When and How to use In-context Learning for modern LVLMs? An Extensive Investigation

Anonymous ACL submission

Abstract

Multi-Modal In-Context Learning (MM-ICL) is commonly employed to enhance performance of Large Vision-Language Models (LVLMs) on downstream tasks. However, as next-generation models like QwenVL and InternVL have been optimized through instruction tuning and alignment, their baseline capabilities have been strengthened. This necessitates a re-evaluation of the effectiveness and optimal configuration of MM-ICL across different models. By evaluating diverse LVLMs spanning different parameter scales (7B–32B) and accessibility settings (open-source and closed-source) across eight cross-domain datasets, this paper systematically investigates the performance of MM-ICL under varying model capabilities and task settings. Experimental results indicate that, beyond validating established patterns, next-generation models exhibit distinct phenomena: in certain logical reasoning and format-sensitive tasks, the introduction of MM-ICL does not necessarily yield improvements and can even lead to performance degradation in specific scenarios. We analyze these observations and, based on extensive experimental data, provide new recommendations for selecting MM-ICL strategies for LVLMs.

1 Introduction

Large Vision-Language Models (LVLMs) (Qin et al., 2025; Alayrac et al., 2022; Laurençon et al., 2024; Liu et al., 2023; Wang et al., 2024a) have evolved rapidly, exhibiting exceptional performance across various multimodal understanding and reasoning tasks (Li and Lu, 2024). To extend these models to new tasks without incurring high fine-tuning costs, Multimodal In-Context Learning (MM-ICL) has emerged as a prevalent paradigm (He et al., 2023; Zhang et al., 2023). By simply prepending a few retrieved image-text demonstration pairs to the query, MM-ICL activates the model's potential, guiding it to capture task intent and standardize output formats.

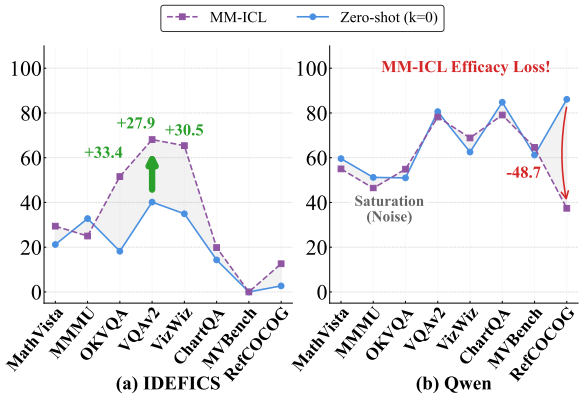

Figure 1: The evolving impact of MM-ICL across model generations. (a) The early model IDEFICS relies on ICL for significant gains. (b) In contrast, the modern Qwen2.5-VL exhibits performance comparable to its zero-shot baseline, indicating diminished ICL utility.

Prior work has demonstrated the effectiveness of MM-ICL on early commonly used LVLMs (e.g., Flamingo and IDEFICS) among traditional vision-language tasks (primarily VQA tasks (Sidorov et al., 2020; Goyal et al., 2017; Marino et al., 2019) and captioning tasks (Lin et al., 2014; Sidorov et al., 2020)). With the rapid advancement of LVLMs, overall model performance has improved substantially and the application scenarios have become increasingly complex. However, whether the MM-ICL paradigm remains effective across modern models and diverse domains has not yet been systematically investigated. Based on our observations using LVLMs with a pilot study, the performance of MM-ICL on next-generation models deviates somewhat from previous empirical findings. Specifically, as shown in Figure 1, when applying traditional MM-ICL methods to newer models, the performance gains on certain tasks were marginal or even deteriorated. This inspires us to revisit the application of MM-ICL on contemporary models and tasks through the lens of ICL theory.

In this paper, we aim to investigate the question:

When and How to use MM-ICL for modern LVLMs? To further answer three critical research questions: **(RQ1)** *When is MM-ICL effective or ineffective*, **(RQ2)** *How to configure MM-ICL for modern models*, **(RQ3)** *How do Post-Training and Explicit Inputs affect MM-ICL?* With these questions in mind, we conducted a large-scale empirical study focused on optimal MM-ICL configuration strategies. Our evaluation is comprehensive in both breadth and depth, covering representative LVLMs spanning diverse capability levels, parameter scales (from 7B to 32B), and accessibility settings (open-source and closed-source), including the latest state-of-the-art models such as GPT-5. We further evaluate these models across eight cross-domain benchmarks, including mathematical reasoning, visual grounding, video understanding, and VQA tasks with varying focuses. Through systematically evaluating mainstream training-free MM-ICL strategies across different tasks and models, we disentangle the relationship between model capabilities, task characteristics, and MM-ICL configurations.

Based on these extensive evaluations, we summarize several key empirical findings corresponding to the preceding three research questions: (1) The role of MM-ICL in contemporary LVLMs appears to be **increasingly task-dependent** rather than uniformly beneficial. MM-ICL helps in distribution-shifted settings like VizWiz, but on complex reasoning tasks such as MathVista, stronger models may see little or even negative benefit from demonstrations. This suggests that for strong zero-shot models, in-context examples can sometimes disrupt rather than improve performance. (2) Our comparative analysis shows that, for several contemporary stronger models evaluated in this study, the performance differences among retrieval strategies become relatively small, which is different from the phenomenon described in early models (Li et al., 2024b). In some settings, simple retrieval strategies and even random sampling achieve performance comparable to more sophisticated methods. (3) Our analysis further suggest that the performance of MM-ICL may vary after the model undergoes post-training. Within the evaluated model families, instruction tuning improves zero-shot instruction-following performance while reducing the additional gains obtained from in-context demonstrations. Preference-aligned variants also exhibit task-dependent changes in zero-shot and few-shot performance.

The main contributions of this paper can be summarized as follows:

- We conduct extensive MM-ICL evaluations across diverse LVLMs and multimodal task domains, providing a systematic comparison of how contemporary models and earlier baselines benefit from in-context demonstrations.
- We identify task- and model-dependent behaviors of MM-ICL, showing that demonstrations remain beneficial in several settings while yielding limited or negative gains for some stronger models on specific reasoning and visual grounding tasks.
- We examine the relationship between post-training stages and MM-ICL behavior, and provide exploratory evidence that how explicit demonstration context contribute to performance degradation in selected reasoning settings.

2 Related Work

In-context learning (ICL) has emerged as a powerful paradigm where large language models (LLMs) can solve new tasks by being given a few example demonstrations in a prompt rather than being fine-tuned (Brown et al., 2020; Rubin et al., 2022; Zhang et al., 2023; Olsson et al., 2022) Recently, investigations on LVLMs demonstrate that the ICL paradigm can handle multimodal tasks (Li et al., 2025a; Awadalla et al., 2023; Baldassini et al., 2024)

In practice, the effectiveness of ICL is influenced by a multitude of factors. Prior research in NLP has extensively investigated the influence of demonstrations, selection (Liu et al., 2022; Rubin et al., 2022), ordering (Lu et al., 2022), and prompt design on ICL performance. Similar studies have recently been extended to Multi-modal ICL (MM-ICL) (Li et al., 2024b; Chen et al., 2025). However, these studies remain confined to traditional pre-trained models. Whether similar conclusions apply to the latest generation of LVLMs, which incorporate extensive post-training techniques, remains an open question.

Complementary to these empirical studies on demonstration selection and prompt design, a growing body of theory interprets ICL as an implicit algorithm executed inside the Transformer.Early analyses on linear models show that Transformer-based in-context learners can implement or ap-

proximate standard estimators such as gradient descent, ridge regression, least-squares regression, and Bayesian predictors, with the effective algorithm depending on depth, noise level, and architecture design (Akyürek et al., 2023; von Oswald et al., 2023). Subsequent work further shows that trained Transformers can learn stable algorithmic procedures at inference time, implement preconditioned gradient descent, and even perform in-context algorithm selection across different data-generating processes (Li et al., 2023; Ahn et al., 2023; Bai et al., 2023). Recent analyses of linear attention and linear Transformer blocks additionally clarify how context length, pretraining diversity, and architectural components such as MLP layers shape the risk and approximation behavior of ICL (Zhang et al., 2024a,b; Lu et al., 2024c). These theoretical results mainly study simplified unimodal regression settings, but they provide a useful lens for understanding MM-ICL: demonstrations may not merely provide superficial examples; they can induce an inference-time update or algorithmic bias that interacts with the model's pretrained and post-trained priors.

To further optimize MM-ICL for enhanced efficacy and stability across an expanding range of novel scenarios, researchers have introduced various optimization frameworks. Some approaches employ lightweight auxiliary models to autoregressively select and arrange demonstrations, capturing task-specific mappings and ordering dependencies for optimal sequence configuration (Yang et al., 2024; Li, 2025; Li et al., 2025b). Some works focus on inference efficiency by compressing demonstrations into learnable latent vectors or attention biases, effectively mimicking the in-context learning mechanism through internal state shifting without processing raw examples (Peng et al., 2024; Jiang et al., 2025). These methods typically necessitate the construction of large-scale, high-quality supervised datasets for training, which imposes significant computational overhead and data engineering challenges. However, these model- or data-specific strategies cause the model to lose generalization and flexibility when deployed in real applications. In this paper, we focus on training-free scenarios and provide practical guidelines regarding adaptability and usage principles by analyzing the performance of MM-ICL across various tasks and LVLMs with different training strategies.

