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

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

## 1. Experimental Setup

**Vision-language Models.** We investigate different models from OpenFlamingo [3], IDEFICS [7] and MMICL [18] with various model sizes as shown in Tab. 1. OpenFlamingo [3] and IDEFICS [7] 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 [14] in OpenFlmaingo and LLaMA [16] in IDEFICS), a frozen visual encoder (*e.g.*, CLIP-ViT [12]) 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 [13] and 43M interleaved image-text sequences in Multimodal C4 [19]. IDEFICS is trained on OBELICS [7] which contains 141M multimodal Engish web documents with 353M images and 115B tokens. Both models achieve competitive performance compared to Flamingo [1]. 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 [9] and LLaVA-Instruct [10]. MMICL [18] uses a different model architecture and treats image and text representations equally. MMICL first uses a ViT to get image representations. Then Q-Former is used to extract visual embeddings and a fully connected layer converts each visual embedding to the same dimension as the text embedding of the LLM. Finally, the visual embeddings of multiple images and text embeddings are combined in an interleaved style and fed into the LLM.

Table 1. Vision-language models studied in this work. `OF` stands for OpenFlamingo [3] and `I` means instructed version.

| Model | Vision Encoder | Language Model |
|---|---|---|
| `OF-3B` | CLIP Vit-L/14 | MPT-1B [14] |
| `OF-3B-I` | CLIP Vit-L/14 | MPT-1B-I [14] |
| `OF-4B` | CLIP Vit-L/14 | RedPajama-3B [15] |
| `OF-4B-I` | CLIP Vit-L/14 | RedPajama-3B-I [15] |
| `OF-9B` | CLIP Vit-L/14 | MPT-7B [14] |
| `IDEFICS-9B` | OpenCLIP Vit-H/14 | LLaMA-7B [16] |
| `IDEFICS-9B-I` | OpenCLIP Vit-H/14 | LLaMA-7B [16] |
| `MMICL` | CLIP ViT-G/14 | FlanT5-XL [18] |

**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 [5] and OK-VQA [11] are adopted. Additionally, we incorporate GQA [6] for visual reasoning and MSCOCO [4] for image captioning. The statistics are in Tab. 2. 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 [17] on the Karpathy-test split is used in MSCOCO. All experiments are conducted on one Nvidia DGX Node with 4 Nvidia A100 (80GB) GPUs, 1TB memory, and 252 CPU Cores.

*corresponding author

Table 2. Dataset Statistics. Four well-known datasets from three popular vision-language tasks are used in this study.

| Task | Dataset | # Images | # Image-text pairs |
|---|---|---|---|
| VQA | VQAv2 [2] | 123.2K | 658.1K |
| | OK-VQA [11] | 14K | 14K |
| Visual Reasoning | GQA [6] | 82.3K | 1087.7K |
| Image Captioning | MSCOCO [4] | 123.2K | 576.8K |

# 2. Additional Results of Importance Investigation on Visual and Textual Information

## 2.1. Importance of Visual Information

To evaluate the importance of visual information, we have designed various demonstration settings as shown in Tab. 3.

- ***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. 5 and Tab. 6. 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. 7 and Tab. 8. The results also suggest 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. Besides, we also conducted experiments on models without the masked cross-attention and the results also indicate the limited influence of demo images as shown in Tab. 9.

## 2.2. Importance of Textual Information

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

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

Performance of `OF-9B` and `IDEFICS-9B` across 4 datasets given randomly selected demonstrations are presented in Tab. 10 and Tab. 11.

