Unlock the Power of Cross-modality In-context Learning for Large Vision-language Models

Anonymous ACL submission

Abstract

While state-of-the-art LVLMs are increasingly capable of handling interleaved multi-image inputs and in-context learning (ICL), the key to fully leveraging these models lies in the careful construction of demonstrations and their integration into effective prompts. However, there is still a lack of strategies informed by a theoretical understanding of LVLMs' ICL processes. To address this gap, we first break down the ICL process into two core components: Task Recognition (TR) and Task Learning (TL) and further refine these components based on the nature of cross-modal interactions. This analysis reveals the crucial role of modality fusion in cross-modal ICL. By applying this concept, we develop a new, generalized demonstration configuration, leading to deeper insights into TR and TL. These insights uncover implicit patterns and tendencies within LVLMs during ICL and underscore the importance of modality fusion. Based on these findings, we design a complete prompt construction process centered on modality fusion, and demonstrate that our approach significantly enhances ICL performance across various vision-language tasks and LVLMs.

1 Introduction

ICL allows LLMs to learn and adapt to new tasks by providing a series of external knowledge, known as demonstrations, without requiring any model parameter updates(Laurençon et al., 2024a)(Zhao et al., 2022)(Alayrac et al., 2022). Simultaneously, as expectations for LLMs to handle increasingly complicated tasks rise, these models have gradually expanded beyond language into new modalities, such as vision, audio, and video (Laurençon et al., 2024b)(Awadalla et al., 2023b)(Zhao et al., 2023)(Peng et al., 2023). The rise of these Multimodal Large Language Models (MLLMs), exemplified by Large Vision-Language Models (LVLMs), follows a similar development trend as LLMs, with the use of ICL becoming a prominent feature(Yang et al., 2024).

However, despite the strong few-shot learning capabilities of SOTA LVLMs, such as InternVL2(Chen et al., 2024) and Qwen2VL(Wang et al., 2024), the construction of demonstrations and prompts still significantly affects their performance on specific tasks as the behavioral logic of these models can be easily influenced by external instructions(Awadalla et al., 2023c)(Laurençon et al., 2024c). While the internal mechanisms of ICL in LLMs have been widely studied and effectively divided into two stages—Task Recognition and Task Learning—the analysis of how these models process interleaved image-text inputs in the multimodal domain remains underexplored(Liu et al., 2024)(Sun et al., 2024).

Our goal is to develop a more systematic understanding of LVLM's ICL based on modality fusion(Zhu et al., 2023), and to leverage this understanding to create a generalized demonstration framework, identify effective retrieval and ordering methods, and optimize interleaved image-text prompts(Laurençon et al., 2024d)(Chen et al., 2023)(Doveh et al., 2024). First, we transfer the concepts of TR and TL to the multimodal domain and refined them for LVLMs. Using these insights, we introduce a new component, **query**, into traditional demonstration construction to enhance modality coordination. We build (I, Q, R) triplets applicable to three different vision-language tasks. We then systematically analyzed the roles of TR and TL in LVLMs based on this new configuration and found that TR is crucial for well-trained LVLMs, with the query playing a central role in this process. Building on this analysis, we designed a retrieval method that fuses both image and query features, creating a cross-modal prompt structure. Furthermore, we further take advantage of modality fusion by proposing demonstration reordering techniques and constructing a special in-

struction $Bridging$ for more efficient prompt optimization. Our methods achieved superior performance across six datasets spanning three different tasks with three different SOTA LVLMs. The complete pipeline of ICL using LVLM and the internal mechanism of the model during this process is shown in Appendix Figure3. In summary, our contributions are:

- We refine the concepts of Task Recognition and Task Learning from LLMs to better align with the nature of interleaved image-text prompts and LVLMs.
- We create a new demonstration configuration that allows us to further investigate the inner workings of LVLM's in-context learning.
- We propose a complete and effective cross-modal prompt construction process, which includes both the retrieval of demonstrations and the optimization of the prompt structure.

2 How do LVLMs learn in-context?

2.1 Vision-language TR and TL

Given that non-parametric external knowledge is embedded into LLMs through prompts, the in-context learning process can be systematically divided into two sub-processes: Task Recognition (TR) and Task Learning (TL) (Pan et al., 2023).