3 Preliminaries

The Paradigm of In-context Learning. The ICL paradigm refers to how LLMs can perform new downstream tasks simply by being given a prompt consisting of a few demonstrations. We formalize this problem as follows: Given the task and its a set of prepared demonstrations $\mathcal{D} = \{d_1..., d_n\}$, each $d_i = (x_i, y_i)$ includes input x_i and output y_i. For a test query x_q, ICL constructs a prompt $\mathcal{P}$ that contains several demonstrations $\{\hat{d}_i\} \subseteq \mathcal{D}$, with which the LLMs generate the corresponding output y_q. In this work, we focus on the multimodal ICL (MM-ICL), where each input x_i has both textual and visual tokens.

The Explanation of In-context Learning. The mechanistic essence of ICL is formally conceptualized as a process of *implicit parameter optimization* (Dherin et al., 2025). Under this framework, a transformer block is viewed as a specific instance of a *contextual block*, which consists of a *contextual layer* $A(\cdot)$ (generalizing structures such as self-attention or RNNs) stacked with a subsequent neural layer M_W (typically an MLP). The contextual layer implicitly transforms the provided demonstrations $C = \{\hat{d}_1, \ldots, \hat{d}_n\}$ into a low-rank weight update for the weight matrix W of the neural layer. This allows the model to behave as if it has undergone a task-specific fine-tuning step at inference time without explicit gradient updates. Formally, the output of a contextual block on query x_q conditioned on context C, denoted as $\mathrm{T}_W(C, x_q)$, is equivalent to the execution of an updated weight matrix $W' = W + \Delta W$ on the query alone:

$$\mathrm{T}_W(C, x_q) = \mathrm{T}_{W+\Delta W}(x_q), \tag{1}$$

where ΔW is a rank-1 implicit update derived from the context:

$$\Delta W = \frac{(W\Delta A)A(x_q)^\top}{\|A(x_q)\|^2}. \tag{2}$$

In this formulation, $A(x_q)$ represents the output of the contextual layer without context, and $\Delta A = A(C, x_q) - A(x_q)$ is the *context vector* that captures task-specific information from the demonstrations.

Crucially, this implicit fine-tuning is not a single global event but is executed independently and sequentially across the stacked layers of the model. As the context information flows through each successive block, the contextual layer of the l-th layer generates a local weight update $\Delta W^{(l)}$ based on its

specific hidden representations. This update reconfigures the mapping rules of the l-th MLP, and the resulting output serves as the refined input for the $(l+1)$-th layer. This layer-wise recursive process identifies the processing of demonstrations in ICL with implicit gradient descent dynamics, whereby the entire model dynamically re-organizes its internal logic to adapt to the intended multimodal task.

4 When and How to use MM-ICL for modern LVLMs?

As LVLMs evolve from pre-trained baselines to sophisticated, instruction-tuned, and human-aligned systems, the prevailing assumption that "more context yields better performance" requires rigorous re-examination. To systematically address the research questions, we establish a three-dimensional analytical framework. This framework deconstructs the interactions among task characteristics, configuration strategies, and the intrinsic priors of strong models, formulating specific hypotheses to guide our experimental validation.

4.1 RQ1: When is MM-ICL effective or ineffective?

While ICL is widely recognized for its ability to facilitate rapid adaptation to downstream tasks, its effectiveness is not uniform across all scenarios. We aim to identify specifically *which* models benefit from ICL and in *which* task domains this paradigm remains applicable.

Based on the theory of implicit weight updates discussed in Section 3, ICL can be characterized as a process of inference-time learning. Since demonstrations do not explicitly provide the ground truth for the test query, the model must rely on implicit parameter updates to rapidly identify the task format and requirements. However, modern LVLMs have already established robust reasoning capabilities and logical priors through extensive post-training, such as instruction tuning and RLHF. Extrapolating from this theory, for the tasks where the model is already highly fitted, the additional weight updates triggered by in-context demonstrations may disrupt the model's established logic, leading to performance degradation. Conversely, for out-of-distribution (OOD) tasks with unique formats where the model's post-training is insufficient, this mechanism serves as a crucial adaptation tool to bridge the domain gap and enhance performance.

To verify this, we curate diverse datasets ranging from General VQA to Complex Reasoning tasks. These benchmarks cover distinct data distributions and cognitive demands. By evaluating varying model architectures against these datasets, we systematically investigate the boundaries of ICL effectiveness, determining where it serves as a catalyst and where it might act as a hindrance.

4.2 RQ2: How to configure MM-ICL for modern models?

How to best configure in-context demonstrations remains a widely discussed topic. Previous efforts have analyzed the factors affecting MM-ICL effectiveness (Qin et al., 2024), and specifically explored demonstration strategies for VQA (Li et al., 2024b). While these studies established foundational paradigms, they were largely restricted to earlier model architectures and traditional tasks. To complement and extend these findings, our study investigates whether such configurations remain effective for next-generation frontier models across a more diverse range of benchmarks, including complex logical reasoning and format-sensitive scenarios. To address this, we conduct extensive studies utilizing multiple retrieval strategies (e.g., random vs. semantic) and varying shot counts. Our goal is to explore the "configuration boundarie" of MM-ICL and determine whether new models exhibit different sensitivities to context quality and quantity compared to their predecessors.

4.3 RQ3: How do Post-Training and Explicit Inputs affect MM-ICL?

Our third research question investigates why different LVLMs may exhibit different responses to in-context demonstrations. Modern LVLMs often undergo multiple post-training stages, such as instruction tuning and preference alignment, which may alter their instruction-following behavior, reasoning patterns, and sensitivity to additional context.

To investigate the specific impact of these post-training stages on ICL performance, we design comparative experiments to isolate the effects of instruction tuning and preference alignment, analyzing whether "alignment" comes at the cost of in-context learnability.

Table 1: Comprehensive performance comparison of LVLMs across five diverse task domains. The "↑↑" metric indicates the absolute gain of the best few-shot performance over the zero-shot baseline. **Bold** and underline denote the highest and second-highest improvements within each benchmark, respectively.

		Large Visual Language Models						
		LLaVa-1.5	**Qwen2.5-VL-7B**	**InternVL3**	**Deepseek-VL**	**IDEFICS**	**Qwen2.5-VL-32B**	**GPT-5**
***TASK 1: Logical Reasoning* (Accuracy %)**								
MathVista	Zero-shot	25.80	60.40	70.00	36.40	21.20	64.40	79.40
	MM-ICL	33.80	57.60	66.60	40.80	36.20	62.40	79.80
	↑↑	8.00	-2.80	-3.40	4.40	**15.00**	-2.00	0.40
MMMU	Zero-shot	33.30	51.20	56.10	35.30	32.80	57.30	62.00
	MM-ICL	43.80	51.60	56.56	35.40	32.80	61.80	63.00
	↑↑	**10.50**	0.40	0.46	0.10	0.00	4.50	1.00
***TASK 2: General VQA* (Consensus Accuracy %)**								
OKVQA	Zero-shot	56.24	50.98	56.20	57.51	18.18	53.69	35.72
	MM-ICL	68.63	63.51	69.23	67.83	62.15	62.40	57.24
	↑↑	12.39	12.53	13.03	10.32	**43.97**	8.71	21.52
VQAv2	Zero-shot	74.18	80.60	79.67	77.88	40.15	78.64	63.50
	MM-ICL	84.67	80.90	85.63	84.80	80.50	79.76	80.02
	↑↑	10.49	0.30	5.96	6.92	**40.35**	1.12	16.52
VizWiz	Zero-shot	59.56	62.54	49.23	49.19	34.94	63.18	53.36
	MM-ICL	71.35	73.91	74.71	72.64	71.70	68.68	65.44
	↑↑	11.79	11.37	25.48	23.45	**36.76**	5.50	12.08
***TASK 3: Chart & Doc Understanding* (Relaxed Accuracy %)**								
ChartQA	Zero-shot	19.40	84.80	78.30	65.80	14.30	83.96	79.90
	MM-ICL	29.30	84.80	83.60	63.30	27.40	84.53	87.24
	↑↑	9.90	0.00	5.30	-2.50	**13.10**	0.57	7.34
***TASK 4: Visual Grounding* (Acc@IoU≥0.5 (%))**								
RefCOCOg	Zero-shot	34.90	86.10	61.60	47.50	2.70	79.22	5.60
	MM-ICL	26.40	61.60	76.90	35.40	13.70	86.10	9.40
	↑↑	-8.50	-24.50	**15.30**	-12.10	11.00	6.88	3.80
***TASK 5: Video Understanding* (Accuracy %)**								
MVBench	Zero-shot	–	60.00	67.00	–	–	55.60	59.60
	MM-ICL	–	65.60	75.60	–	–	63.30	64.40
	↑↑	–	5.60	**8.60**	–	–	7.70	4.80
Average Improvement		7.80	0.36	8.84	4.37	**22.88**	4.12	8.43

5 Experimental Setup

Models and Tasks. We evaluate seven LVLM configurations from six model families, including open-weight models ranging from 7B to 32B and the closed-source frontier model. specifically IDEFICS (Laurençon et al., 2023), LLaVA-1.5 (Liu et al., 2024), Deepseek-VL (Lu et al., 2024a), Qwen2.5-VL-7B, Qwen2.5-VL-32B (Bai et al., 2025), and InternVL3 (Zhu et al., 2025) and GPT-5. Our evaluation spans five distinct task categories: General VQA (Marino et al., 2019; Goyal et al., 2017; Bigham et al., 2010), Reasoning (Lu et al., 2024b; Yue et al., 2024), Visual Grounding (Mao et al., 2016), Chart Reasoning (Masry et al., 2022), and Video Understanding (Li et al., 2024a), utilizing task-specific metrics for assessment. Note that video understanding tasks are exclusively conducted on models equipped with native video capabilities.