# 3. More Details of Understanding Multimodal Information Flow

The ability to handle interleaved text and image sequences makes ICL possible [1]. An illustration is presented in Fig. 1, with two demos and a query, each of which contains an image and corresponding text such as $I_1$ and $T_1$ in the first demo. The masked cross-attention layer enables the language models to incorporate visual information for the next-token prediction. This layer also limits the visual tokens the model can see at each text token. Specifically, at a given text token, the model only attends to the visual tokens of the last preceding image, rather than to all previous images in the interleaved sequence. For example, text embedding $\mathbf{T_q}$ can only attend to the query image $\mathbf{I_q}$ in the masked cross-attention layer, as shown in the last row of $A_c$ in Fig. 1. Therefore, demonstration images $I_1$ and $I_2$ cannot directly pass their visual information to the query text embedding $\mathbf{T_q}$, as $\mathbf{T_q}$ is limited to interacting with the query image representation $\mathbf{I_q}$ in the masked cross-attention layer. Only in the subsequent self-attention layer can $\mathbf{T_q}$ indirectly access the information from $I_1$ and $I_2$ through the demo text embeddings $\mathbf{T_1}$ and $\mathbf{T_2}$. Because

Table 3. 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 in the demonstration 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? |
| ***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 textual demonstration settings with one demonstration and one query. The differences compared to the `standard` setting are highlighted in blue.

| 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? |

they have already processed the visual information from $I_1$ and $I_2$ in the masked cross-attention layer. We argue that the masked cross-attention mechanism with such per-image attention masking [1] diminishes text tokens' dependency on all previous images. In other words, relying solely on the self-attention layer for transferring visual information to text tokens is difficult. Thus, it is observed that the generated output tokens primarily focus on the latest image, *i.e.*, the query image, and largely disregard the visual information of the previous images.

Masked cross-attention enables the processing of interleaved text and visual sequences, allowing for in-context few-shot learning to be possible [1]. As depicted in Fig. 1, the visual information from demonstration images $I_1$ and $I_2$ cannot directly influence the query text embedding $\mathbf{T_q}$. This is because $\mathbf{T_q}$ only interacts with the query image representation $\mathbf{I_q}$ in the masked cross-attention layer. To assess the impact of visual information from the demonstration on the generated content, 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. 2.
- ***hide query visual embedding*** refers the case where the visual embeddings from query images are masked, as shown in the right side of Fig. 2.

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. 3) and the attention weights in the last layer. We

Table 5. 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 | 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 `IDEFICS-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 |

then compute the cosine similarity between these extracted values and their counterparts in the *standard* setting.

# 4. More Results on the ICL Performance Improvement

## 4.1. More Results

We have conducted experiments using various models and VL datasets, which are listed in Table 1 and Table 2. 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. 23 presents examples selected by MMICES and RICES.

Table 7. 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 | 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 `IDEFICS-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 `Qwen-VL` on VQAv2. We observe similar trend where the removal of images lead to no performance decrease.

| Model | Setting | 4 | 8 | 16 |
|---|---|---|---|---|
| Qwen-VL | standard | 74% | 74% | 72.3% |
| | demo w/o img | 75.6% | 74.3% | 75.9% |

## 4.2. 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.

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 | 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 11. The performances of `IDEFICS-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 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 | 43.45 (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)** |

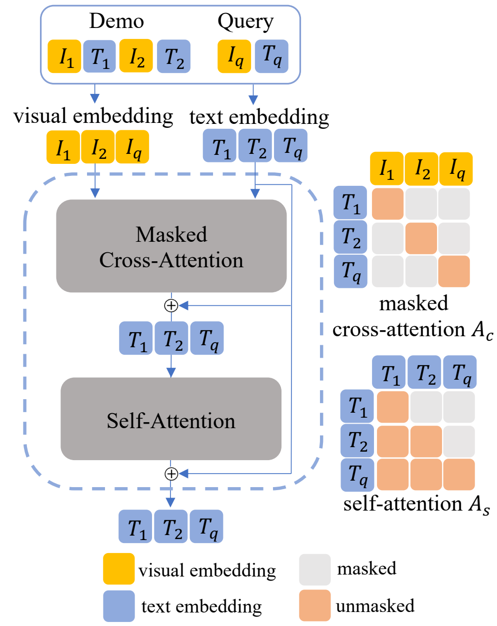



Figure 1. Model block supporting interleaved image-text inputs. Visual and language information, *i.e.*, $I$ and $T$, are first fused using a masked cross-attention layer, where each text token is only conditioned on the last preceding image. Visual embeddings $\mathbf{I_1}$ and $\mathbf{I_2}$ from demonstration 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 $\mathbf{A_c}$.