We further split TR for LVLMs into two processes: **framework construction** and **cross-modal adapting**. Compared to pure text input, TR builds a finer framework for multimodal input to integrate knowledge. In this phase, LVLM scans the structure of all input demonstrations to build a task-specific framework that encapsulates task format, input distribution and label space. Cross-modal adapting identifies key attributes of the input from each modality and then uses the constructed framework to narrow down the potential answer space using information from both modalities to better adapt to it. TL fills external multimodal knowledge into the framework and try to deepen LVLM's understanding of the mappings. During this phase, the model discerns innovative logic patterns for generating the desired results through implicit learning. We can use prompts to further guide it in selecting the optimal pattern when processing the current input.

2.2 Configuring Demonstrations

Demonstration libraries for LVLMs in ICL are typically composed of multiple (I, T) pairs, where I denotes the image and T denotes the overall textual information required for different tasks. This structure, restricted by the dataset's original form, may weaken the LVLM's ability to perform both TR and TL. Given a well-trained LVLM with multi-graph reasoning capability, we set an alternating versatile n-shot in-context sequence format $S = \{(I_1, Q_1, R_1), (I_2, Q_2, R_2), ..., (I_n, Q_n, R_n); (\hat{I}, \hat{Q}, \hat{R})\}$ where each shot is a triplet (I, Q, R) composed of an image, a query, and a result. The first n triplets serve as demonstrations used in ICL, and the final represents a query sample. The components Q and R are filled and adjusted according to the specific formats of the three different vision-language tasks: open-ended VQA, image captioning, and image classification, proving the configuration's potential for cross-task generalization. The details of configuring for the three tasks is provided in the Appendix B.1.

2.3 Disentangling TR and TL

Following (Pan et al., 2023), we developed three settings based on our configuration to examine the model's performance on TR or TL separately within open-ended VQA and image classification tasks by manipulating demonstrations: **Standard, Random, and Dislocation**, which are used to reflect the LVLM's overall ICL ability, isolated TR, and isolated TL, respectively. In **Random** and **Dislocation**, we specifically avoid altering both Q and R, allowing us to compare the individual importance of Q and R to the mechanisms of ICL. Details of the demonstration manipulations can be found in the Appendix A.2. We randomly sample the n-shot demonstrations following a uniform distribution.

As shown in Figure1,**for LVLMs, TR is more important than TL** because their extensive fine-tuning equips parametric memory to fill TL gaps. However, this may lead to conflicts between internal and external knowledge, emphasizing the need for a strong TR framework. TR is more critical for Qwen2VL-7B compared to OFv2-9B(Awadalla et al., 2023a), further indicating that **a backbone with broader task applicability will increase LVLM's reliance on fine-grained cross-modal mapping in TR.**

TR is more important than TL because the differences between queries and results are greater,

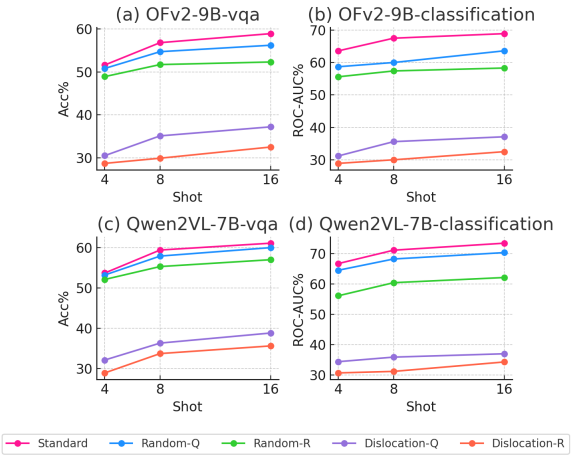

Figure 1: Results of five settings across two tasks and two LVLMs which represent different parts of LVLM's in-context learning.

making the mappings within each demonstration and between different demonstrations harder to interpret. This implies that **the more complex the vision-language task, the stronger the need for TR, while the demand for TL is prone to be influenced by the LVLM itself**. Both in TR and TL, Q is more important than R, confirming that **strong performance in LVLMs is closely tied to well-constructed semantic guidance information.**

3 Method

3.1 Retrieving Demonstrations

For a given query sample $(\hat{I}, \hat{Q}, \hat{R})$ we compute the joint similarity between $(\hat{I}, \hat{Q})$ and each candidate demonstration's (I_i, Q_i) using **cosine similarity**. We first generate CLIP embeddings for images and their corresponding queries in couples. Then we use **Joint Embedding Similarity (JES)** to compute the joint image-text similarity, which concatenates the image and query embeddings to form a comprehensive vector, and use this unified representation to compute the similarity between $(\hat{I}, \hat{Q})$ and each (I_i, Q_i) as a whole. We retrieve $4n$ demonstrations with the highest similarity scores to construct the optimal sequence S.