MM-ICL Configuration. Following mainstream strategies (Li et al., 2024b; Qin et al., 2024), we construct the final input sequence by combining a system prompt, retrieved demonstrations, and the user query, all formatted according to each model's specific conversation template. To investigate the impact of example selection, we compare five strategies: *Random Select* (RS); *Visual Retrieval* (VR) based on image similarity; *Text Retrieval* (TR) utilizing question semantics; *Semantic Pair Retrieval* (SPR) focusing on QA logic; and *Multimodal Retrieval* (MR) via joint embeddings. Comprehensive experimental settings are detailed in Appendix A.

6 Results Analysis

6.1 When is MM-ICL Effective or Ineffective?

Overall effects of MM-ICL across models. Table 1 presents the MM-ICL performance of different LVLMs across five task categories. The results show that the overall effect of MM-ICL differs substantially across models. IDEFICS, an earlier model with relatively limited zero-shot capability, obtains a large average improvement of 22.88 points after introducing in-context demonstrations. In comparison, the gains are smaller and

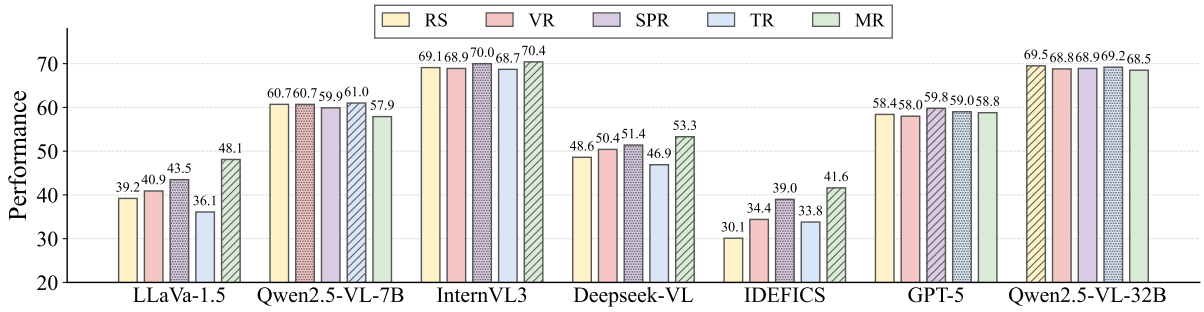

Figure 2: Model Performance Across Different Strategies. bars with a **dotted pattern** indicate the best result, while bars with a **diagonal hatched pattern** represent the second-best result within each model group

more task-dependent for several stronger models. Qwen2.5-VL-7B shows only a marginal average improvement of 0.36 points, while Qwen2.5-VL-32B and GPT-5 achieve average improvements of 4.12 and 8.43 points, respectively. These results indicate that MM-ICL is not uniformly beneficial for stronger models. While its overall gains are limited for models such as Qwen2.5-VL-7B and Qwen2.5-VL-32B, GPT-5 still benefits noticeably on several tasks.

LoRA-based Implicit-Update Proxy. Following §3, we construct a LoRA-based implicit-update proxy to approximate the adaptation effect induced by demonstrations. Specifically, the retrieved demonstrations are used to perform a lightweight LoRA update on the model's MLP layers, while inference is conducted without including demonstration tokens in the input context. We compare this proxy with standard MM-ICL on 20 randomly selected MathVista validation samples using Qwen2.5-VL-7B and InternVL3-8B. As shown in Table 2, the LoRA-based proxy does not improve over the zero-shot baseline for either model: Qwen2.5-VL-7B remains at 55.0%, while InternVL3-8B decreases from 60.0% to 55.0%. Moreover, for Qwen2.5-VL-7B, standard MM-ICL further reduces accuracy to 40.0%, whereas the LoRA-based proxy preserves the zero-shot result. These results suggest that, on this reasoning subset, demonstrations provide limited additional adaptation benefits for the two strong models, while explicitly retaining them in the inference context can further impair performance in some cases.

The Task-specific Suitability of MM-ICL. As shown by the cross-row comparisons in Table 1, the effectiveness of MM-ICL varies substantially across task categories. In logical reasoning tasks, MM-ICL provides limited benefits for several stronger models and can even lead to performance degradation. On MathVista, Qwen2.5-VL-7B, Qwen2.5-VL-32B, and InternVL3-8B decrease by 2.80, 2.00, and 3.40 points, respectively, while GPT-5 achieves only a marginal gain of 0.40 points. On MMMU, Qwen2.5-VL-7B, InternVL3-8B, and GPT-5 obtain relatively small gains of 0.40, 0.46, and 1.00 points, respectively, whereas Qwen2.5-VL-32B improves by 4.50 points. These results suggest that, for reasoning-intensive tasks, additional demonstrations do not consistently provide substantial benefits to models with strong zero-shot performance.

Table 2: Preliminary comparison between standard MM-ICL and LoRA-based demonstration adaptation on a 20-sample MathVista subset.

Model	Method	Accuracy
Qwen2.5-VL-7B	Zero-shot (Baseline)	55.0%
	MM-ICL (4-shot)	40.0%
	LoRA-based Adaptation (4-shot)	55.0%
InternVL3-8B	Zero-shot (Baseline)	60.0%
	MM-ICL (4-shot)	60.0%
	LoRA-based Adaptation (4-shot)	55.0%

In contrast, MM-ICL remains beneficial in several general VQA and distribution-shifted settings. On VizWiz, which contains low-quality images and unanswerable questions, Qwen2.5-VL-7B, InternVL3-8B, Qwen2.5-VL-32B, and GPT-5 improve by 11.37, 25.48, 5.50, and 12.08 points, respectively. GPT-5 also achieves substantial gains on OKVQA and VQAv2, improving by 21.52 and 16.52 points. Moreover, on MVBench, MM-ICL improves Qwen2.5-VL-7B, InternVL3-8B, Qwen2.5-VL-32B, and GPT-5 by 5.60, 8.60, 7.70, and 4.80 points, respectively. These results indicate that MM-ICL remains useful when demonstrations help models accommodate distribution shifts or satisfy task-specific output requirements.

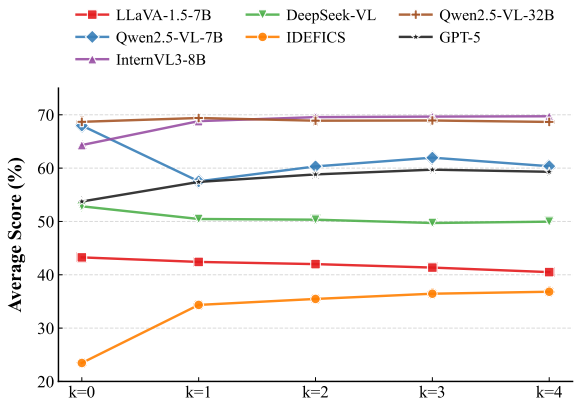

Figure 3: Average performance over available benchmarks of different models across varying k-shot settings.

Summary. In summary, MM-ICL is strongly dependent on both model and task. While earlier models like IDEFICS benefit clearly, stronger models show more selective gains. It often helps little or even hurts on reasoning-heavy tasks, but remains useful on VQA and distribution-shifted tasks such as VizWiz, even for strong models like GPT-5. Therefore, MM-ICL should be applied selectively rather than assumed to be universally beneficial.

6.2 How to configure MM-ICL for modern models?

Retrieval Strategy. Figure 2 presents a comparative analysis of various LVLMs across different retrieval strategies. The experimental results show that MR and SPR generally achieve strong performance, ranking as either the best or second-best strategies across multiple models. This indicates that demonstrations retrieved using multimodal and semantic information remain effective choices for MM-ICL. Meanwhile, the influence of retrieval strategy becomes less pronounced for stronger models. Specifically, the gap between the best and worst strategies is **11.5** points for IDEFICS, but decreases to **3.1**, **1.7**, **1.8**, and **1.0** points for Qwen2.5-VL-7B, InternVL3-8B, GPT-5, and Qwen2.5-VL-32B, respectively. These results suggest that, although MR and SPR remain generally reliable strategies, modern LVLMs are less sensitive to the specific retrieval method used.