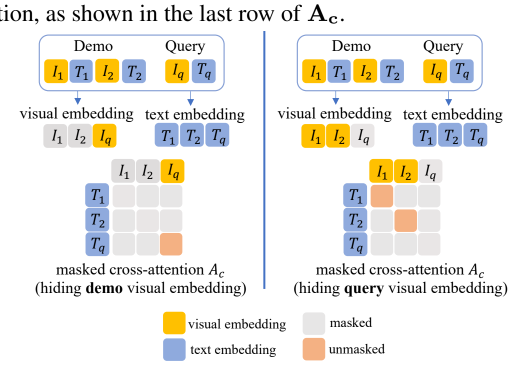



Figure 2. Compared with the standard setting, we hide demo visual embedding and query visual embedding respectively to explore the influence of different visual embeddings.

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 | 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)** |

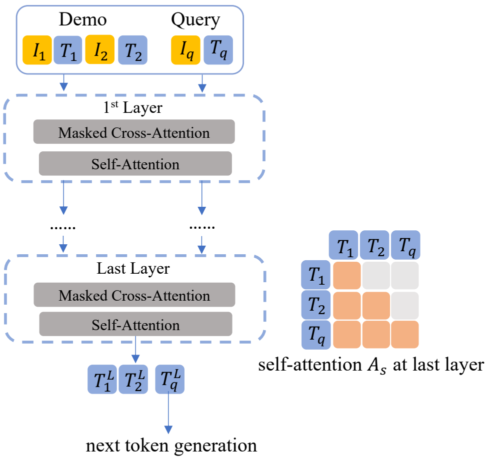



Figure 3. We compute the cosine similarity on the last row of hidden states,*i.e.*, $\mathbf{T}_q^L$ in this figure, and attention weights, *i.e.*, $\mathbf{A}_s$ in this figure, in the last decoder layer for each generation forward and then average the results over the whole dataset.

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 | 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 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 | 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) |

**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 numbers 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 select-

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 | 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 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 | 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 18. The performances of random selection, RICES, and MMICES on `IDEFICS-9BI`. MMICES achieves the best performance in most 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 19. Performance of MMICES given different $K$.

| 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 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.

| 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 |

ing 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.

## 5. 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. 1). 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. 4 to Fig. 6. 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. 4) and visual reasoning (Fig. 5). 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

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.

| 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 |

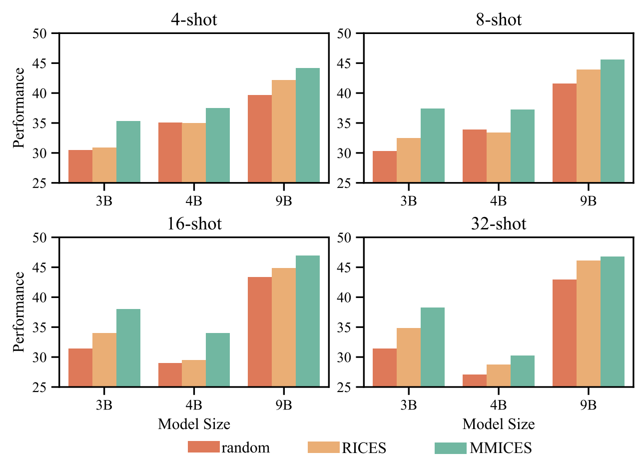



Figure 4. The performance of ICL (on OK-VQA) is consistently enhanced by MMICES on OpenFlamingo with different sizes.

gained from MMICES is consistent across different models, as shown in Tab. 22 and Fig. 7 to Fig. 9. 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 [7].