3.2 Prompt Construction

We construct the prompt presented to the LVLM based on the optimal sequence S. The details of prompt construction, especially regarding special tokens for multi-image inputs required by different LVLMs, can be found in the Appendix C.2.

Reordering We further optimize the sequence S and its related prompt using an image-query fusion-based self-evaluation approach. For each demonstration D, we concatenate its visual features I_i with $\hat{I}$ of the sample, and similarly concatenate the query Q_i with $\hat{Q}$. This combined input is passed through the LVLM, and the resulting logits are converted into a probability distribution. The likelihood score for the demonstration is computed as the sum of the log probabilities for each token in the ground truth answer:

$$\log \text{Score}(D) = \sum_{t=1}^{T} \log P(d_t \mid [I_i, \hat{I}], [Q_i, \hat{Q}])$$

where d_t denotes the token of the ground truth answer sequence. The $4n$ retrieved demonstrations are filtered by selecting the top n based on their likelihood scores within S, and then ordered in either **ascending(asc)** or **descending(desc)** order.

Bridging For the reordered sequence S, we insert a bridging instruction $Brid$ between the final demonstration and the query sample, forming $S' = \{(I_1, Q_1, R_1), ..., (I_n, Q_n, R_n); Brid; (\hat{I}, \hat{Q}, \hat{R})\}$. To generate $Brid$, we prompt GPT-4o with specific descriptions detailing the format of the data and the task, its objectives, and the requirements for the instruction.

4 Experiment

We conduct a comprehensive evaluation using our proposed strategies for novel well-trained LVLMs.

Dataset and Implementation Details We use six datasets across three VL tasks: three open-ended VQA datasets, VQAv2(Goyal et al., 2017), VizWiz(Gurari et al., 2018), and OK-VQA(Marino et al., 2019); two image captioning datasets, Flickr30K(Young et al., 2014) and MSCOCO(Lin et al., 2014); and one classification dataset, HatefulMemes(Kiela et al., 2020). We employed three LVLMs that support multi-image input and have demonstrated strong few-shot performance: OpenFlamingo-9B v2(OFv2-9B), InternVL2-8B, and Qwen2VL-7B. During demonstration retrieval, we use clip-vit-base-patch16 to convert them into embeddings and compute the similarity. The baselines we used will be detailed in the Appendix C.3.

4.1 Results and Analyses

Table 1 shows that our modality fusion strategy significantly outperforms the use of image or query alone across all three tasks. Additionally, our method surpasses the approach of retrieving based on concatenated query and result, further proving

Methods(n=4)	VQAv2	VizWiz	OK-VQA	Flickr30k	MSCOCO	HatefulMemes
No Demo	51.8	37.1	57.9	54.3	79.1	62.5
RS	52.4	44.5	59.1	59.1	85.6	65.7
I2I	53.2	51.1	62.3	61.1	81.2	67.2
Q2Q	57.9	54.3	65.4	83.9	91.5	72.8
SQPR(Li et al., 2024)	57.1	55.6	68.3	81.4	93.1	75.5
$IQ2IQ$	63.8	61.3	72.4	85.7	99.4	77.2
$IQ2IQ$(+asc)	64.9	61.6	73.1	**88.3**	**104.0**	80.3
$IQ2IQ$(+desc)	**66.3**	**62.1**	**75.6**	**88.3**	103.9	**82.4**

Table 1: Average performance on OFv2-9B, Intern2VL, and QwenVL2. $IQ2IQ$ is the retrieval method we propose, with (+asc) and (+desc) indicating ascending and descending reordering added, respectively. Baselines and the specific metric for each dataset are detailed in the Appendix C.3 and C.1.

that in multimodal ICL, it is essential to consider both image and textual information together, as cross-modality adapting plays a critical role in TR.

As shown in Table 1, in the VQA task, descending order consistently outperforms ascending order, with an average accuracy improvement of 0.67. In the classification task, descending order performs 2.1 points higher than ascending order. **Descending order allows the LVLM to quickly build an accurate framework during TR** with the most valuable demonstration presented first. Additionally, the presence of $Brid$ further reinforces this reasoning chain. As for captioning, the difference between descending and ascending order is minimal, with a maximum gap of only 0.1, since this task's query is simpler and the framework construction is less complex, leading to a smaller effect of demonstration ordering.