Demonstration Number. Figure 3 presents the average scores achieved by different models under various k-shot configurations. Model behaviors vary significantly across different architectures. **IDEFICS** shows consistent improvement as the number of demonstrations increases. Conversely, models with high zero-shot baselines, such as **Qwen2.5-VL-7B**, experience a performance decline. This suggests that external examples may conflict with their internal priors. The most significant performance shift occurs during the transition from zero-shot to one-shot. Beyond $k = 1$, performance gains diminish or stabilize for most models. Therefore, a single demonstration is often sufficient to define the task format, while additional examples provide limited benefit.

Table 3: Comparison of performance gains for IDEFICS2-Base and IDEFICS2-Instruction under Zero-shot with instruction (Inst. Gain) and 2-shot (ICL Gain) settings.

Dataset	Inst. Gain (Zero-shot)		ICL Gain (2-shot)	
	Base	Instruct	Base	Instruct
VQAv2	-29.9	+0.4	+22.4	-1.7
OKVQA	-13.6	-1.6	+30.2	+0.6
VizWiz	-12.9	+16.7	+50.7	+30.0
MMMU	-2.4	+5.4	-14.2	+2.2
ChartQA	0.0	0.0	+10.5	+1.1
RefCOCOg	+0.4	+1.8	+2.1	-0.5
MathVista	+3.0	-2.5	+5.5	+0.8
Average	**-7.9**	**+2.9**	**+15.3**	**+4.6**

Summary. In conclusion, for modern LVLMs, the dependency on specific retrieval algorithms decreases as model capabilities scale. We therefore suggest prioritizing generally robust retrieval strategies as a sufficient baseline. Instead, the focus should shift to strictly limiting demonstration quantity. Our findings show that heavily instruction-tuned models perform best with a 1-shot setup, as additional examples tend to introduce contextual noise and degrade zero-shot priors. Thus, the "Robust Retrieved 1-shot" configuration emerges as the most efficient and stable strategy for real-world applications.

6.3 How Post-Traing influence the MM-ICL performance

MM-ICL under Instruction Tuning. To quantify the influence of instruction tuning on ICL performance, we conduct a comparative analysis using the base and instruction-tuned versions of the IDEFICS2 model family. As illustrated in Table 3, when examining zero-shot performance with instructions alone, the base model suffers an average decline of 7.9%, while the instruction-tuned model yields a 2.9% gain. This indicates that the instruction tuning indeed significantly enhances the model's ability to follow natural language instructions. However, the instruction-tuned model

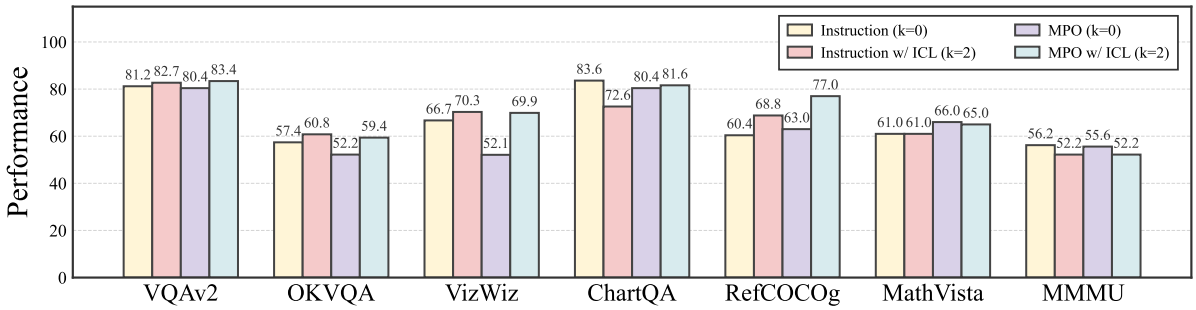

Figure 4: Comparison of Instruction Tuning Model and MPO Alignment Model. MPO stands for InternVL3 with MPO alignment; k denotes the number of demonstrations ($k = 0$ means non-ICL).

demonstrates a significantly lower performance gain from MM-ICL compared to the base model. Specifically, the IDEFICS2-Base model achieves a significant average gain of 15.3% under the MM-ICL setting, whereas the IDEFICS2-Instruction model shows a limited improvement of only 4.6%. This comparison suggests that instruction tuning improves instruction-following performance while reducing the additional gains obtained from MM-ICL within the IDEFICS2 model family.

MM-ICL under Preference Alignment. To further investigate the impact of preference alignment on MM-ICL, we perform a comparative analysis between the original instruction-tuned and MPO versions of InternVL3 models (Zhu et al., 2025), the latter of which incorporates an additional preference alignment stage (Wang et al., 2024b). As illustrated in Figure 4, in the zero-shot setting ($k = 0$), InternVL3-MPO exhibits a performance decline compared to the instruction version on several tasks, especially the format-sensitive tasks (i.e., VizWiz and ChartQA). For instance, the VizWiz score drops notably from 66.7 to 52.1, and the ChartQA score drops from 83.6 to 80.4. Conversely, on the reasoning-intensive task MathVista, the MPO-aligned model achieves a superior zero-shot score of 66.0, compared to 61.0 for the instruction-tuned model. However, introducing MM-ICL ($k = 2$) significantly alters this landscape: on tasks like VizWiz, ChartQA and RefCOCOg, InternVL3-MPO gains substantial improvements (e.g., VizWiz rising to 69.9), or even surpassing the instruction baseline (e.g., RefCOCO reaching 77.0). These results indicate that while the MPO alignment process enhances specific reasoning capabilities, it may temporarily weaken zero-shot proficiency in other domains. Crucially, MM-ICL can serve as a key activator, helping the MPO model recapture its suppressed capabilities in certain tasks. This activation diminishing returns in domains where the model is already highly optimized.

Summary. Overall, post-training changes how LVLMs benefit from in-context demonstrations. Instruction tuning reduces the additional MM-ICL gains in the IDEFICS2 comparison, while preference alignment changes the task-wise behavior of InternVL3 and leaves room for MM-ICL to recover performance on selected tasks. These findings are consistent with our observations in Section 6.1: for modern LVLMs, the usefulness of MM-ICL depends on the model state and the target task.

7 Conclusion

This study systematically investigates the effectiveness and configuration of Multi-Modal In-Context Learning (MM-ICL) for modern Large Vision-Language Models (LVLMs). Our experiments across seven LVLM configurations from six model families and eight cross-domain benchmarks show that MM-ICL remains beneficial for several out-of-distribution and format-sensitive tasks, while providing limited or even negative gains on some complex reasoning tasks for stronger models. Furthermore, we find that as model capabilities scale, the performance disparity between advanced retrieval strategies and random sampling diminishes, with a demonstration often being sufficient to define task intent.Meanwhile, our controlled analyses suggest that post-training stages influence how models benefit from in-context examples We hope these findings will provide practical recommendations for selecting MM-ICL strategies and guide future research in understanding the evolving interaction between context and model capabilities.

Limitation

Although this study evaluates MM-ICL across LVLMs with different capability levels and accessibility settings, several limitations remain. First, our experiments mainly focus on short-form multimodal tasks, such as VQA, visual grounding, and structured reasoning. Whether the observed model- and task-dependent patterns generalize to a broader range of frontier LVLMs or open-ended multimodal interaction scenarios remains to be investigated. Second, we employ standard input-output demonstrations without considering more structured prompting strategies, such as Chain-of-Thought, which may influence MM-ICL performance on reasoning-intensive tasks. Finally, the LoRA-based Implicit-Update Proxy in our auxiliary experiment only provides an approximate analysis of demonstration-induced adaptation. Given its limited evaluation scale, it cannot fully characterize the internal inference dynamics of standard MM-ICL.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. 2023. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Proc. NeurIPS 2022*.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, pages 1539–1550. IEEE.

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and 1 others. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proc. annual ACM symposium on User interface software and technology*, pages 333–342.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proc. NeurIPS*.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. 2025. Can multimodal large language models truly perform multimodal in-context learning? In *Proc. WACV*, pages 6000–6010.

Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, and Javier Gonzalvo. 2025. Learning without training: The implicit dynamics of in-context learning. *ArXiv preprint*, abs/2507.16003.

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 *Proc. CVPR*, pages 6325–6334.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. ICL-D3IE: in-context learning with diverse demonstrations updating for document information extraction. In *Proc. ICCV*, pages 19428–19437.

Yuchu Jiang, Jiale Fu, Chenduo Hao, Xinting Hu, Yingzhe Peng, Xin Geng, and Xu Yang. 2025. Mimic

in-context learning for multimodal tasks. In *Proc. CVPR*, pages 29825–29835.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELICS: an open web-scale filtered dataset of interleaved image-text documents. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? In *Proc. NeurIPS 2022*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025a. Otter: A multi-modal model with in-context instruction tuning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(9):7543–7557.