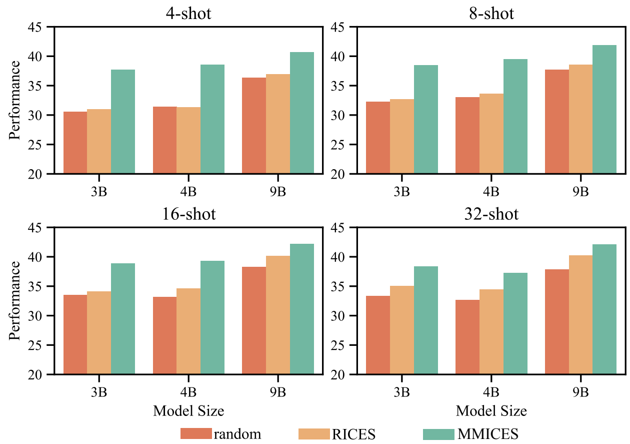



Figure 5. The performance of ICL (on GQA) is consistently enhanced by MMICES on OpenFlamingo with different sizes.

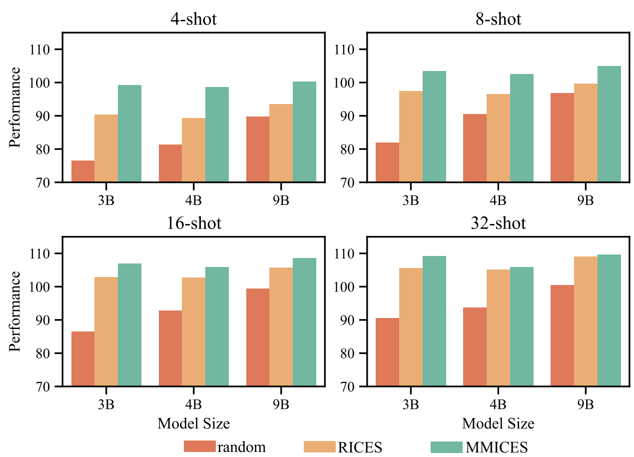



Figure 6. The performance of ICL (on COCO) is consistently enhanced by MMICES on OpenFlamingo with different sizes.

Table 22. MMICES achieves better performance across three different MLLMs. The performance is the accuracy evaluated on OK-VQA.

| Model | Method | 4-shot | 8-shot | 16-shot | 32-shot |
|---|---|---|---|---|---|
| OF-9B | Random | 39.62 | 41.56 | 43.40 | 42.97 |
| | RICES | 42.13 | 43.87 | 44.90 | 46.15 |
| | MMICES | **44.18** | **45.61** | **46.93** | **46.79** |
| IDEFICS-9B | Random | 49.24 | 49.54 | 50.89 | 51.86 |
| | RICES | 48.82 | 50.55 | 52.42 | 53.22 |
| | MMICES | **49.63** | **52.16** | **53.65** | **54.16** |
| MMICL | Random | 49.37 | 48.90 | 48.32 | 47.29 |
| | RICES | 49.77 | 49.87 | 49.24 | 48.34 |
| | MMICES | **52.43** | **52.21** | **51.20** | **49.39** |

## 5.1. Ablation Study

The number of pre-filtered samples, *i.e.*, $K$, selected by visual similarity is a hyperparameter in MMICES. Additionally, as MMICES considers both visual and language modalities, we also investigate the ICL performance when the examples are retrieved 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. 10 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 modality mixture, the results are consistent with our analysis. Retrieval based on a single modality, such as RICES on visual, underperforms mixed modality retrieval. Besides, MMICES consistently achieves better results compared to *text-image*.

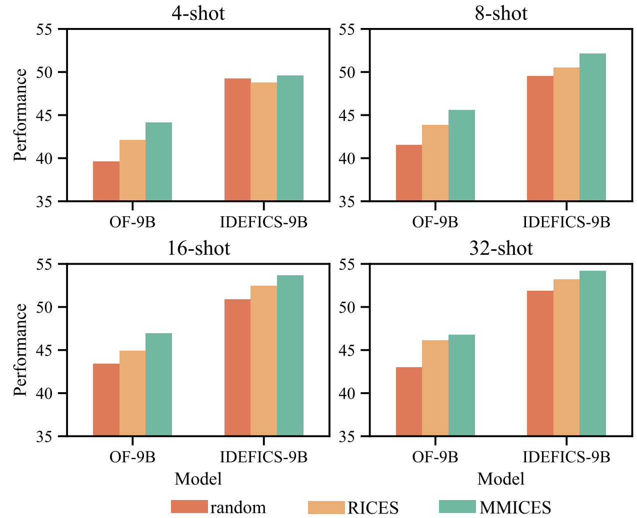



Figure 7. The performance of ICL (on OK-VQA) is consistently enhanced by MMICES across different models.

