# Provoking Multi-modal Few-Shot LVLM via Exploration-Exploitation In-Context Learning

## Supplementary Material

This appendix provides the training algorithm of our proposed framework in Section A. The details of stochastic beam search and prompting are given in Section B and Section C. In Section D, we analyze the cross-domain performance of our method compared with heuristic methods, which have strong interpretability and potential in transfer. Next, we provide an ablation study of beam search width. We additionally visualize more demonstration cases in Section F.

## A. Training Algorithm

The training algorithm of the proposed method is elaborated in Algorithm 1. In the algorithm, we describe: 1) the initialization of the networks $f_{\{f,d\}}$, embeddings $\mathbf{f}_{\{\mathbf{t}_i,\mathbf{I}_i\}}$; 2) the multi-modal interactive encoding of each query to get the interactive features $\mathbf{M}$; 3) the exploration process with stochastic beam search; 4) the exploitation process with responsive reinforcement and LVLMs inference. The completed pipeline consists of four parts above.

**Algorithm 1:** Training Algorithm

**Data:** task $\mathcal{T}$, candidate demonstrations $\mathcal{D}$, size of demonstration sets $m$, beam width $c$, multi-modal encoder $f_e$, interactive encoder $f_f$, auto-regressive decoder $f_d$
**Result:** policy $\pi_\theta$

```
/* init policy and embeddings     */
```
init $f_f, f_d$ randomly;
prompt each demo in $\mathcal{D}$;
encode prompts of each demo in $\mathcal{D}$ to get $\mathbf{f}_{\{\mathbf{t}_i,\mathbf{I}_i\}}$;
**for** $(\mathbf{t}_q, \mathcal{I}_q, \mathbf{y}) \sim \mathcal{T}$ **do**
```
  /* multi-modal interactive
     encoding                    */
```
  $\mathbf{f}_{\{\mathbf{t}_q,\mathbf{I}_q\}} \leftarrow f_e(\mathbf{t}_q), f_e(\mathcal{I}_q)$;
  $\mathbf{M} = f_f(\begin{bmatrix}\mathbf{f}_{\mathbf{t}_q} & \mathbf{f}_{\mathbf{I}_q} & \mathbf{f}_{\mathbf{t}_1} & \mathbf{f}_{\mathbf{I}_1} & \mathbf{f}_{\mathbf{t}_2} & \mathbf{f}_{\mathbf{I}_2} & \cdots\end{bmatrix})$;
```
  /* stochastic beam search      */
```
  $B_0 \leftarrow \{(1, \varnothing)\}$;
  **for** $i \leftarrow 1$ **to** $m$ **do**
    $B \leftarrow \varnothing$;
    **for** $(s, \mathcal{E}) \in B_{t-1}$ **do**
      $\mathbf{q}_i \leftarrow f_d(\mathbf{M}, \{q_j \mid \mathbf{d}_j \in \mathcal{E}\})$;
      $\mathbf{s} \leftarrow$ calculate the similarities between $\mathbf{q}_i$ and each $\mathbf{f}_{\{\mathbf{t},\mathbf{I}\}}$;
      **for** $\mathbf{d}_j \in \mathcal{D}$ **do**
        $B \leftarrow B \cup \{(s \cdot \mathbf{s}_i, \mathcal{E} \cup \mathbf{d}_j)\}$;
      **end**
    **end**
    $S \leftarrow 0$;
    **for** $(s, \mathcal{E}) \in B$ **do**
      $S \leftarrow S + |s|$;
    **end**
    $s \leftarrow \frac{s}{|S|}$ for each $(s, \mathcal{E})$ in $B$;
    sample from $B$ with possibility $s$ to get $B_i = \{(s_j, \mathcal{E}_j)\}_{j=1}^{c}$;
  **end**
  $\{(s_i, \mathcal{E}_i)\}_{i=1}^{c} \leftarrow B_m$;
```
  /* inference                   */
```
  $\mathbf{o}_i \leftarrow \mathcal{F}(P(\mathbf{t}_q, \mathcal{E}_i), \bigcup_{\mathcal{I}_j \in \mathcal{E}_i} \mathcal{I}_j)$;
  $A_i \leftarrow r(\mathbf{o}_i, \mathbf{y})$;
  $\hat{A}_i \leftarrow \frac{A_i - \text{mean}(\mathbf{A})}{\text{var}(\mathbf{A})}$;
  Optimize $f_f, f_d$ with $\mathcal{L}$;
**end**

