What Makes Good In-context Demonstrations in Multimodal Large Language Model?

Anonymous ACL submission

Abstract

Recently, multimodal large language models (MLLM) are beginning to exhibit the capability of in-context learning (ICL), enabling them to learn a new task by conditioning solely on some in-context examples, without updating the model parameters. However, existing studies on MLLM often randomly sample a subset of in-context examples and then order these examples randomly. It is still unclear what makes good in-context demonstrations in MLLM. In this paper, we empirically explore the impact of two key factors on the performance of ICL in MLLM to fill this gap: the selection and the order of demonstration examples. We conduct extensive experiments on three multimodal tasks including VQA, image captioning and multimodal image-text classification. Our experimental results show that the above two factors dramatically impact the performance of ICL. Additionally, we summarize our findings and provide takeaway suggestions on how to construct effective demonstrations in MLLM.

1 Introduction

One of the most surprising behaviors observed in foundation models is in-context learning (ICL; Brown et al. (2020)). ICL is an ability of a foundation model to condition on a prompt sequence consisting of in-context examples (input-output pairs from some task) along with a new query input, and generate the corresponding output. Notably, ICL is a post-training approach and does not require backward gradients and parameter updates, which makes it the most popular strategy for interacting with LLMs (Oniani and Wang, 2023).

The research on ICL starts from Brown et al. (2020), which shows that the LLM such as GPT-3 can condition on a list of input-output pairs (demonstration) to learn a new task. From then on, there are more and more studies devoted on this topic (Dong et al., 2023; Dai et al., 2023; Yang et al., 2023). Among these studies, some researchers find that the performance of ICL in LLMs is sensitive to the selection of in-context examples (Liu et al., 2022) , and order of examples Lu et al. (2022).

Recently, several MLLMs, such as CM3 (Aghajanyan et al., 2022), Flamingo (Awadalla et al., 2023), Kosmos-1 (Huang et al., 2023), PaLM-E (Driess et al., 2023), and multimodal GPT-4 (OpenAI, 2023), have demonstrated ICL capabilities akin to LLMs. Meanwhile, theses studies have highlighted the significance of ICL in enhancing the performance of MLLMs across various multimodal tasks. However, these existing studies on MLLMs often randomly sample a subset of in-context examples from a pool of training examples and then order these examples randomly to construct the demonstration prompt. This leads to a question: "Is random selection and random arrangement appropriate in MLLM?". In other word, a fundamental question "*What Makes Good In-context Demonstrations in MLLM?*" is still unexplored.

To comprehensively address this fundamental question, we analyze the design space of in-context demonstrations, and mainly pay attention to two aspects of in-context demonstrations in this paper: the selection and the order. To this end, we conduct an experimental study on three popular multimodal tasks including visual question answering, image captioning and multimodal image-text classification. Specifically, we mainly investigate the following two research questions (RQs): (1) What kind of selection methods are useful for ICL in MLLM? (2) How should demonstration examples be arranged for ICL in MLLM? To answer the RQ1, we compare a wide range of demonstration selection methods, such as random selection, similarity based selection, and diversity based selection. Besides, We also analyze the impact of different retrieval patterns for ICL in MLLM. To answer the RQ2, we compare random ordering with two ordering methods, namely similarity and similarity

reverse, towards investigating the impact of different ordering approach.

From the experimental results, we have the following key findings: (1) Random demonstration selection is a strong baseline among different model sizes ranged from 3B to 9B; (2) Among different retrieval pattern in demonstration selection, using image as the pivot could be a better strategy rather leveraging text information or multimodal information. (3) In the demonstration arrangement, employing similarity ordering or similarity reverse ordering in practice is a better strategy.

2 Preliminary

In-context learning only requires a list of input-output pairs to solve a task. Formally, we can denote a demonstration as $\text{demo}_i = (x_i, y_i)$, where x_i is the input instance and y_i is the output label. For a new test instance x_{test}, its corresponding label y_{predict} is generated via a given MLLM as follows in K-shot ICL:

$$\text{MLLM}(y_{\text{predict}}|\text{demo}_1, \text{demo}_2, \cdots, \text{demo}_K, x_{\text{test}}) \quad (1)$$

In this paper, we further clarify that there are two types of demonstration in ICL: **task-level demonstration** and **instance-level demonstration**. The task-level demonstration uses the same demonstration examples for all test samples and does not take the difference of each test sample into consideration, while the instance-level demonstration selects different demonstration examples for different test samples.

