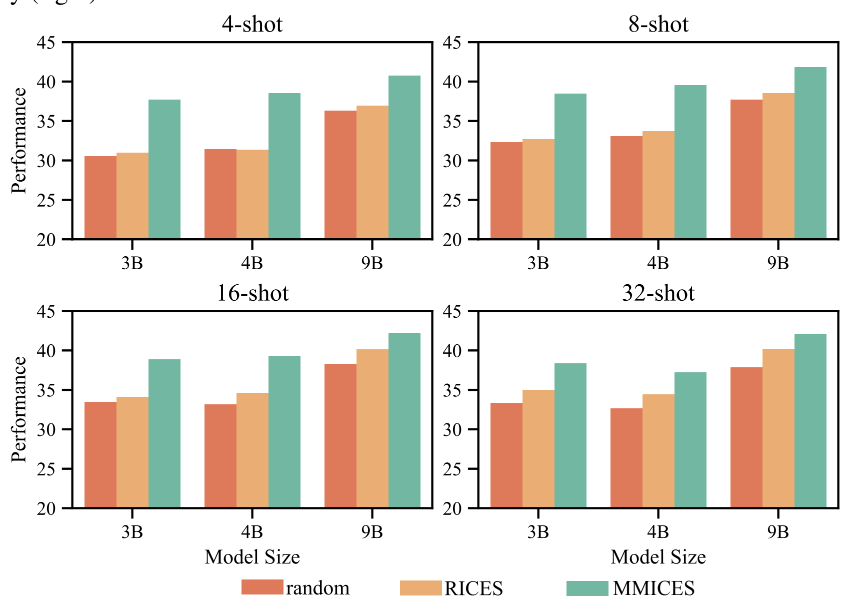


Figure 16: MMICES consistently enhances the ICL performance across models of varying sizes. MMICES on smaller models can even outperform RICES on larger models. Results here are from GQA and more results are in Appendix. E.

only by text similarity (termed as *text*), and when retrieved by first comparing language and then selecting based on image similarity (termed as *text-image*). Fig. 15 shows the performance comparison on OK-VQA. A larger K leads to more candidate demonstrations filtered by visual similarity and is more useful when the number of shots is small. Regarding the mixture of modalities, the results are consistent with our analysis in Sec. 3. Retrieval based on a single modality, such as RICES on visual and *text*, underperforms mixed modality retrieval. Besides, MMICES consistently achieves better results compared to *text-image* (more analysis in Appendix. F).

Query	Method	Demo 1	Demo 2	Demo 3	Model Generation
Who makes the guitar on the wall?	MMICES	Who makes the luggage in this room? samsonite	Who invented the device pictured? steve job	Who manufactures this bag? ll bean	fender
	RICES	What kind of suitcase is this? carry on	Where can you buy these luggages? walmart	What items would you typically find in these bags? cloth	yamaha
Name the material used to make this umbrella shown in this picture?	MMICES	What material are the umbrellas made of? straw	What is the pattern on the umbrella? striped	What do you call this type of window covering? blind	plastic
	RICES	What causes high and low tides? moon	What is the orange triangle in the road called? cone	When is it bad luck to open the black and pink object in the photo? inside	rubber
What were they fixing?	MMICES	What happened here? accident	What safety precaution did both of these people take? helmet	What is being done on this road? construction	power line
	RICES	What purpose does the white and red striped bar in the picture serve? stop traffic	Is this person crossing illegally or legally? legally	What is the job title for the man pictured here? electrician	light pole
What food item do you think this ornament resembles?	MMICES	What food is this? carrot cake	What food is this? cake	What food is this? candy apple	donut
	RICES	What can this make you become if you eat a lot of it? fat	What type of computer is shown in this image? desktop	What are donuts topped with? ice	cookie
What is the purpose of the elephant here?	MMICES	What is the elephant doing? paint	Why does the elephant go to the water? thirsty	Why are they riding an elephant? for fun	decoration
	RICES	What country was this photograph taken in? thailand	How tall do these animals typically grow to be? 11 feet	When was this type of vehicle with two equal sized wheels invented? 1850	park meter
What color is the taxi?	MMICES	What is the name of the body style of the grey vehicle? minivan	What make and model is the car pictured? toyota avalon	What liquid makes the vehicle in the picture move? gasoline	yellow
	RICES	What is the use of that pink object over her head? keep dry	What photo technique is being used? sephia	Who invented the blue item in this picture? samuel fox	black
Name a metal shown?	MMICES	What is the silver tool called? tong	What type of jewelry uses a term similar to one of these veggies? carrot	Which of these items depicted grows underground? potato	stainless steel
	RICES	When would i eat this? dinner	How is the the meat in this dish prepared? grilled	What food group is mostly represented? meat	copper
What do these animals eat?	MMICES	What do these animals eat? plant	What do you feed these animals? hay	What is a staple of the diet of these animals? fish	grass
	RICES	What type of food does this animal eat? berry	What is a staple of the diet of these animals? fish	what do these animals do in the winter? hibernate	berry

Table 22: Examples of demonstrations selected by MMICES and RICES on OK-VQA. Model generations in green are correct and red means wrong prediction.

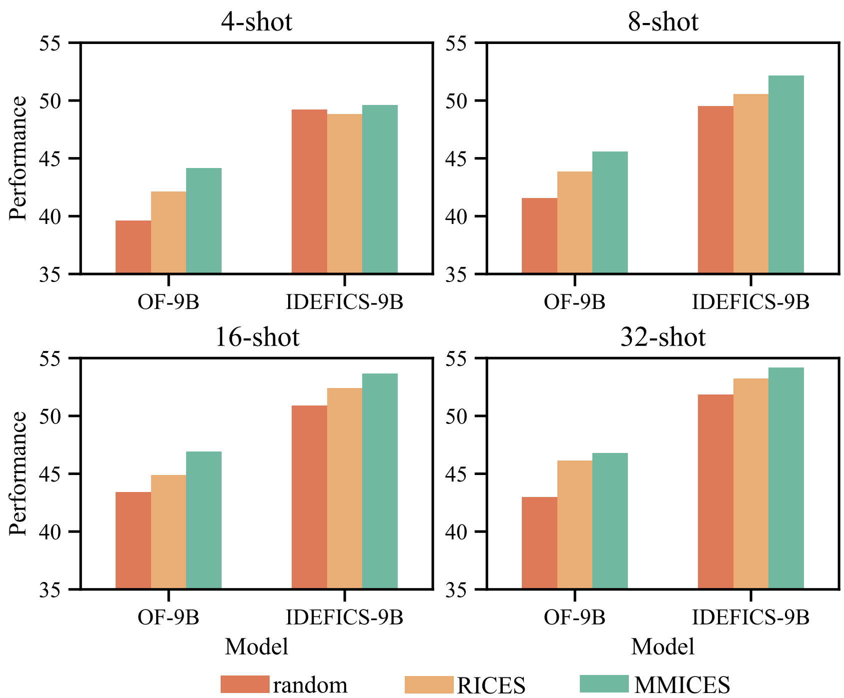


Figure 17: The performance of ICL (on OK-VQA) is consistently enhanced by MMICES across different models.