## B. Discussion of Learnable Selection Methods

We are not the first to propose learnable selection methods for in-context learning, and previous works have explored this area [3, 4, 6, 7]. However, our method differs significantly: 1) Modality: Prior methods focus solely on unimodal LLM, making them unsuitable for multi-modal inputs [3, 4, 6, 7]. In contrast, our method incorporates interactive encoding that effectively leverages information from multiple modalities while considering the interactions between inputs. 2) Evaluation: Previous methods treat LLM as a direct evaluator, assessing the validity of demo-query pairs [6, 7]. This standard deviates considerably from LLM's actual reasoning results. Our method directly supervises policy learning using the reasoning outcomes of LLM. 3) Perspective and modeling: Previous methods independently selected demonstrations without considering their interrelationships [4]. Our method examines the ICL demonstration selection from a novel perspective, introducing new ideas at the level of demonstration combinations and further advancing exploration in this field.

## C. Details of Stochastic Beam Search

In Section 3.3 of the manuscript, we introduce a modified version of Beam Search, transforming its sampling from deterministic to stochastic. The pytorch implementation is listed below. First, we compute the logits for each expand-

ing token like the conventional beam search algorithm. Subsequently, the standard beam search applies softmax to the logits and selects the top-k candidates for the next expansion. To make this process stochastic, we normalize all logits after the softmax step to serve as the sampling probabilities, followed by random sampling. In this manner, superior tokens can retain their advantages, while the suboptimal may be selected, thereby balancing exploration efficiency with comprehensiveness.

```
logit = llm(tokens)
prob = F.normalize(softmax(logit),
 dim = -1)
sum = prob.cumsum(dim = -1)
ids = torch.searchsorted(sum,
 torch.rand(*sum.shape[:-1], c))
p = torch.gather(tokens, -1, ids)
# sequences: ids,
# possibilities: p
```

## D. Details of Prompting

When using LVLM inference, it is necessary to construct prompts. In our method, the parts that require prompt construction include: the prompts during LLM inference and the question-answer text embedding for both queries and demonstrations. We list all prompts below. As prompts are not the contributions of our work, we keep all prompts minimal. In practical tasks, the prompts can be further adjusted to improve performance.

**Prompt of Text Embeddings** For demonstrations, the prompt consists of images, questions, and answers:

```
<Image>
<Question>
<Answer>
```

For queries, the prompt consists of image and question only:

```
<Image>
<Question>
```

The prompted queries and demonstrations are encoded by $f_e$ to get the features.

**Prompt of LVLM Inference** For inference, we follow in-context learning and adopt the minimal prompt:

```
<#1 Demo Image>
<#1 Question>
<#1 Answer>
<#2 Demo Image>
<#2 Question>
<#2 Answer>
...
```