Jian Li and Weiheng Lu. 2024. A survey on benchmarks of multimodal large language models. *ArXiv preprint*, abs/2408.08632.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024a. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proc. CVPR*, pages 22195–22206.

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. 2024b. How to configure good in-context sequence for visual question answering. In *Proc. CVPR*, pages 26700–26710.

Yanshu Li. 2025. Advancing multimodal in-context learning in large vision-language models with task-aware demonstrations. *ArXiv preprint*, abs/2503.04839.

Yanshu Li, Jianjiang Yang, Tian Yun, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025b. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proc. EMNLP*, pages 736–763.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 19565–19594. PMLR.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Proc. ECCV*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proc. NeurIPS 2022*.

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

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, and 1 others. 2024a. Deepseek-vl: towards real-world vision-language understanding. *ArXiv preprint*, abs/2403.05525.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024b. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proc. ICLR*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proc. ACL*, pages 8086–8098, Dublin, Ireland.

Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. 2024c. Asymptotic theory of in-context learning by linear attention. *CoRR*, abs/2405.11751.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proc. CVPR*, pages 11–20.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proc. CVPR*, pages 3195–3204.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the ACL*, pages 2263–2279, Dublin, Ireland.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

Yingzhe Peng, Chenduo Hao, Xinting Hu, Jiawei Peng, Xin Geng, and Xu Yang. 2024. LIVE: learnable in-context vector for visual question answering. In *Proc. NeurIPS*.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. 2024. What factors affect multimodal in-context learning? an in-depth exploration. In *Proc. NeurIPS*.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. A survey of multilingual large language models. *Patterns*, 6(1):101118.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2655–2671. Association for Computational Linguistics.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: A dataset for image captioning with reading comprehension. In *Proc. ECCV*, volume 12347, pages 742–758.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 35151–35174. PMLR.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv preprint*, abs/2409.12191.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024b. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *CoRR*, abs/2411.10442.

Xu Yang, Yingzhe Peng, Haoxuan Ma, Shuo Xu, Chi Zhang, Yucheng Han, and Hanwang Zhang. 2024. Lever LM: configuring in-context sequence to lever large vision language models. In *Proc. NeurIPS*.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proc. CVPR*, pages 9556–9567.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. 2024a. Trained transformers learn linear models in-context. *J. Mach. Learn. Res.*, 25:49:1–49:55.

Ruiqi Zhang, Jingfeng Wu, and Peter L. Bartlett. 2024b. In-context learning of a linear transformer block: Benefits of the MLP component and one-step GD initialization. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023. What makes good examples for visual in-context learning? In *Proc. NeurIPS*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv preprint*, abs/2504.10479.

A Implementation Details

A.1 Dataset

This section details the dataset splitting strategies and specific evaluation sample sizes employed in our study. For datasets possessing standard partitions (i.e., Training/Validation or Validation/Test splits)—specifically VQAv2, VizWiz, OK-VQA, ChartQA, and RefCOCOg—we designate the original Training or Validation sets as the Support Set for retrieving in-context demonstrations. Consequently, the original Validation or Test sets serve as the final Evaluation Set to assess ICL performance.

For comprehensive benchmarks with specific structures, we adopt the following splitting protocols:

- **MMMU:** We utilize the *Dev* split as the Support Set and the *Val* split as the Evaluation Set, aggregating samples across different categories.
- **MathVista:** The *TestMini* subset is randomly partitioned into Support and Evaluation sets at a 1:1 ratio.
- **MVBench:** Samples from all tasks are mixed and subsequently split with a ratio of 70% for the Support Set and 30% for the Evaluation Set.

To balance evaluation comprehensiveness with computational efficiency, the specific sample sizes used for evaluation are detailed in Table 4.

Dataset	Evaluation Size	Support Size
VQAv2	5,000	443,757
OK-VQA	5,000	9,009
VizWiz	4,319	23,954
ChartQA	1,920	28,299
RefCOCOg	1,000	7,573
MMMU	900	150
MathVista	500	500
MVBench	500	2,660

Table 4: Overview of the evaluation sample sizes across different datasets.

A.2 ICL Configuration Strategies

To systematically investigate the impact of demonstration selection, we implement five distinct retrieval strategies. These strategies range from stochastic selection to advanced multimodal matching:

- **Random Select (RS):** As a baseline, examples are randomly sampled from the support corpus without considering the input query.
- **Visual Retrieval (VR):** This method utilizes a pre-trained visual encoder (e.g., CLIP) to identify and retrieve examples that are visually coherent with the query image or video.
- **Text Retrieval (TR):** We compute semantic vector similarity between the textual component of the user's query and the questions in the support set to find semantically relevant examples.
- **Semantic Pair Retrieval (SPR):** To capture both the task logic and the expected answer format, this strategy calculates the similarity between the user's input question and the candidate "Question-Answer" pairs.
- **Multimodal Retrieval (MR):** This approach employs a unified multimodal encoder to map image-text pairs into a shared embedding space, facilitating the retrieval of examples based on holistic multimodal similarity.

Encoder Implementation. To enable the retrieval strategies described above, we employ specific encoders tailored to different modalities:

- **Text-Encoder:** We utilize the **Qwen3-embedding** model to encode textual queries and captions.
- **Visual-Encoder:** For static images, we employ **CLIP (ViT-L/14)** for feature extraction. For video inputs, features are extracted using **InternVideo2**.
- **Multi-modal Encoder:** We leverage **BLIP-2** to extract joint embeddings for image-text pairs, ensuring alignment across modalities.

A.3 Prompt Construction and Inference Settings

We explicitly design distinct instructions tailored to different types of tasks to ensure clarity and adherence to the model's capabilities.

General VQA. For standard visual question answering tasks such as **VQAv2** and **OKVQA**, the instruction emphasizes conciseness. The prompt is:

> Instruction: provide an answer to the question. Use the image to answer. Answer with a single word or phrase.

For the **VizWiz** dataset, which contains unanswerable visual queries, we append a specific clause to handle uncertainty:

> Instruction: provide an answer to the question. Use the image to answer. Answer with a single word or phrase. If you don't know the answer, just answer 'unanswerable'.

Video Understanding. For video tasks like **MVBench**, the model is required to attend to temporal dynamics and specific object details. The system prompt is:

> Carefully watch the video and pay attention to the cause and sequence of events, the detail and movement of objects, and the action and pose of persons. Based on your observations, select the best option that accurately addresses the question.

Visual Grounding. For **RefCOCOg**, strict formatting constraints are applied to the output coordinates. The prompt specifies the bounding box normalization and ordering:

> Bounding box coordinates are specified in the format (top-left x, top-left y, bottom-right x, bottom-right y). All values are floating point numbers bounded between 0 and 1. Only answer with box coordinates in this format:(0.20, 0.35, 0.80, 0.55) Please provide the bounding box coordinate of the region this sentence describes:

Reasoning Tasks. For **MMMU**, the instruction varies by question type. For multiple-choice questions, we use:

> Answer with the option's letter from the given choices directly.

Whereas for open-ended questions, the prompt is simply "*Answer the question using a single word or phrase.*".

For **MathVista**, we employ dynamic prompts based on the required answer format. For questions requiring an **integer** answer, the prompt is:

> Please answer the question requiring an integer answer and provide the final value, e.g., 1, 2, 3, at the end.

For questions requiring **floating-point** precision (e.g., two decimal places), we explicitly instruct the model:

> Please answer the question requiring a floating-point number with two decimal places and provide the final value, e.g., 1.23, 1.34, 1.45, at the end.

Similarly, customized prompts are applied for text-based choices, 1-decimal floats, and list outputs to match the evaluation scripts strictly.

Decoding Settings. To eliminate the influence of sampling stochasticity on the evaluation of ICL strategies and to ensure the reproducibility of our results, we employ **Greedy Decoding** (i.e., temperature set to 0) across all model generation phases.

A.4 LoRA-based Implicit-Update Proxy

Building on the analysis in Sec. 3, which posits that ICL is equivalent to implicit parameter updates, we conducted Implicit ICL experiments on Qwen2.5-VL-7B and InternVL3.

LoRA-based Implicit-Update Proxy Configuration Motivated by theoretical analyses that interpret in-context learning as inducing low-rank, inference-time adaptation in simplified settings, we construct a LoRA-based proxy to examine whether demonstration-induced adaptation alone can account for the observed performance changes. This proxy is designed as an operational approximation rather than a direct reconstruction of the internal dynamics of standard MM-ICL. Its configuration is as follows:

- **Target Modules:** We restrict the trainable parameters to the MLP layers within the Transformer blocks. This design follows the implicit-update perspective introduced in Section 3, while keeping the adaptation lightweight and controlled.

- **Rank Constraint:** To approximate the low-rank nature of the hypothesized demonstration-induced update, we employ a Rank-1 LoRA configuration ($r = 1$).