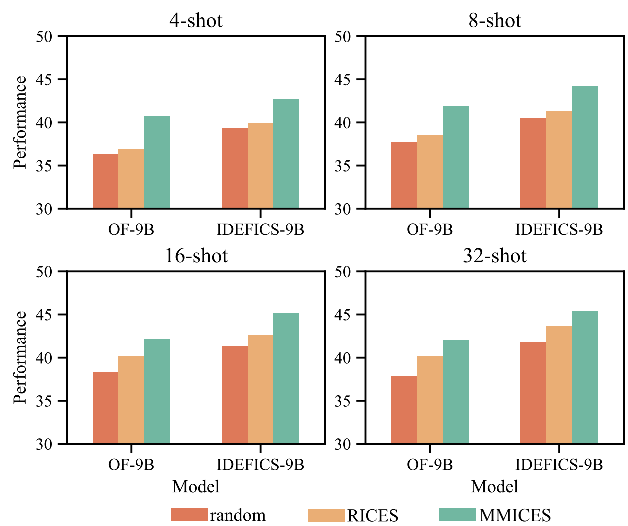



Figure 8. The performance of ICL (on GQA) is consistently enhanced by MMICES across different models.

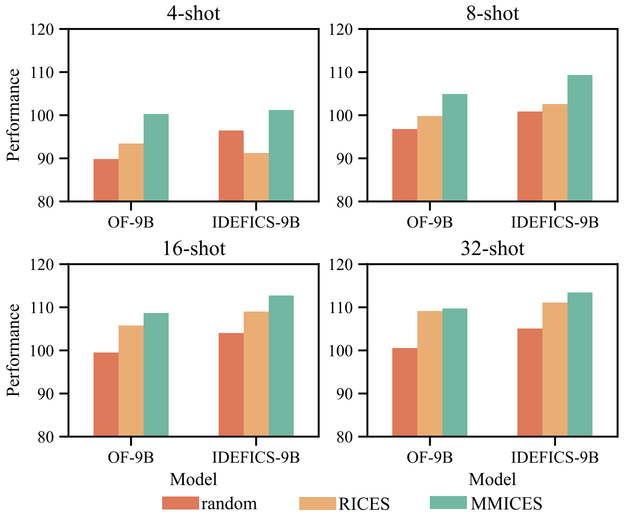



Figure 9. The performance of ICL (on COCO) is consistently enhanced by MMICES across different models.

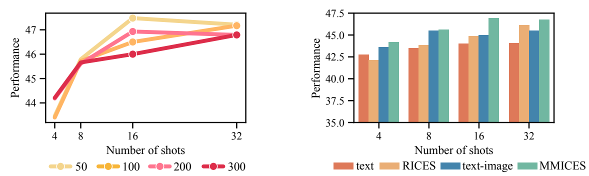



Figure 10. Comparison of performance on OK-VQA given different $K$ (left) and different mixture of modality (right).