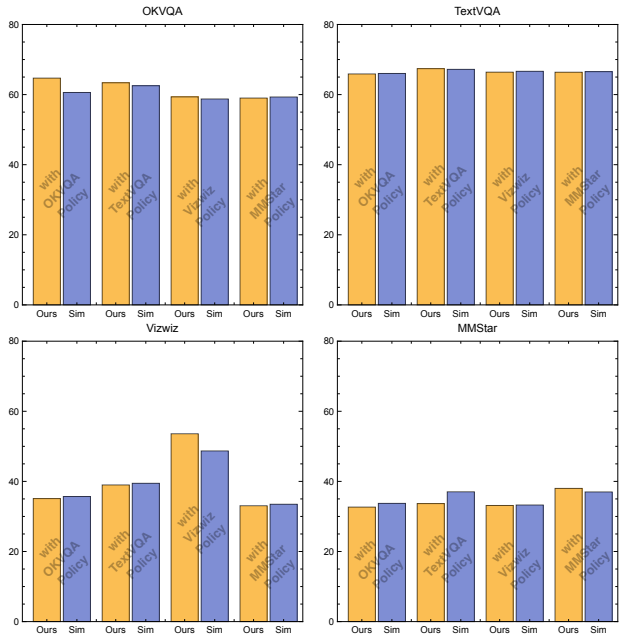


Figure 1. Transfer performance of our proposed method and similarity-based method on four benchmarks. We evaluate the policies trained on four datasets on each of the four. The performance of all methods decreases when using cross-dataset policies. Due to larger training data, the performance of the similarity-based method has advantages.

```
<Query Image>
<Query Question>
```

We adopt the base model for inference, which infers and outputs predictions based on the query. In practical scenarios, if a fine-tuned model is utilized, the prompts can be adjusted, *e.g.*, by introducing instructions, to achieve better results.

## E. Cross-Domain Performance

In addition to the performance on tasks of the same type, the performance on cross-domain tasks and different models, *i.e.*, the transferability, is also an important aspect of evaluating the policy. Although heuristic algorithms generally perform worse than learnable methods in some tasks, their strong interpretability often provides an advantage in transfer performance. We conduct experiments to explore the performance of our method after data transfer on OKVQA [5], TextVQA [8], Vizwiz [1] and MMStar [2] benchmarks. The results are shown in Figure 1. As shown in figure, 1) whether using our method or similarity-based methods, cross-data policies perform worse after transferring compared to the one on the original dataset. For example, significant performance degradation occurs in cross-data policies on Vizwiz and MMStar benchmarks; 2) the advantages

Table 1. Detailed transfer performance of our proposed method on each sub-task of MMStar benchmarks. CP: coarse perception. FP: fine-grained perception. IR: instance reasoning. LR: logic reasoning.

| Source | Target - MMStar | | | | | |
|---|---|---|---|---|---|---|
| | CP | FP | IR | LR | ST | MA |
| OKVQA | 50.4 | 28.4 | 43.6 | 25.2 | 19.6 | 28.8 |
| TextVQA | 48.0 | 34.0 | 46.0 | 25.2 | 20.8 | 28.0 |
| Vizwiz | 50.8 | 34.0 | 43.6 | 23.6 | 18.8 | 28.0 |
| MMStar | 54.6 | 38.0 | 49.8 | 36.7 | 20.1 | 28.4 |

of our method become less pronounced after transferring. This discrepancy may be attributed to the data distribution. Heuristic similarity-based methods exhibit stable lower performance across various data, while our method is more affected by distribution changes yet yields better performance. To obtain a policy suitable for general scenarios, our method may require more diverse data for training.

To further explore the cross-domain performance, we report detailed results on MMStar in Table 1. Although we observe the performance degradation in Figure 1, the detailed results reveal that the strategies do exhibit that transferability and universal strategies are feasible. The fine-grained visual reasoning in TextVQA and Vizwiz transfer well, leading to better performance on Fine-grained Perception (FP) and Instance Reasoning (IR) compared to OKVQA's strategies.

## F. Ablation Study of Beam Search Width

An important hyperparameter in beam search is the width, which determines the breadth of the tree searching and affects the exploration range of the policies. We conduct ablation experiments to study its impact on performance. The results are shown in Figure 3. As illustrated in the figure, the performance exhibits a trend of first decreasing and then increasing. We believe that when the width is low, the exploration range is limited, leading to unstable performance, as evidenced by the fluctuations observed at widths of 2 and 4. As the width increases, the exploration range expands, allowing the method to more readily identify superior policies, thereby resulting in a stable enhancement of performance.