- **Data Selection:** We randomly select 20 query samples from the MathVista evaluation set. For each query sample, we retrieve four demonstrations using the SPR strategy. These demonstrations constitute the query-specific adaptation set used by the proxy model.

Test-Time Adaptation Pipeline The LoRA-based proxy follows a query-specific test-time adaptation pipeline. For each query sample x_q, the procedure is conducted independently as follows:

1. **Initialization:** A temporary LoRA adapter is attached to the frozen base model. The adapter is initialized independently for each query sample, while all original model parameters remain frozen.

2. **Query-Specific Adaptation:** The model is adapted using the four retrieved demonstrations associated with x_q ($k = 4$). We use the **SGD optimizer** with a learning rate of 1×10^{-3}. The demonstrations are processed sequentially, with one gradient update performed for each demonstration, resulting in four adaptation steps for each query sample.

3. **Inference without Demonstration Context:** After adaptation, the resulting LoRA adapter is used to perform inference on x_q without explicitly including the retrieved demonstrations in the inference prompt. This setting allows us to separate the effect of the proxy adaptation from the effect of retaining demonstration tokens in the input context.

4. **Reset:** After inference on x_q, the temporary adapter is removed and discarded. The frozen base model is restored before processing the next query sample, ensuring that no adaptation information is transferred across evaluation instances.

Scope of the Proxy Analysis The proposed LoRA-based proxy is intended to provide an exploratory comparison with standard MM-ICL. Since it introduces explicit optimization steps and is evaluated on a limited subset of MathVista samples, it should not be interpreted as an exact implementation of the internal inference process induced by in-context demonstrations. Instead, the comparison is used to examine whether removing demonstration tokens from the inference context, while retaining an approximate adaptation effect, is associated with different performance behavior in the selected reasoning setting.

B Detailed Experimental Results

This appendix reports the complete per-dataset results underlying the aggregate comparisons in the main paper. Tables 5 and 6 cover reasoning tasks; Tables 7–9 cover general visual question answering; and Tables 10–12 cover structured visual understanding, visual grounding, and video understanding. For all image-based benchmarks, we report seven LVLM configurations under five demonstration selection strategies and $k \in \{0, 1, 2, 3, 4\}$. For MVBench, we evaluate the four video-capable models with $k \in \{0, 1, 2\}$. Since the full tables are large and are intended primarily for verification, they are presented as a consecutive table catalogue; a concise takeaway is placed directly below each table rather than separated into intervening prose sections.

Table 5: **Reasoning Task: MathVista.** Full results under different demonstration selection strategies and shot settings.

Strategy	Shots (k)	LLaVA-1.5-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	InternVL3-8B	Deepseek-VL	IDEFICS	GPT-5
RS	$k = 0$	25.80	60.40	64.40	70.00	36.40	21.20	79.40
	$k = 1$	24.40	53.80	62.00	65.20	37.40	23.00	77.60
	$k = 2$	26.60	52.60	59.80	66.20	36.60	23.40	76.00
	$k = 3$	24.60	52.60	60.20	66.40	38.20	23.60	73.60
	$k = 4$	26.20	52.80	60.80	65.20	37.20	25.00	76.20
	Avg.	**25.52**	**54.44**	**61.44**	**66.60**	**37.16**	**23.24**	**76.56**
VR	$k = 0$	25.80	60.40	64.40	70.00	36.40	21.20	79.40
	$k = 1$	29.20	53.20	62.20	65.00	35.80	27.20	76.00
	$k = 2$	30.00	56.40	59.80	63.80	38.20	27.00	76.80
	$k = 3$	28.40	57.60	59.80	63.40	38.20	27.80	78.20
	$k = 4$	28.80	54.80	59.40	63.20	40.40	27.80	66.80
	Avg.	**28.40**	**56.32**	**61.28**	**65.12**	**37.88**	**26.16**	**76.08**
TR	$k = 0$	25.80	60.40	64.40	70.00	36.40	21.20	79.40
	$k = 1$	29.20	53.80	62.40	65.40	35.40	27.20	79.80
	$k = 2$	29.40	55.40	60.20	65.40	38.20	25.20	78.80
	$k = 3$	28.20	57.00	59.80	64.00	37.00	29.20	79.40
	$k = 4$	30.40	57.60	59.40	63.40	39.80	26.60	77.40
	Avg.	**28.48**	**56.56**	**61.36**	**65.64**	**37.44**	**25.84**	**78.40**
SPR	$k = 0$	25.80	60.40	64.40	70.00	36.40	21.20	79.40
	$k = 1$	33.80	54.20	62.40	66.60	39.40	30.00	78.60
	$k = 2$	32.40	56.20	60.00	65.80	40.80	28.40	79.20
	$k = 3$	30.40	54.60	59.80	62.60	39.60	29.80	78.60
	$k = 4$	28.80	54.80	61.20	62.40	39.40	28.20	66.80
	Avg.	**30.12**	**55.88**	**61.64**	**65.56**	**39.16**	**27.52**	**76.52**
MR	$k = 0$	25.80	60.40	64.40	70.00	36.40	21.20	79.40
	$k = 1$	32.80	47.80	61.80	60.00	34.40	26.80	68.40
	$k = 2$	32.80	46.20	59.80	58.00	34.60	36.20	68.80
	$k = 3$	30.00	45.80	60.00	58.00	33.00	32.20	66.60
	$k = 4$	28.40	44.40	59.60	57.40	32.60	34.40	66.20
	Avg.	**29.84**	**48.76**	**61.16**	**60.76**	**34.24**	**30.16**	**69.88**

Takeaway: MathVista highlights the potential interference of demonstrations in complex reasoning: while selected examples benefit weaker models, Qwen2.5-VL-7B, Qwen2.5-VL-32B, InternVL3-8B, and GPT-5 do not exceed their zero-shot results.

Table 6: **Reasoning Task: MMMU.** Full results under different demonstration selection strategies and shot settings.

Strategy	Shots (k)	LLaVA-1.5-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	InternVL3-8B	Deepseek-VL	IDEFICS	GPT-5
RS	$k = 0$	33.30	51.20	57.30	56.10	35.30	32.80	62.00
	$k = 1$	33.80	50.50	57.80	53.40	31.60	28.00	61.40
	$k = 2$	34.60	47.50	57.80	54.80	31.80	24.50	63.00
	$k = 3$	32.50	47.70	54.80	54.70	28.10	27.30	62.10
	$k = 4$	32.40	50.10	58.20	53.80	29.00	25.00	62.50
	Avg.	**33.32**	**49.40**	**57.18**	**54.56**	**31.16**	**27.52**	**62.20**
VR	$k = 0$	33.30	51.20	57.30	56.10	35.30	32.80	62.00
	$k = 1$	34.30	48.50	55.20	54.00	33.80	24.80	63.00
	$k = 2$	33.50	48.30	56.10	53.40	34.00	22.80	62.00
	$k = 3$	32.80	48.20	55.20	54.60	29.50	26.10	63.00
	$k = 4$	33.30	47.30	56.00	55.50	29.10	25.40	61.30
	Avg.	**33.44**	**48.74**	**55.96**	**54.74**	**32.34**	**26.38**	**62.40**
TR	$k = 0$	33.30	51.20	57.30	56.10	35.30	32.80	62.00
	$k = 1$	33.30	51.60	56.20	56.10	35.40	32.80	61.60
	$k = 2$	33.30	51.20	57.10	56.10	35.30	32.80	61.70
	$k = 3$	33.30	51.30	56.10	56.00	35.40	32.80	60.60
	$k = 4$	33.30	51.00	58.00	56.00	35.30	32.80	61.00
	Avg.	**33.30**	**51.24**	**56.94**	**56.08**	**35.36**	**32.80**	**61.48**
SPR	$k = 0$	33.30	51.20	57.30	56.10	35.30	32.80	62.00
	$k = 1$	34.10	46.80	56.00	53.40	33.10	27.60	61.40
	$k = 2$	34.50	45.50	54.40	52.30	32.70	24.30	62.50
	$k = 3$	35.10	46.80	54.60	53.60	28.80	23.10	62.40
	$k = 4$	33.30	47.40	55.80	54.50	29.40	25.50	61.70
	Avg.	**34.06**	**47.54**	**55.62**	**53.98**	**31.88**	**26.66**	**62.06**
MR	$k = 0$	33.30	51.20	57.30	56.10	35.30	32.80	62.00
	$k = 1$	38.00	33.40	61.80	54.00	33.80	25.40	60.80
	$k = 2$	41.60	34.40	43.30	54.44	31.40	24.40	61.40
	$k = 3$	43.80	35.60	43.60	55.44	30.90	25.50	62.10
	$k = 4$	42.30	36.30	45.70	56.56	30.30	25.10	61.10
	Avg.	**39.80**	**38.18**	**50.34**	**55.31**	**32.36**	**26.64**	**61.58**

Takeaway: MMMU shows a more model-dependent pattern: MR notably improves LLaVA-1.5-7B, whereas the stronger models obtain only small gains or remain close to their zero-shot baselines.

Table 7: **General VQA Task: OK-VQA.** Full results under different demonstration selection strategies and shot settings.