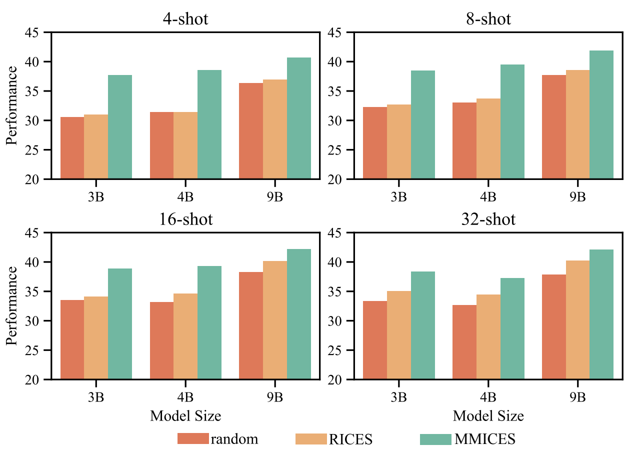



Figure 11. MMICES consistently enhances the ICL performance across models of varying sizes. MMICES on smaller models can even outperform RICES on larger models. Results here are from GQA and more results are in Supplementary Section 4.

More analysis is presented in Supplementary Section 5.

## 6. Limitations

This paper primarily focuses on MLLMs with the masked cross-attention mechanism, *i.e.*, Flamingo [1, 3] and IDEFICS [7]. These are the first batch of models that support interleaved input and in-context learning, making

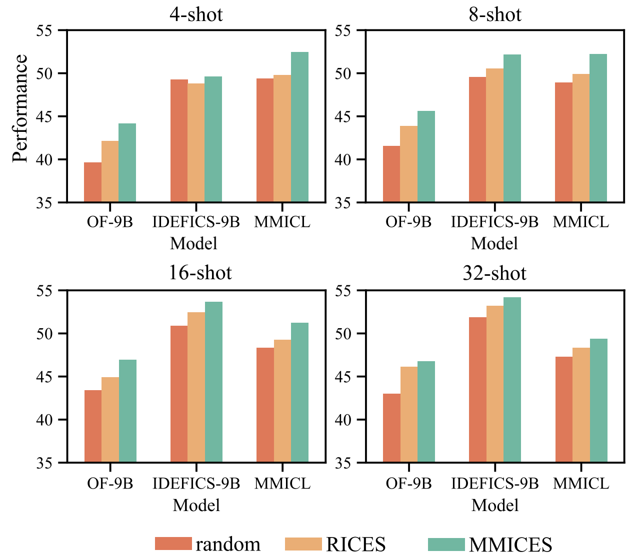



Figure 12. The performance of ICL (on OK-VQA) is consistently enhanced by MMICES across different models, including OpenFlamingo [3], IDEFICS [7], and MMICL [18].

them significant for the community and the primary focus of this study. However, recent developments have introduced more MLLMs with various architectures, such as IDEFICS-2 [8], which uses auto-regressive structure and possesses in-context learning capabilities. Evaluating these models further is part of our future agenda. Additionally, the evaluation tasks in this study are traditional vision-language tasks such as visual question answering [5] and image captioning [4]. Although these tasks are not specifically designed for in-context learning, they are typically used to demonstrate the in-context learning ability of MLLMs, which is why they were chosen for this study. Incorporating more tasks and datasets is another step we plan to take in the future.

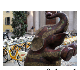
| Query | Method | Demo 1 | Demo 2 | Demo 3 | Model Generation |
|---|---|---|---|---|---|
|  Who makes the guitar on the wall? | MMICES | Who makes the luggage in this room?<br>samsonite | Who invented the device pictured?<br>steve job | Who manufactures this bag?<br>ll bean | fender |
| | RICES | What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

Table 23. Examples of demonstrations selected by MMICES and RICES on OK-VQA. Model generations in green are correct and red means wrong prediction.

## References


[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 1, 2, 3, 9

[2] 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*, pages 2425–2433, 2015. 2

[3] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1, 9

[4] 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. 1, 2, 9

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1, 9

[6] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 2

[7] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023. 1, 7, 9

[8] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 9

[9] 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. $M^3$it: A large-scale dataset towards multi-modal multilingual instruction tuning. *arXiv preprint arXiv:2306.04387*, 2023. 1

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1

[11] 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. 1, 2

[12] 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*, pages 8748–8763. PMLR, 2021. 1

[13] 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. 1

[14] 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. 1

[15] 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. 1

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

[17] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[18] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 1, 9

[19] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae 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*, 2023. 1