The relationship between shot number and performance is generally positive, but the gains decrease as the number further increases as shown in Figure . This once again confirms that TR is shown to be more critical than TL. TR relies on key sample attributes to form a task-specific framework, which doesn't need excessive demonstrations. Once the framework is established, additional samples provide minimal improvement to TR and thus the overall performance, though TL continues to benefit from increased demonstrations for more accurate task mapping.

In addition to optimizing the reasoning chain, $Brid$ plays a key role in LVLM's handling of interleaved image-text inputs as it increases the density of textual information to improve the model's awareness of mappings within demonstrations. $Brid$ is further discussed in the Appendix C.4.

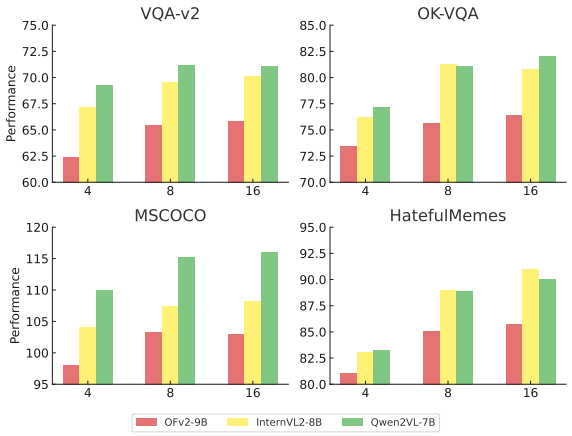

Figure 2: Results of experiments under the $IQ2IQ$(+desc) setting with different shot numbers (4, 8, 16) across four datasets and three LVLMs. The performance improvement is more noticeable when increasing the shot number from 4 to 8 compared to the increase from 8 to 16.

5 Conclusion

In this study, we interpret the in-context learning (ICL) process of LVLMs from the perspective of modality fusion and identify the crucial role of Task Recognition (TR) in cross-modal ICL. Based on this finding, we introduce a new component into the demonstration configuration to emphasize its modality fusion properties. Using this configuration, we conduct a deeper theoretical analysis of LVLMs and demonstrate the importance of modality fusion mapping for effective ICL. As a result, we propose a comprehensive strategy for constructing optimal few-shot prompts for LVLMs, from retrieval to prompt design, and demonstrate its generalization and effectiveness across different tasks and models through experiments.

Limitations

In this study, although we selected six datasets across three vision-language tasks and tested on three high-performing LVLMs, this is far from sufficient to establish a fully generalized image-text cross-modal learning framework. To bridge the gap between current methodologies and real-world applications, we need to further optimize in-context learning strategies by targeting more complex tasks and integrating more advanced LVLMs. Additionally, since we worked with pre-existing datasets, applying this strategy to real-world scenarios will require the design of more comprehensive data processing steps.

Moreover, during our research, we observed that evaluating LVLMs on individual datasets can be overly simplistic, potentially leading to a misjudgment of the effectiveness of our proposed methods. There is still a lack of a comprehensive benchmark for assessing the few-shot learning capabilities of LVLMs that support multi-image inputs. Designing such benchmarks is essential for developing more meaningful methods, and it is one of our future goals. At the same time, the conflicts between internal and external knowledge introduced by ICL cannot be ignored, and we have yet to address this issue in our work.

References

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

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

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

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

Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. 2023. Understanding and improving in-context learning on vision-language models. *arXiv preprint arXiv:2311.18021*, 1(2).

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

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

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

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

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.

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

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

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

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

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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

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

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada. Association for Computational Linguistics.

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

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409.

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

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

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

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

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

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. 2022. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*.

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

A Vision-language TR and TL

A.1 LVLM Reasoning

When the LVLM is given a prompt P that incorporates few-shot external knowledge, it generates the desired result $DR = \{d_1, d_2, ..., d_n\}$, , where the t-th word is drawn from a word distribution $P(\cdot)$ computed by the LVLM as

$$P(d_t|P, d_{t-1})$$

When the LVLM performs reasoning using the sequence S, the in-context learning process can be represented as

$$P(R|I, Q) = P(\mathbf{T}|S) \times P(\mathbf{L}|S) \times P(R|I, Q, \mathbf{T}, \mathbf{L})$$

where **T** denotes the task type and objective identified by the LVLM, and **L** denotes the learned input-output mapping for the given task. The LVLM must sequentially analyze the query sample based on the results of TR and Task Learning TL, aligning the task-specific characteristics with the query.

A.2 Disentangling TR and TL

(1) **Standard**: The correct demonstrations (I_i, Q_i, R_i) are used as input to reflect both the TR and TL. (2) **Random**: For a given sequence S, all triplets' Q or R are replaced by the Q or R from one randomly selected demonstration within the sequence. This setting only reflects TR. The two subcategories are represented as Random-Q and Random-R. (3)**Dislocation**: In this setting, either Q or R in the sequence is modified with content that introduces semantic elements related to the image captioning task, resulting in (I_i, Q_i*, R_i) or (I_i, Q_i, R_i*). This setting only reflects TL. The two subcategories are represented as Dislocation-Q and Dislocation-R.

B Prompt Configuring

B.1 Demonstration Configuring Details

(a)**Open-ended VQA**: The query Q_i is the single question associated with the image I_i, while the result Ri is the answer to the question, provided as a short response. For the query sample, $\hat{Q}$ represents the question related to the image $\hat{I}$, and $\hat{R}$ is the expected output from the model.

(b)**Image Captioning**: Both Q_i and $\hat{Q}$ are set as short prompts instructing the LVLM to generate a caption for the given image, such as "Please write a caption to describe the given image." The result R_i corresponds to the actual caption of the image.

(c)**Image Classification**: Both Q_i and $\hat{Q}$ provide the textual information paired with the image followed by a directive requiring the model to classify based on the provided image-text pairs. The result R_i is the predefined class label.

For all the three tasks mentioned above, since the ground truth answers are not visible to the LVLM during reasoning, all the $\hat{R}$ are set as non-semantic toekns <DR> to represent the desired results.

B.2 Retrieving Strategies

Previous works have typically focused on calculating similarity between either the image or parts of the textual information in the query sample and the demonstrations from the library in isolation. However, this approach can lead to insufficient use of demonstrations by the LVLM, as discussed in Section 3. To address this issue, we propose a fusion-based retrieval strategy *IQ2IQ(image-query to image-query)*, which contains two implementation methods:

(1)**Averaged Modality Similarity (AMS)**: calculate the similarity between $\hat{I}$ and each I_i, and between $\hat{Q}$ and each Q_i, then take the average of these two similarities;

(2)**Joint Embedding Similarity (JES)**:compute the joint image-text similarity, which concatenates the image and query embeddings to form a comprehensive vector, and use this unified representation to compute the similarity.

B.3 Prompt Details

To assist the LVLM in distinguishing and interpreting the contexts, we prepend every text-containing component with a hint ending in a colon. The image information, on the other hand, is indicated by image tokens as required by the specific LVLM. Answer token <DR> in the query sample is replaced with an empty space when added to the end of the prompt, guiding the model to start reasoning. Depending on the input format required by different LVLMs, we may also include special tags at the beginning and end of the prompt.

Table 2 provides an overview of the prompt details used for the different models in our experiments. Each model, including OpenFlamingoV2, InternVL2, and Qwen2-VL, employs a structured approach to engage with image-text pairs. The two-phase task requires models to first absorb information from a series of prompts before utilizing that context to answer subsequent questions related to new images. This method allows for enhanced