## G. More Visualization

In Figure 4, we list additional visualizations of selected demonstrations by ours, similarity-based method, and random sampling. In Figure 4 (a), the demonstrations selected by our method focus on describing object properties. Although they are semantically far from the query, the question structure is consistent with it, providing better cues.

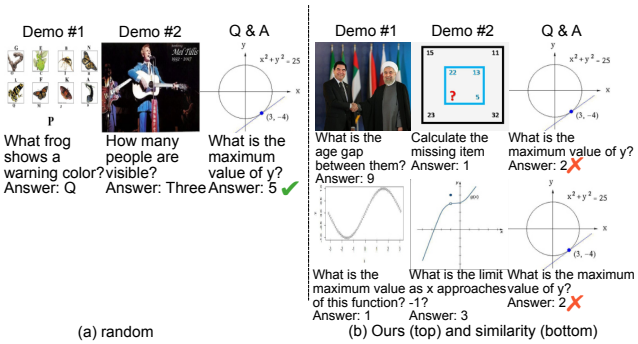


Figure 2. The visualization of bad predictions on math tasks by ours. Both the similarity-based strategy and our approach fail, while random selection succeeds.

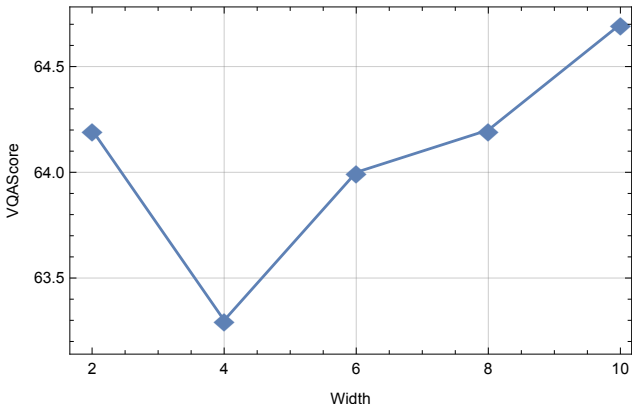


Figure 3. VQAScore on OKVQA benchmark with different beam search width. Performance first decreases and then increases.

The LVLMs predict the correct results on random samples. In contrast, while similarity policy selects similar demonstrations, their questions are completely different, which mislead LVLMs into predicting incorrect results. A similar situation also occurs in (c). In (d), both predictions of LVLMs are correct, one focusing on the scene and the other on human actions. However, when combined with image semantics, we can understand that the question asks about the former, which is more informative. Our method selects appropriate demonstrations, allowing LVLMs to focus on predicting the scene semantics, thereby obtaining the correct answer.

Next, to understand the lower performance on math tasks, we collect the outputs of LVLMs on those questions. One of the results is shown in Figure 2. Math problems follow rigid rules where minor variations can lead to completely different solution paths. Related but non-isomorphic demonstrations interfere with LVLMs' understanding.

These examples further reveal that demonstration selection is challenging and cannot be simply solved based on a single factor. After the few-shot demonstrations improve

the correctness of the output format, further performance enhancement needs to take into account the complex interaction between the demos and the queries. This interaction is difficult for humans to understand and design. Our method can autonomously explore and exploit this, alleviating the issue.

## References


[1] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and Tom Yeh. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, page 333–342, New York, NY, USA, 2010. Association for Computing Machinery. 2

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 2

[3] Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 1

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

[5] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 2

[6] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States, 2022. Association for Computational Linguistics. 1