3 Experimental Setup

Evaluation Tasks In this paper, our evaluation tasks include VQA, image captioning and multimodal image-text classification. For the VQA task, we use VQA v2.0 (Agrawal et al., 2016),OK-VQA (Marino et al., 2019) and TextVQA (Singh et al., 2019) as the testbed and use the VQA accuarcy (Agrawal et al., 2016) as the metric. For the image captioning task, we choose Flickr30K (Plummer et al., 2015) dataset as our benchmark and employ CIDEr score (Vedantam et al., 2015) to evaluate models. For the multimodal image-text classification task, our evaluation dateset is Hateful Memes (Kiela et al., 2020) and we report AUC ROC.

Foundation Multimodal Model In this paper, we select OpenFlamingo (Awadalla et al., 2023) as our foundation model. OpenFlamingo combines a pretrained vision encoder and a language model using cross attention layers. In OpenFlamingo, different size models share CLIP ViT-L/14 (Radford et al., 2021) as the vision encoder, while the language model can be chosen from MPT-1B (MosaicML, 2023), RedPajama-3B (Together.AI, 2023) and MPT-7B (MosaicML, 2023), which correspond to OpenFlamingo of 3B, 4B, and 9B.

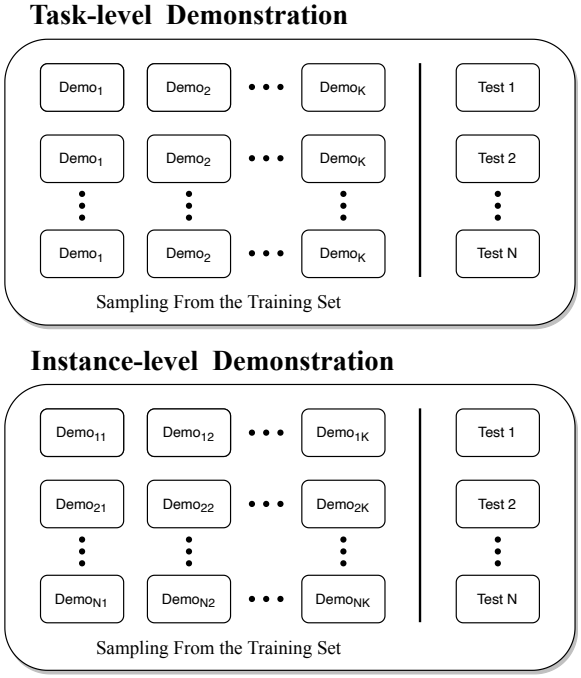

Figure 1: Illustration of task-level demonstration and instance-level demonstration.

Task-level Demonstration Selection Method For the task-level demonstration selection, we need to select a group of demonstration examples for the whole test set. There are two methods we used in task-level demonstration selection, which are **Herding** (Welling, 2009) and **K-centering** (Sener and Savarese, 2018).

Herding: The Herding method selects data points based on the distance between the demonstration set center and original dataset center in the feature space. The method incrementally and greedily adds one sample each time into the demonstration set that can minimize distance between two centers. Herding is a similarity based selection method. **K-centering**: Different from computing a single center in Herding, K-centering selects the training examples that are maximally separated, in other words, K-centering is a diversity based selec-