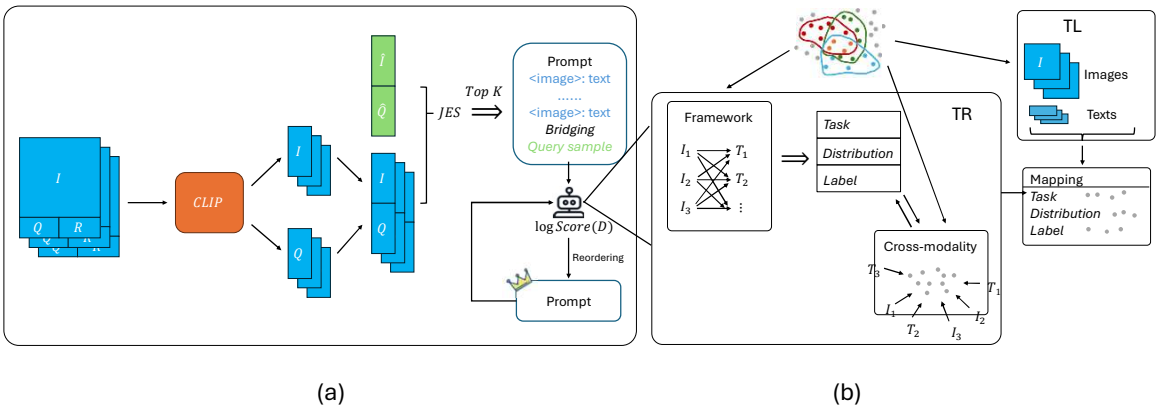

Figure 3: Illustration of our proposed pipeline for optimizing interleaved image-text prompts in LVLMs. (a) We build (I,Q,R)(I, Q, R)(I,Q,R) triplets, representing images (I), queries (Q), and responses (R), which are processed by a CLIP model. The joint embedding space retrieves top-K matches for the prompt construction. Bridging queries are introduced to enhance cross-modal coordination. (b) Task Recognition and Task Learning components are transferred to the multimodal domain, with TR further divided into **framework construction** and cross-modal adapting.

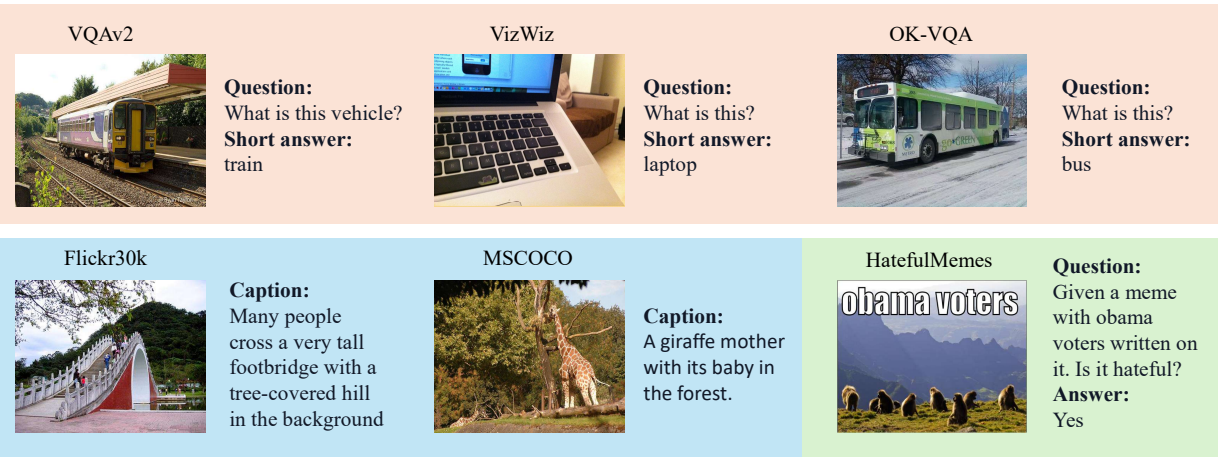

Figure 4: Examples from vision-and-language datasets. VQA tasks are highlighted in red (VQAv2: train, VizWiz: laptop, OK-VQA: bus), captioning tasks in blue (Flickr30k: footbridge, MSCOCO: giraffes), and classification tasks in green (HatefulMemes: meme labeled as hateful)

Models	Prompt details
OpenFlamingoV2	<img><IMG_CONTEXT>...<IMG_CONTEXT><|endofchunk|> Question: In what country can you see this? Short answer: vietnam <img><IMG_CONTEXT>...<IMG_CONTEXT><|endofchunk|> Question: Is this a buggy or car? Short answer: buggy You will be engaged in a two-phase task. 1: Absorb the information from a series of image-text pairs which are provided above. Phase 2: Use that context, combined with an upcoming image and your own database of knowledge, to accurately answer a subsequent question. <img><IMG_CONTEXT>...<IMG_CONTEXT><|endofchunk|> Question: What is this? Short answer:
InternVL2	<img><IMG_CONTEXT>...<IMG_CONTEXT></img> Question: In what country can you see this? Short answer: vietnam <img><IMG_CONTEXT>...<IMG_CONTEXT></img> Question: Is this a buggy or car? Short answer: buggy You will be engaged in a two-phase task. 1: Absorb the information from a series of image-text pairs which are provided above. Phase 2: Use that context, combined with an upcoming image and your own database of knowledge, to accurately answer a subsequent question. <img><IMG_CONTEXT>...<IMG_CONTEXT></img> Question: What is this? Short answer:
Qwen2-VL	<|im_start|>system You are a helpful assistant.<|im_end|> <|im_start|>user <|vision_start|><|image_pad|><|vision_end|>The question with this image up here is:In what country can you see this? Short answer: vietnam <|vision_start|><|image_pad|><|vision_end|>The question with this image up here is: Is this a buggy or car? Short answer: buggy You will be engaged in a two-phase task. 1: Absorb the information from a series of image-text pairs which are provided above. Phase 2: Use that context, combined with an upcoming image and your own database of knowledge, to accurately answer a subsequent question. <|vision_start|><|image_pad|><|vision_end|>The question with this image up here is: What is this? Short answer: <|im_end|> <|im_start|>assistant

Table 2: Prompt details for different models used in the experiments. The table outlines how OpenFlamingoV2, InternVL2, and Qwen2-VL format their image-text interactions, including examples of image-based questions and short answers. Each model follows a multi-phase task structure, where context is absorbed from previous image-text pairs to answer subsequent questions.

Datasets	VQAv2	VizWiz	OK-VQA	Flickr30k	MSCOCO	HatefulMemes
metrics	Accuracy	Accuracy	Accuracy	CIDEr	CIDEr	ROC-AUC

Table 3: Evaluation metrics used for each dataset. Accuracy is used for VQA datasets (VQAv2, VizWiz, OK-VQA), CIDEr(Vedantam et al., 2015) for image captioning datasets (Flickr30k, MSCOCO), and ROC-AUC for the HatefulMemes classification task.

understanding and reasoning based on prior knowledge and context, which is essential for accurate question answering in vision-and-language tasks.

C Experiment

C.1 Dataset

In our study, we explore various vision-and-language tasks that leverage diverse datasets to evaluate model performance. As illustrated in Figure 4, we utilize VQA datasets such as VQAv2, VizWiz, and OK-VQA, which test the models' abilities in question-answering scenarios. Additionally, we incorporate image captioning datasets like Flickr30k and MSCOCO to assess descriptive accuracy, along with the HatefulMemes dataset for classification tasks focused on hate speech detection. This comprehensive approach allows us to thoroughly evaluate the models across different tasks and data types.

For the Open-ended VQA task, we utilize the following datasets: VQAv2, which contains images from the MSCOCO dataset and focuses on traditional question-answering pairs, testing the model's ability to understand both the image and the question. VizWiz, on the other hand, presents a more challenging setting with lower-quality images and questions along with a lot of unanswerable questions, pushing models to handle uncertainty and ambiguity. OK-VQA is distinct in that it requires the model to leverage external knowledge beyond the image content itself to generate correct answers, making it a benchmark for evaluating models' capacity to integrate outside information.

For the Image Captioning task, we use the Flickr30k and MSCOCO datasets. The Flickr30k dataset consists of images depicting everyday activities, with accompanying captions that provide concise descriptions of these scenes. The MSCOCO dataset is a widely-used benchmark featuring a diverse range of images with detailed and richly descriptive captions, ideal for evaluating image captioning models.

For the Image Classification task, we use the HatefulMemes dataset, which is an innovative dataset designed to reflect real-world challenges found in internet memes. It combines both visual and textual elements, requiring the model to jointly interpret the image and the overlaid text to detect instances of hate speech.

For each dataset, we constructed the demonstration library using the training set and used the data from the validation set as query samples to evaluate the in-context learning capabilities.

The datasets in our experiments are evaluated using task-specific metrics, as summarized in Table 3. For the VQA tasks (VQAv2, VizWiz, OK-VQA), we use accuracy as the primary metric to assess the models' ability to provide correct answers. For the image captioning tasks (Flickr30k, MSCOCO), we use the CIDEr score, which measures the similarity between generated captions and human annotations. Finally, for the HatefulMemes classification task, we evaluate performance using the ROC-AUC metric, which reflects the model's ability to distinguish between hateful and non-hateful content.

C.2 LVLMs

In recent advancements of Large Vision-Language Models (LVLMs), efficient processing of multimodal inputs, especially images, has become a critical focus. Models like OpenFlamingoV2, InternVL2, and Qwen2-VL implement unique strategies to manage and process visual data alongside textual inputs.

OpenFlamingoV2 handles visual inputs by dividing images into patches and encoding them with a Vision Transformer. Each image patch generates a number of visual tokens, which are then processed alongside text inputs for multimodal tasks. To manage multi-image inputs, the model inserts special tokens <image> and <|endofchunk|> at the beginning and end of the visual token sequences. For example, an image divided into 4 patches produces 4 x 256 visual tokens, with the additional special tokens marking the boundaries before the tokens are processed by the large language model. This approach allows OpenFlamingoV2 to flexibly place images at any position in the sequence, ensuring seamless integration with text.