[7] Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing, 2022. 1

[8] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 2

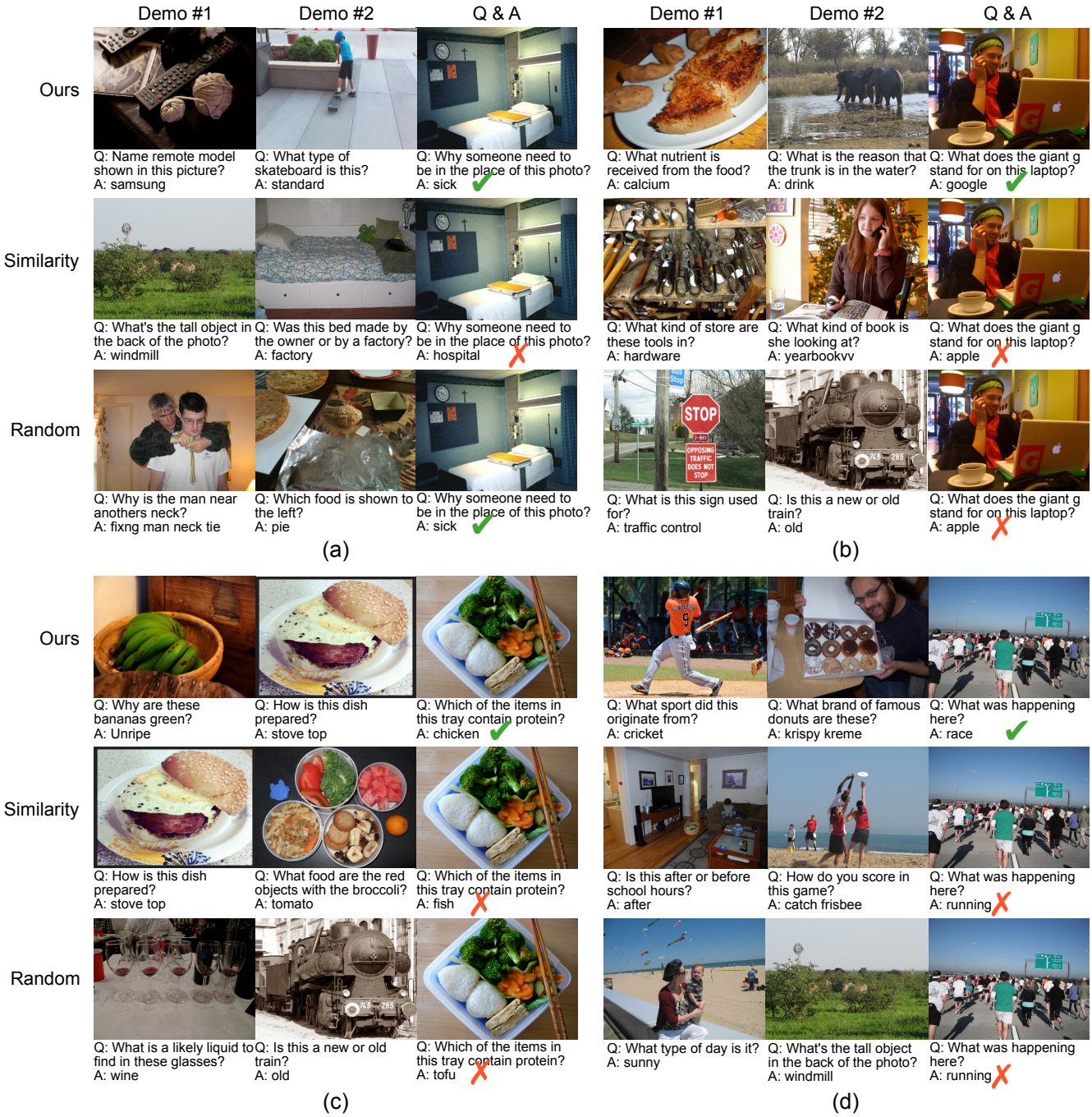


Figure 4. The visualization of selected demonstrations by ours, similarity-based method, and random sampling on the OKVQA dataset. (a) Our demos focus on describing object properties, while similar demos mislead models to predict locations. (b) Our demos correct the wrong predictions. (c) Our demos correct the wrong predictions, and similar demos result in non-existent objects. (d) Our demos suggest models to predict the type of sports, which is more informative.