Strategy	Shots (k)	LLaVA-1.5-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	InternVL3-8B	Deepseek-VL	IDEFICS	GPT-5
RS	$k = 0$	56.24	50.98	53.69	56.20	57.51	18.18	35.72
	$k = 1$	54.12	56.70	56.73	57.05	57.50	39.02	45.64
	$k = 2$	54.33	53.75	56.88	57.28	57.44	39.55	45.36
	$k = 3$	53.38	52.52	57.16	57.51	57.69	40.07	47.48
	$k = 4$	52.94	42.12	56.24	57.79	57.66	40.39	48.00
	Avg.	**54.20**	**51.21**	**56.14**	**57.17**	**57.56**	**35.44**	**44.44**
VR	$k = 0$	56.24	50.98	53.69	56.20	57.51	18.18	35.72
	$k = 1$	54.08	57.02	56.60	58.00	56.92	39.80	43.00
	$k = 2$	54.20	53.80	56.46	57.98	57.72	40.90	45.20
	$k = 3$	53.91	51.85	56.79	58.35	57.61	41.73	46.76
	$k = 4$	53.15	43.42	56.07	54.85	58.38	42.66	48.68
	Avg.	**54.32**	**51.41**	**55.92**	**56.81**	**57.63**	**36.65**	**43.80**
TR	$k = 0$	56.24	50.98	53.69	56.20	57.51	18.18	35.72
	$k = 1$	43.93	56.54	56.65	56.48	52.15	43.77	46.28
	$k = 2$	44.52	52.26	56.45	57.63	54.69	45.00	49.56
	$k = 3$	53.82	50.59	55.64	57.75	55.43	46.54	51.00
	$k = 4$	43.11	38.80	55.00	58.28	56.02	46.94	49.32
	Avg.	**48.32**	**49.83**	**55.49**	**57.00**	**55.16**	**40.09**	**46.36**
SPR	$k = 0$	56.24	50.98	53.69	56.20	57.51	18.18	35.72
	$k = 1$	61.42	57.78	56.30	60.26	62.45	49.64	45.60
	$k = 2$	60.55	54.33	56.71	62.64	61.72	51.94	49.64
	$k = 3$	59.16	52.53	57.05	63.54	60.71	53.10	51.72
	$k = 4$	56.38	38.97	56.57	63.80	59.93	52.94	52.60
	Avg.	**58.75**	**50.92**	**56.06**	**61.02**	**60.46**	**45.16**	**46.88**
MR	$k = 0$	56.24	50.98	53.69	56.20	57.51	18.18	35.72
	$k = 1$	68.63	57.03	56.59	61.47	67.83	52.94	46.40
	$k = 2$	68.37	59.78	59.20	66.48	66.68	57.22	52.88
	$k = 3$	65.62	62.70	61.21	68.48	65.00	60.96	56.48
	$k = 4$	61.67	63.51	62.40	69.23	63.08	62.15	57.24
	Avg.	**64.11**	**58.80**	**58.62**	**64.10**	**64.02**	**50.29**	**49.70**

Takeaway: MM-ICL is broadly effective on OK-VQA, with MR producing substantial gains for several models, suggesting that demonstrations provide useful knowledge and answer-pattern cues.

Table 8: **General VQA Task: VQAv2.** Full results under different demonstration selection strategies and shot settings.

Strategy	Shots (k)	LLaVA-1.5-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	InternVL3-8B	Deepseek-VL	IDEFICS	GPT-5
RS	$k = 0$	74.18	80.60	78.64	79.67	77.88	40.15	63.50
	$k = 1$	72.62	80.90	78.50	81.20	77.77	60.88	74.36
	$k = 2$	72.60	80.34	78.23	81.76	77.98	62.62	77.06
	$k = 3$	72.65	80.35	78.40	81.70	78.07	62.89	80.02
	$k = 4$	71.51	80.36	78.34	81.97	78.11	62.86	78.48
	Avg.	**72.71**	**80.51**	**78.42**	**81.26**	**77.96**	**57.88**	**74.68**
VR	$k = 0$	74.18	80.60	78.64	79.67	77.88	40.15	63.50
	$k = 1$	72.35	80.26	78.36	81.17	76.91	60.69	71.96
	$k = 2$	71.91	79.50	78.15	81.13	77.27	62.56	73.80
	$k = 3$	71.69	79.49	78.34	80.96	77.55	62.99	73.84
	$k = 4$	71.46	79.60	78.37	80.96	77.81	63.25	76.34
	Avg.	**72.32**	**79.89**	**78.37**	**80.78**	**77.48**	**57.93**	**71.21**
TR	$k = 0$	74.18	80.60	78.64	79.67	77.88	40.15	63.50
	$k = 1$	57.81	79.27	77.55	80.41	63.63	59.45	73.08
	$k = 2$	56.68	78.79	77.14	80.41	67.06	60.76	74.80
	$k = 3$	56.23	78.98	77.25	80.09	68.38	61.33	76.20
	$k = 4$	55.53	79.02	77.36	80.37	68.73	61.55	77.98
	Avg.	**60.09**	**79.33**	**77.59**	**80.19**	**69.14**	**56.65**	**71.92**
SPR	$k = 0$	74.18	80.60	78.64	79.67	77.88	40.15	63.50
	$k = 1$	78.28	77.78	77.00	82.12	78.75	66.76	74.74
	$k = 2$	74.96	77.96	77.40	82.50	76.62	68.31	75.92
	$k = 3$	74.32	78.77	78.14	82.52	76.56	69.24	78.50
	$k = 4$	72.07	79.29	78.50	83.15	75.09	69.76	79.72
	Avg.	**74.76**	**78.88**	**77.94**	**81.99**	**76.98**	**62.84**	**73.61**
MR	$k = 0$	74.18	80.60	78.64	79.67	77.88	40.15	63.50
	$k = 1$	82.91	63.23	75.94	82.83	84.15	70.52	73.32
	$k = 2$	84.67	77.28	76.72	84.26	84.58	74.78	76.90
	$k = 3$	84.24	78.38	78.31	84.83	84.80	78.03	78.10
	$k = 4$	84.09	79.52	79.76	85.63	84.80	80.50	79.98
	Avg.	**81.87**	**75.80**	**77.87**	**83.44**	**83.24**	**68.80**	**73.78**

Takeaway: On VQAv2, demonstrations particularly help models with greater zero-shot headroom, whereas Qwen2.5-VL models receive relatively limited additional benefit.

Table 9: **General VQA Task: VizWiz.** Full results under different demonstration selection strategies and shot settings.

Strategy	Shots (k)	LLaVA-1.5-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	InternVL3-8B	Deepseek-VL	IDEFICS	GPT-5
RS	$k = 0$	59.56	62.54	63.18	49.23	49.19	34.94	53.36
	$k = 1$	55.16	65.98	67.43	65.74	45.79	34.42	59.22
	$k = 2$	55.42	66.44	67.49	66.62	47.34	37.98	60.78
	$k = 3$	55.85	67.14	65.48	67.12	47.37	40.12	62.36
	$k = 4$	55.46	66.50	60.95	67.11	46.65	41.14	62.22
	Avg.	**56.29**	**65.72**	**64.91**	**63.16**	**47.27**	**37.72**	**59.59**
VR	$k = 0$	59.56	62.54	63.18	49.23	49.19	34.94	53.36
	$k = 1$	60.34	66.97	68.68	66.61	57.86	51.11	58.60
	$k = 2$	60.41	66.56	66.91	68.82	59.12	54.06	59.16
	$k = 3$	60.98	67.52	65.29	68.92	58.06	55.34	61.90
	$k = 4$	60.14	67.25	59.42	68.92	56.73	56.57	61.58
	Avg.	**60.29**	**66.17**	**64.70**	**64.50**	**56.19**	**50.40**	**58.91**
TR	$k = 0$	59.56	62.54	63.18	49.23	49.19	34.94	53.36
	$k = 1$	52.04	66.42	68.28	66.27	49.69	43.98	58.06
	$k = 2$	52.04	66.94	66.81	67.22	50.56	47.93	61.22
	$k = 3$	52.89	67.43	64.26	68.04	52.51	51.06	61.90
	$k = 4$	54.01	67.33	60.81	68.57	50.04	52.20	61.98
	Avg.	**54.11**	**66.13**	**64.67**	**63.87**	**50.40**	**46.02**	**59.34**
SPR	$k = 0$	59.56	62.54	63.18	49.23	49.19	34.94	53.36
	$k = 1$	68.24	67.89	68.41	69.61	67.79	62.32	58.44
	$k = 2$	68.27	68.73	67.67	71.85	69.65	66.64	60.30
	$k = 3$	67.06	69.91	66.49	72.78	69.15	67.21	62.26
	$k = 4$	65.98	70.26	63.24	73.67	67.54	67.65	62.40
	Avg.	**65.82**	**67.87**	**65.80**	**67.43**	**64.66**	**59.75**	**59.28**
MR	$k = 0$	59.56	62.54	63.18	49.23	49.19	34.94	53.36
	$k = 1$	69.44	69.86	68.37	68.87	71.00	65.24	57.58
	$k = 2$	71.35	69.87	67.27	71.60	72.44	69.42	62.46
	$k = 3$	70.72	73.91	66.83	73.78	72.64	70.72	64.60
	$k = 4$	69.57	73.06	62.68	74.71	72.32	71.70	65.44
	Avg.	**68.13**	**69.85**	**65.67**	**67.64**	**67.52**	**62.40**	**60.62**

Takeaway: VizWiz presents consistently positive MM-ICL behavior, especially under MR and SPR, supporting the value of demonstrations for difficult or distribution-shifted visual inputs.