Method	Pattern	#Param	VQAv2	OK-VQA	Text-VQA	Flickr30k	Hateful Memes
Herding	M-M	3B	0.09	3.18	4.53	-16.13	1.01
		4B	-0.18	0.88	3.84	-28.51	4.13
		9B	0.94	-0.91	7.41	6.22	4.03
	T-T	3B	-0.91	3.01	4.52	-16.40	1.49
		4B	-3.23	1.28	3.84	-29.12	0.96
		9B	-1.10	0.75	6.90	9.02	4.89
	I-I	3B	-0.31	2.12	4.00	5.78	1.17
		4B	-1.70	1.27	5.47	0.35	3.93
		9B	1.20	-0.48	7.53	-1.20	5.42
K-Centering	M-M	3B	0.29	3.17	3.60	-3.79	1.24
		4B	-4.69	1.13	4.41	-5.43	3.89
		9B	-4.39	1.16	6.35	-7.00	4.20
	I-I	3B	-0.26	3.31	2.13	2.78	1.90
		4B	-1.42	-5.22	1.47	2.91	3.42
		9B	0.17	0.33	6.35	3.11	4.63
	T-T	3B	0.09	2.57	3.67	-5.65	1.72
		4B	-2.82	0.39	4.20	-5.11	2.72
		9B	-0.93	0.00	6.96	-3.91	4.28
RICES	I-I	3B	1.63	4.26	2.32	-10.78	15.08
		4B	0.54	1.62	1.83	-6.94	10.48
		9B	1.63	2.13	5.98	-8.43	17.01
	T-T	3B	-0.07	4.50	1.01	*	12.34
		4B	-4.02	2.27	-0.26	*	6.04
		9B	-5.51	-0.10	3.62	*	16.68
	I-T	3B	-1.74	3.23	0.55	-0.08	8.10
		4B	-3.08	1.63	-0.05	-4.41	1.46
		9B	-2.62	0.31	6.05	-5.83	11.08
	T-I	3B	-1.34	2.53	0.99	*	4.94
		4B	-2.91	2.27	-0.28	*	-0.86
		9B	-1.17	-0.06	3.61	*	10.49
	M-M	3B	1.14	6.60	0.84	*	19.56
		4B	0.28	3.78	-0.74	*	12.24
		9B	-3.08	2.49	4.00	*	21.60

Table 1: Experimental results on the demonstration selection methods. Pattern denotes retrieval pattern, and #Param denotes the number of model parameters. To obtain this table, we conduct experiments for each model and each retrieval pattern in the 0, 4, 8, 16 and 32 shot setting, then we average these results in various shot setting and only report the difference values between these methods and the random selection method for presenting these results more clearly. The details of these experiment can be find in the Appendix. Note that for RICES on the image captioning task (Flickr30k dataset), we can not use the text (caption) in test example as the key to retrieval the whole training set.

tion method. In detail, We randomly select a single sample as the initial demonstration set, and then we add a new sample that is furthest in Euclidean distance within feature space from the nearest sample in the demonstration set until selecting K demonstrations.

Instance-level Demonstration Selection Method For the instance-level demonstration selection, we need to select different demonstration examples for different test samples. In this paper, we employ the Retrieval-based In-Context Example Selection (RICES) method (Yang et al., 2022) to this end, which is still a similarity based method. In detail, RICES selects the top-K most similar training examples as demonstrations, with similarity being determined by cosine similarity in the feature space.

Demonstration Ordering Method In this paper, we utilize three ordering methods: random ordering, similarity based ordering, and reverse similarity based ordering. As for the similarity based ordering, the demonstration with higher similarity to the test example are placed closer to it. Conversely, in the reverse similarity based ordering, the demon-

Method	VQAv2	OK-VQA	Text-VQA	Flickr30K	Hateful Memes
Random	51.29	41.93	34.19	54.52	**70.09**
Similarity	**54.08**	41.82	**34.79**	52.64	69.71
Similarity reverse	53.87	**42.25**	34.28	**55.60**	69.60

Table 2: Experimental results of different demonstration ordering methods.

stration examples are arranged in descending order based on their similarity to the test sample.

Feature Space of Demonstration Selection and Ordering To project both text and images into a shared feature space, we utilize CLIP ViT-L/14 to extract image features and employ the text encoder in CLIP to extract text features. After that, we design five distinct retrieval pattern : Image-to-Image (I-I), Image-to-Text (I-T), Text-to-Image (T-I), Text-to-Text (T-T) and Multimodal-to-Multimodal (M-M) [1], where the first element represents the query feature, while the second element represents the key feature.