InternVL2 also dynamically divides images into 1 to 4 patches based on their aspect ratio. A Vision Transformer then extracts visual features from each patch, followed by a pixel shuffle operation and a mlp, producing 256 visual tokens for each patch. Additionally, special tokens <img> and </img> are inserted at the beginning and end of the sequence. So, an image divided into 3 patches will produce 3 x 256 + 2 tokens before entering LLM.

Qwen2-VL reduces the number of visual tokens per image through a compression mechanism that condenses adjacent tokens. A ViT first encodes an image (e.g., with a resolution of 224 x 224 and a patch size of 14), producing a grid of tokens, which

is then reduced by employing a simple MLP to compress 2 x 2 tokens into a single token. Special <|vision_start|> and <|vision_end|> tokens are inserted at the start and end of the compressed visual token sequence. For example, an image that initially generates 256 visual tokens is compressed to just 66 tokens before entering the LLM. This reduces the complexity of visual processing while maintaining key spatial and visual information.

C.3 Baseline

Various baseline methods are used to evaluate the model's performance, ranging from no demonstrations to different retrieval strategies. Below is a description of the baselines used in our experiments.

1.**No Demo**: This method does not utilize any demonstrations. The model's input consists solely of the query sample

2.**Random Sampling (RS)**: In this approach, a uniform distribution is followed to randomly sample n demonstrations from the library. These demonstrations are then directly inserted into the prompt to guide the model in answering the query.

3.**Image2Image (I2I)**: During the retrieval process, only the image embeddings I_i from each demonstration $(I_i, Q_i, R_i$ are used. These embeddings are compared with the query image embedding $\hat{I}$ and retrieval is based on the similarity between the images.

4.**Query2Query (Q2Q)**: During the retrieval process, only the query embeddings Q_i from each demonstration $(I_i, Q_i, R_i$ are used. These embeddings are compared with the query image embedding $\hat{Q}$ and retrieval is based on the similarity between the queries.

5.**Retrieving via Similar Question & Pseudo Result (SQPR)**: This baseline starts by using the RS to generate a pseudo result $\hat{R}^P$. The pseudo result is then concatenated with $\hat{Q}$ and retrieval is based on the similarity of this combined query and result.

C.4 $Brid$

The $Brid$ generated by GPT-4o in the main experiment is "You will be engaged in a two-phase task. 1: Absorb the information from a series of image-text pairs which are provided above. Phase 2: Use that context, combined with an upcoming image and your own database of knowledge, to accurately answer a subsequent question." This content demonstrates great orderliness and can act as a good semantic transition and guide between demonstrations and query sample.

In Table 4, we use GPT-4o to rewrite the $Brid$, placing it at both the beginning and the end of the sequence S, altering its semantic structure accordingly. Additionally, we generate two simplified versions of $Brid$ to explore whether the level of semantic detail impacts the performance of ICL. The experiments show that placing $Brid$ between the demonstrations and the query sample yields the most effective results, and the more detailed the semantic information in $Brid$, the greater its contribution to ICL.

$Brid$ can serve as a special yet vital demonstration, differentiating itself to avoid the diminishing returns of in-context learning. Furthermore, for latest LVLMs that support more complex inputs and tasks, sensitivity to the prompt increases, meaning that a high-quality $Brid$ leads to more noticeable performance gains. Therefore, as LVLMs become more advanced in integrating multimodal information, the prompt designing becomes increasingly crucial for optimal performance.

Bridge position	Bridge details
Middle	You will be engaged in a two-phase task. 1: Absorb the information from a series of image-text pairs which are provided above. Phase 2: Use that context, combined with an upcoming image and your own database of knowledge, to accurately answer a subsequent question.
Beginning	In this task, you will first absorb information from a series of image-text pairs provided. Then, using that context and your own knowledge, you will accurately answer questions based on an upcoming image.
End	After analyzing the series of image-text pairs provided above, you will use that context, combined with your own knowledge, to accurately answer a subsequent question regarding a new image.
Middle (abbreviated)	Engage in two phases: first, absorb information from the provided image-text pairs; then, use that context and your knowledge to answer questions based on a new image.
Middle (most abbreviated)	First, absorb the information from the image-text pairs; then use that context to answer questions about a new image.

Table 4: Summary of different bridge positions and their corresponding details used in the prompt structure for image-text tasks. Each entry demonstrates how the phrasing and position of the bridge statement can affect the clarity and flow of the instructions given to the model.