Table 10: **Structured Visual Task: ChartQA.** Full results under different demonstration selection strategies and shot settings.

Strategy	Shots (k)	LLaVA-1.5-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	InternVL3-8B	Deepseek-VL	IDEFICS	GPT-5
RS	$k = 0$	19.40	84.80	83.96	78.30	65.80	14.30	79.90
	$k = 1$	17.08	83.60	83.59	82.10	61.70	16.60	81.41
	$k = 2$	17.03	82.10	83.18	83.60	61.80	17.10	81.15
	$k = 3$	16.30	84.80	83.44	83.20	63.00	16.80	81.61
	$k = 4$	16.30	82.80	83.23	82.80	63.30	16.50	81.98
	Avg.	**17.22**	**83.62**	**83.48**	**82.00**	**63.12**	**16.26**	**81.21**
VR	$k = 0$	19.40	84.80	83.96	78.30	65.80	14.30	79.90
	$k = 1$	18.02	82.80	83.96	81.60	58.70	17.50	82.40
	$k = 2$	17.80	81.30	84.17	82.60	61.30	17.20	82.40
	$k = 3$	18.20	79.50	83.91	82.20	60.40	18.00	82.66
	$k = 4$	17.50	80.30	83.91	82.90	61.20	18.20	82.86
	Avg.	**18.18**	**81.74**	**83.98**	**81.52**	**61.48**	**17.04**	**82.04**
TR	$k = 0$	19.40	84.80	83.96	78.30	65.80	14.30	79.90
	$k = 1$	17.90	82.70	84.53	81.10	53.60	18.40	83.54
	$k = 2$	17.20	79.00	83.85	79.80	56.30	19.20	84.32
	$k = 3$	16.00	76.30	84.27	79.30	55.90	18.50	84.79
	$k = 4$	16.10	78.50	84.17	80.70	56.30	19.00	85.47
	Avg.	**17.32**	**80.26**	**84.16**	**79.84**	**57.58**	**17.88**	**83.60**
SPR	$k = 0$	19.40	84.80	83.96	78.30	65.80	14.30	79.90
	$k = 1$	21.30	82.70	83.96	81.30	55.90	19.50	83.96
	$k = 2$	21.10	78.40	83.49	80.60	57.80	20.10	85.26
	$k = 3$	20.30	76.10	84.22	79.30	57.10	19.90	85.57
	$k = 4$	19.50	78.60	84.27	80.00	57.80	20.00	86.04
	Avg.	**20.32**	**80.12**	**83.98**	**79.90**	**58.88**	**18.76**	**84.15**
MR	$k = 0$	19.40	84.80	83.96	78.30	65.80	14.30	79.90
	$k = 1$	28.40	82.20	83.80	83.00	61.60	20.80	84.84
	$k = 2$	29.30	79.50	83.44	81.40	61.90	23.60	86.09
	$k = 3$	27.00	76.20	83.75	81.20	63.30	25.30	86.98
	$k = 4$	25.30	79.50	83.80	80.50	62.60	27.40	87.24
	Avg.	**25.88**	**80.44**	**83.75**	**80.88**	**63.04**	**22.28**	**85.01**

Takeaway: ChartQA shows mixed effects: demonstrations improve LLaVA-1.5-7B, IDEFICS, and GPT-5, but do not improve Qwen2.5-VL-7B over its strong zero-shot baseline.

Table 11: **Visual Grounding Task: RefCOCOg.** Full results under different demonstration selection strategies and shot settings.

Strategy	Shots (k)	LLaVA-1.5-7B	Qwen2.5-VL-7B	Qwen2.5-VL-32B	InternVL3-8B	Deepseek-VL	IDEFICS	GPT-5
RS	$k = 0$	34.90	86.10	79.22	61.60	47.50	2.70	5.60
	$k = 1$	19.00	16.70	82.27	74.50	30.70	3.60	2.80
	$k = 2$	17.80	38.50	84.52	75.70	27.50	2.90	4.00
	$k = 3$	15.20	55.00	86.00	74.90	26.00	3.40	5.00
	$k = 4$	17.70	54.10	86.10	75.80	27.30	3.50	3.20
	Avg.	**20.92**	**50.08**	**83.62**	**72.50**	**31.80**	**3.22**	**4.12**
VR	$k = 0$	34.90	86.10	79.22	61.60	47.50	2.70	5.60
	$k = 1$	26.40	29.10	79.63	73.90	35.40	11.30	4.80
	$k = 2$	20.00	39.00	81.95	76.30	30.10	12.70	6.00
	$k = 3$	17.00	50.50	83.31	76.90	27.80	13.00	7.80
	$k = 4$	16.60	49.40	83.55	74.60	25.80	13.70	7.40
	Avg.	**22.98**	**50.82**	**81.53**	**72.66**	**33.32**	**10.68**	**6.48**
TR	$k = 0$	34.90	86.10	79.22	61.60	47.50	2.70	5.60
	$k = 1$	21.00	14.40	81.02	73.60	33.10	3.20	2.40
	$k = 2$	18.30	57.90	84.24	74.70	23.20	3.00	3.00
	$k = 3$	16.70	61.60	85.66	76.40	23.80	3.00	2.60
	$k = 4$	15.30	45.40	86.10	75.30	29.10	3.00	4.60
	Avg.	**21.24**	**53.08**	**83.25**	**72.32**	**31.34**	**2.98**	**3.56**
SPR	$k = 0$	34.90	86.10	79.22	61.60	47.50	2.70	5.60
	$k = 1$	21.00	23.80	78.22	72.40	29.10	12.60	5.40
	$k = 2$	16.70	39.00	81.23	76.20	24.60	12.40	7.60
	$k = 3$	14.50	49.40	83.13	75.00	23.60	12.80	9.00
	$k = 4$	14.50	49.80	83.31	74.80	24.70	13.10	9.40
	Avg.	**20.32**	**49.62**	**81.02**	**72.00**	**29.90**	**10.72**	**7.12**
MR	$k = 0$	34.90	86.10	79.22	61.60	47.50	2.70	5.60
	$k = 1$	19.90	17.90	78.80	73.60	25.90	5.30	2.80
	$k = 2$	15.70	36.30	83.34	75.20	22.50	4.60	3.80
	$k = 3$	14.50	50.00	84.21	74.50	18.70	6.30	4.00
	$k = 4$	14.60	48.70	84.70	75.80	25.20	6.10	4.80
	Avg.	**19.92**	**47.80**	**82.05**	**72.14**	**27.96**	**5.00**	**4.32**

Takeaway: RefCOCOg is highly unstable under demonstrations: InternVL3-8B and Qwen2.5-VL-32B improve, whereas several other models degrade substantially, indicating sensitivity to grounding-format interference.

Table 12: **Video Understanding Task: MVBench.** Full results under different demonstration selection strategies and shot settings.

Strategy	Shots (k)	Qwen2.5-VL-7B	Qwen2.5-VL-32B	InternVL3-8B	GPT-5
RS	$k = 0$	60.00	55.60	67.00	59.60
	$k = 1$	61.40	57.20	64.60	60.00
	$k = 2$	58.60	56.60	61.00	60.00
	Avg.	**60.00**	**56.47**	**64.20**	**59.87**
VR	$k = 0$	60.00	55.60	67.00	59.60
	$k = 1$	63.00	60.20	66.20	64.40
	$k = 2$	57.20	59.30	61.40	60.60
	Avg.	**60.13**	**59.03**	**64.93**	**61.67**
TR	$k = 0$	60.00	55.60	67.00	59.60
	$k = 1$	59.20	53.60	58.20	63.20
	$k = 2$	56.20	54.80	57.80	62.80
	Avg.	**58.60**	**55.40**	**61.73**	**61.80**
SPR	$k = 0$	60.00	55.60	67.00	59.60
	$k = 1$	63.60	60.20	75.60	–
	$k = 2$	65.60	63.30	69.60	61.80
	Avg.	**63.47**	**60.17**	**71.27**	**61.70**
MR	$k = 0$	60.00	55.60	67.00	59.60
	$k = 1$	63.20	61.60	68.00	60.60
	$k = 2$	60.80	61.40	66.00	63.40
	Avg.	**61.67**	**60.00**	**66.80**	**61.40**

Takeaway: On MVBench, a small number of appropriately selected examples is sufficient to yield improvements; SPR is particularly effective for the open-weight video-capable models.