4 Results on Demonstration Selection

We present the experimental results on demonstration selection methods in the Table 1. Based on theses results, we find that:

Random demonstration selection is a strong baseline among different model sizes and no human-designed selection methods mentioned in this paper can consistently surpass it. As shown in Table 1, we find that in most datasets, involving sample selection and random selection yields mixed results. However, there are exceptions, on Hateful Memes, introducing sample selection, whether the task-level or the instance level demonstration selection methods, can significantly outperforms random selection. This result is consistent with the studies in LLM (Rubin et al., 2022; Li et al., 2023), which show that classification tasks are more likely to benefit from sample selection.

Among different retrieval pattern, Image-to-Image (I-I) performs quite well in most cases. We find that except for the Flickr30K dataset, Image-to-Image retrieval pattern demonstrated good performance across all other datasets, achieving the best scores in VQAv2, OK-VQA, and Text-VQA. The Multimodal-to-Multimodal (M-M) retrieval pattern also shows a similar trend of Image-Image pairs. We speculate that the good performance of M-M pattern is primarily due to the influence of I-I pattern.

[1]We concatenate the text feature and image feature as the multimodal feature

5 Results on Demonstration Arrangement

We present the experimental result on demonstration arrangement in the Table 2. From the table, we can find that: (1) Demonstration arrangement can dramatically impact the performance of ICL. Take VQAv2 as a example, great performance boosts as we move from random arrangement to similarity based arrangement. (2) Across datasets, there is no shared knowledge on demonstration arrangement to draw upon. Even within a single task, like VQA, none of the ordering methods mentioned in this paper can consistently outperforms the others. (3) Unlike random selection serves as a strong baseline in demonstration selection, random arrangement is not a wise option in practice. In the Hateful Memes, random arrangement can slightly outperform than similarity based ordering and reverse similarity based ordering, but in the other testbeds, random arrangement usually performs poorly.

6 Discussion and Conclusion

Our research presents a timely investigation into an emerging capability termed ICL for the MLLM. We systematically explore the influence of in-context examples on downstream performance, revealing a crucial insight: the effectiveness of ICL is intricately tied to the design of demonstrations. Surprisingly, there is no apparent methodology in which ICL consistently improves performance across all tasks for the MLLM. Our study also uncovers intriguing and anomalous phenomena. Notably, we demonstrate that, unlike ICL in LLMs, where a good in-context example ought to exhibit semantic similarity to the test example, however, the ICL in MLLM presents a divergent scenario. We further identify that different retrieval pattern exert varying impacts on distinct tasks, and certain retrieval pattern may even detrimentally affect MLLM performance. This observation prompts a critical conclusion: the conventional approaches in ICL may be unreliable in MLLM.

Limitations

The choice of MLLMs. Due to the unavailability of open-access resources for the most contemporary MLLM, our experiments is confined to employing OpenFlamingo. As a result, we can not definitively ascertain whether the ICL capabilities would exhibit variation in other MLLMs which employ distinct methods for incorporating textual and visual information.

The choice of testbeds. We have selected five datasets that encompass the three primary tasks of MLLM. For both the image captioning task and the multimodal image-text classification tasks, we only utilized a single dataset each. The varying results across these diverse datasets indicate a need for a broader range of datasets to fully understand the impact of ICL. Therefore, in our future work, we intend to incorporate additional datasets as testbeds to delve deeper into the impact of ICL.

References

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. 2022. Cm3: A causal masked multimodal model of the internet.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4005–4019.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

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

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. 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. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge.

MosaicML. 2023. Introducing mpt-7b: a new standard for open-source, commercially usable llms.

David Oniani and Yanshan Wang. 2023. In-context learning functions with varying number of minima.

OpenAI. 2023. Gpt-4 technical report.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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

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

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: A core-set approach.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read.

Together.AI. 2023. Releasing 3b and 7b redpajama-maincite family of models including base, instruction-tuned chat models.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation.

Max Welling. 2009. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. 2023. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5443–5456, Singapore. Association for Computational Linguistics.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa.