Why Multimodal In-Context Learning Lags Behind? Unveiling the Inner Mechanisms and Bottlenecks

Yu Wang, Sharon Li
Department of Computer Sciences, University of Wisconsin-Madison,
{yuwang, sharonli}@cs.wisc.edu

Abstract

In-context learning (ICL) enables models to adapt to new tasks via inference-time demonstrations. Despite its success in large language models, the extension of ICL to multimodal settings remains poorly understood in terms of its internal mechanisms and how it differs from text-only ICL. In this work, we conduct a systematic analysis of ICL in multimodal large language models. Using identical task formulations across modalities, we show that multimodal ICL performs comparably to text-only ICL in zero-shot settings but degrades significantly under few-shot demonstrations. To understand this gap, we decompose multimodal ICL into task mapping construction and task mapping transfer, and analyze how models establish cross-modal task mappings, and transfer them to query samples across layers. Our analysis reveals that current models lack reasoning-level alignment between visual and textual representations, and fail to reliably transfer learned task mappings to queries. Guided by these findings, we further propose a simple inference-stage enhancement method that reinforces task mapping transfer. Our results provide new insights into the mechanisms and limitations of multimodal ICL and suggest directions for more effective multimodal adaptation. Our code is available here.

1 Introduction

In-context learning (ICL) enables models to adapt to new tasks at inference time by using a small number of demonstrations, without updating parameters (Zhang et al., 2023b; Min et al., 2022; Zhou et al., 2024; Cho et al., 2025). Originally studied in text-based LLMs, this paradigm has now been extended to Multimodal LLMs (MLLMs) (Li, 2025; Li et al., 2025b; Huang et al., 2024), allowing models to incorporate both visual and textual information in demonstrations. Multimodal ICL has shown promise on various vision-language tasks, such as visual question answering (Li et al., 2025b; Huang et al., 2024), offering more flexible and generalizable multimodal understanding.

Recent studies have investigated multimodal ICL from different perspectives (Qin et al., 2024a; Li et al., 2025b; Gao et al., 2024; Chen et al., 2025b), such as the underutilization of visual context (Li et al., 2025c; Chen et al., 2025a) and sensitivity to demonstration configurations (Qin et al., 2024a; Chen et al., 2024a). In parallel, recent works have introduced token pruning to improve efficiency (Li et al., 2025a; Gao et al., 2024). Despite these advances, the underlying mechanisms of ICL within the multimodal domain remain poorly explored. It remains an open question *how* MLLMs perform ICL and *whether* this process differs fundamentally from textual ICL. Motivated by these gaps, this paper proposes a systematic analysis framework for multimodal ICL.

First, to investigate the differences between multimodal and textual ICL, we construct an identical ICL dataset (see Fig. 1(a–b); details in Sec.3). For the same ICL problem, we provide both textual and multimodal formulations (Nikankin et al., 2025), enabling a controlled comparison of model performance under the two settings (More details see Sec.3). Evaluation on Qwen2.5-VL (Bai et al., 2025) and Gemma-3 (Team et al., 2025) (Fig. 1(c)) reveals a striking contrast: while zero-shot performance is comparable across modalities, few-shot performance significantly degrades with multimodal demonstrations. This gap indicates that MLLMs struggle to leverage multimodal context as effectively as text.

To delve into why multimodal ICL diverges from its textual counterpart and identify the bottlenecks constraining its performance, we take a closer look at the multimodal ICL pipeline. Specifically, we decompose multimodal ICL into two steps (Li et al., 2025c; Cho et al., 2025), as illustrated in Fig. 1(d): *Task Mapping Construction*, where the

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13670–13685
July 2-7, 2026

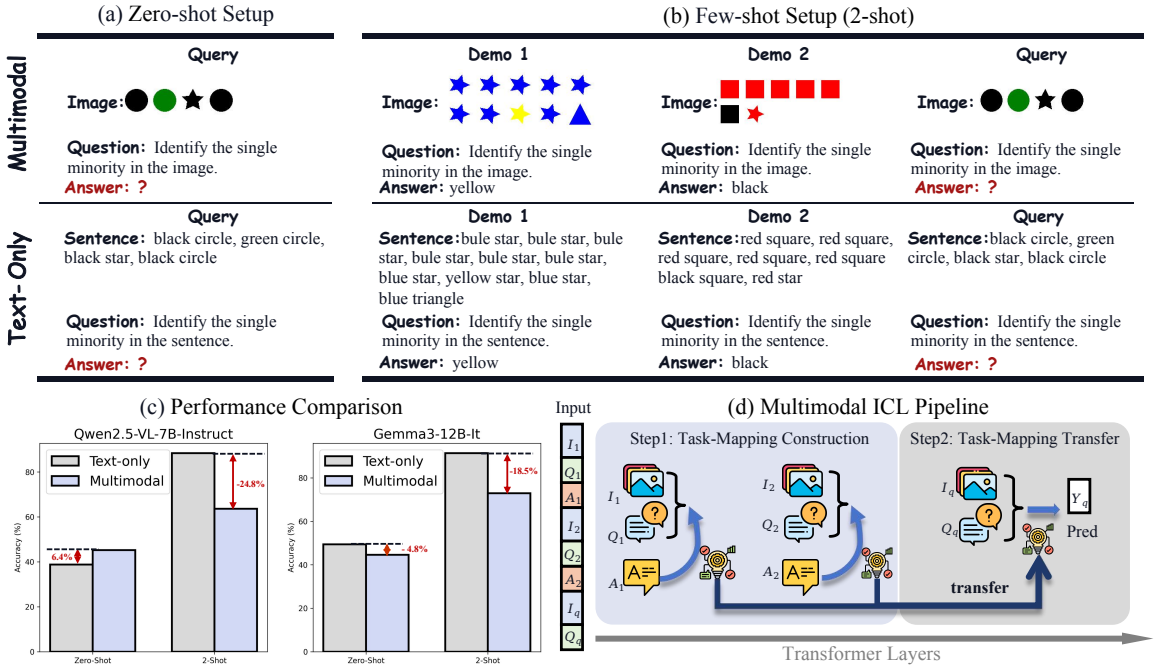


Figure 1: (**a–b**) Illustration of constructed outlier detection task (Chen et al., 2025a), where (a) shows the zero-shot setup and (b) shows the few-shot setup. In the few-shot scenario (2-shot), the model must infer from the demonstrations whether the query should be solved based on shape or color, and then apply this inferred rule to identify the outlier item in either the image or the sentence. (**c**) Performance comparison of Qwen2.5-VL-7B and Gemma-3-12B, showing that multimodal ICL suffers a clear degradation under the few-shot setting (4-shot) relative to text-only ICL. (**d**) Illustration of the multimodal ICL pipeline, including task mapping construction from demonstrations and task mapping transfer to the query.

model learns and builds the task mapping within the demonstrations, and *Task Mapping Transfer*, where the learned mapping is applied to the query for final prediction. Here, task mapping refers to inference mechanism that transforms questions into their corresponding answers within the MLLMs latent space.

Our analysis yields two pivotal findings. First, we find that MLLMs are capable of constructing task mappings from multimodal demonstrations: in middle layers, demonstration labels are grounded to the correct visual evidence, indicating successful task mapping construction. Notably, this grounding emerges even in cases where final predictions on the query are incorrect, suggesting that task mapping construction alone is insufficient for successful multimodal ICL. Second, we show that these constructed task mappings at mid-layer are not reliably transferred to guide query-time reasoning, due to a *critical cross-modal misalignment between visual perception and textual reasoning during ICL*. Specifically, attention from the query's last token to the demonstration labels and their grounded visual evidence increases sharply in later layers. However, at these later layers, the task mappings induced by the demonstrations are no longer accurate or reliably represented, preventing the model from applying the inferred task structure during reasoning. Consequently, late-layer predictions are driven primarily by perceptual signals in the query image rather than by the task mappings inferred from demonstrations. In other words, *task mapping construction and task mapping transfer are decoupled across different layers.*

Finally, to access the practical implications of this analysis, we introduce a simple yet effective inference-stage ICL enhancement method. Our approach explicitly extracts the task mapping encoded within the demonstrations and reinforces its transfer to the query by dynamically amplifying attention toward task-relevant evidence while suppressing irrelevant regions. This intervention serves as an exploratory probe that validates the causal role of task mapping transfer in multimodal ICL. In summary, our main contributions are:

1. We present a controlled experimental framework that enables a direct comparison between text-only and multimodal ICL under identical task formulations, revealing a substantial perfor-

mance gap in few-shot settings.

2. We conduct a systematic mechanistic analysis of multimodal ICL that decomposes the process into task mapping construction and task mapping transfer, uncovering the mechanisms of multimodal ICL and identifying the bottlenecks that constrain multimodal ICL performance.

3. We explore a lightweight inference-time intervention that reinforces task mapping transfer, providing empirical evidence for the causal relevance of the identified bottleneck and demonstrating the practical utility of our analysis.

2 Background and Related Work

Multimodal In-Context Learning. In multimodal ICL, a model $\mathcal{M}$ receives an input sequence consisting of n multimodal demonstrations and a query, and predicts the target label $\hat{Y}_q$, as follows:

$$\hat{Y}_q = \mathcal{M}\big((I_1, Q_1, Y_1), \ldots, (I_n, Q_n, Y_n), (I_q, Q_q)\big),$$

Here, each demonstration is a triplet (I_i, Q_i, Y_i), where I_i is the image, Q_i is the question, and Y_i is the ground-truth label. The query instance consists of an image-question pair (I_q, Q_q).

Related Work on ICL. While the mechanisms of text-only ICL have been extensively studied (Cho et al., 2025; Xie et al., 2021; Dai et al., 2023; Wang et al., 2023; Han et al., 2023a; Jeon et al., 2024; Zheng et al., 2024) through pre-training data (Singh et al., 2024b,a; Han et al., 2023b), feature attribution of inputs (Pan, 2023; Kossen et al., 2024), and functional reduction (Zhang et al., 2023a; Han et al., 2023a), understanding its multimodal counterpart remains an open challenge. For instance, Cho et al. (Cho et al., 2025) decompose text-only ICL inference into three primary operations: input text encoding, semantic merging, and feature retrieval and copying.

Recently, research on multimodal ICL has witnessed significant advancements. One line of work primarily investigates demonstration configurations to enhance ICL performance (Chen et al., 2023; Li et al., 2023c; Yang et al., 2024; Baldassini et al., 2024; Qin et al., 2024b; Xu et al., 2024; Wu et al., 2025; Luo et al., 2024; Li et al., 2025c; Qin et al., 2024a; Chen et al., 2024a; Huang et al., 2025), and establishes comprehensive benchmarks (Zong et al., 2024; Chen et al., 2025a). Another strand focuses on fine-tuning the MLLMs for better multimodal ICL ability (Zhao et al., 2023; Doveh et al., 2024; Li et al., 2023b,a; Gao et al., 2025; Li et al., 2023b; Jia et al., 2024; Chen et al., 2025a), while a separate line of work investigates inference-stage methods that improve multimodal ICL without additional training (Li et al., 2025b). In parallel, some studies focus on improving efficiency by addressing the issue of token redundancy in ICL (Li et al., 2025a; Gao et al., 2024). However, the internal latent mechanisms governing multimodal ICL–and how they fundamentally differ from text-only ICL–remain largely unexplored. Our work fills this gap by conducting the first in-depth analysis to uncover the structural bottlenecks in cross-modal task mapping and transfer.

3 Multimodal ICL is More Challenging

This section provides a systematic comparison between multimodal and text-only ICL under controlled conditions. Inspired by Chen et al. (2025a); Nikankin et al. (2025), we design a controlled outlier detection task. The objective is to identify a single minority instance that deviates from the majority based on a specific feature (shape or color). The label space consists of 4 shape categories (circle, triangle, square, star) and 10 color categories (yellow, blue, green, red, black, orange, purple, pink, brown, gray). To isolate the effect of modality, we instantiate each problem in this task in two distinct formats: text-only and multimodal. As illustrated in Fig. 1, the text-only format presents all task-relevant information in natural language, whereas the multimodal format requires the model to integrate multimodal information, particularly visual evidence, for correct task completion. This design ensures that only the input modality varies while keeping task structure and supervision fixed.

Performance Gap. We evaluate representative multimodal large language models under zero-shot and few-shot settings to show that multimodal ICL is fundamentally more difficult than textual ICL. The results in Tab. 1 reveal a clear performance gap under few-shot conditions. For example, Qwen2.5-VL-7B model exhibits an accuracy drop of over **24**% when moving from text-only to multi-modal demonstrations. *This degradation indicates that MLLMs struggle to effectively leverage multimodal demonstrations for task induction and transfer.*

Case Study. We further analyze the error cases of text-only and multimodal ICL. Specifically, we first check whether the model correctly identifies

Model	Size	ICL Type	Zero-Shot	4-shot
Qwen2.5-VL	7B	Text Only	38.80	**88.40**
		Multimodal	**45.20** (↑ 6.40)	63.60 (↓ 24.80)
	32B	Text Only	**50.00**	**80.47**
		Multimodal	43.60 (↓ 6.40)	76.13 (↓ 4.34)
Gemma-3	12B	Text Only	**49.40**	**91.40**
		Multimodal	44.60 (↓ 4.80)	72.93 (↓ 18.47)
	27B	Text Only	**50.00**	**89.67**
		Multimodal	48.00 (↓ 2.00)	78.80 (↓ 10.87)

Table 1: Comparison between text-only and multimodal ICL across model families and sizes.

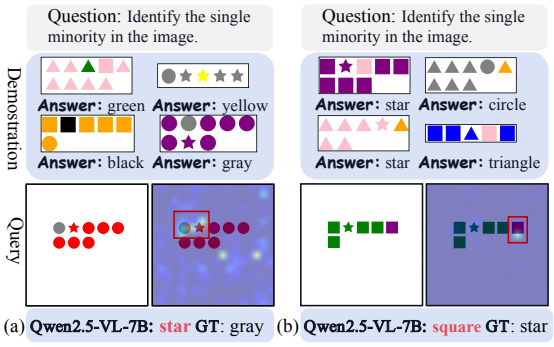


Figure 2: **Qualitative examples of error cases**. (a) False task recognition: the model misreads the task mapping in the demonstrations and outputs the wrong attribute (shape star) while the true minority is color gray. (b) Correct task recognition but false answer: although the model identifies the correct task mapping (detecting the OOD sample by shape feature), it still predicts the wrong minority because it attends to incorrect regions in the image (square instead of star).

the task type, i.e., whether it performs outlier detection based on the correct feature (color or shape). This is determined by string-matching the model output to see whether the prediction refers to color or shape. We then examine whether the prediction matches the ground-truth label. Based on this analysis, we categorize the errors into two types: (Case 1) *False Task Recognition* and (Case 2) *Correct Task Recognition but Incorrect Answer*, where Fig. 2 provides representative examples of each error type. Case 1 occurs when the model fails to recognize the intended task, stemming from either a failure to infer the correct mapping from demonstrations or an inability to apply it to the query (e.g., detecting the outlier by shape when the demonstrations imply color). Case 2 occurs when the model correctly infers the task rule but fails to produce a correct final prediction (e.g., predicting `square` instead of `star`).

Understanding the Gap. We also report the proportions of each error type in Fig. 3. As shown, multimodal ICL exhibits significantly higher rates

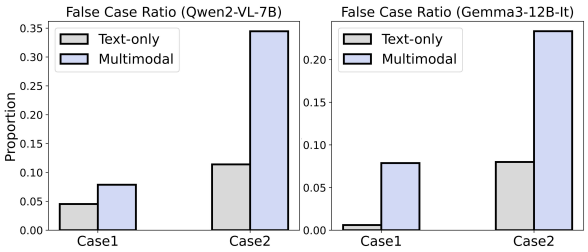


Figure 3: Proportion of error types in text-only ICL and multimodal ICL. Case 1 corresponds to incorrect task recognition, and Case 2 corresponds to correct task recognition but an incorrect answer.

for both error types compared to the text-only baseline. Specifically, these errors suggest distinct internal failure modes: (i) failures to construct a stable task mapping from multimodal demonstrations, and (ii) failures to accurately transfer the recognized task to the query. Understanding these errors is important because effective multimodal ICL requires both capabilities–learning the task structure from demonstrations and reusing this structure during query inference by integrating visual and textual cues into a unified reasoning process.

However, existing analyses based solely on input-output behavior do not reveal how task mappings are formed *internally* or *why* they fail to generalize from demonstrations to the query (Li et al., 2025c; Qin et al., 2024a; Chen et al., 2024a). To move beyond performance-level observations, we next examine the internal mechanisms underlying multimodal ICL and identify the structural bottlenecks that hinder performance.

4 Understanding Multimodal ICL Mechanisms

To understand the process of multimodal ICL, we conduct a systematic analysis of its internal behaviors. Our goal is to uncover the mechanisms of multimodal ICL and identify the bottlenecks that constrain multimodal ICL performance. Guided by the errors identified in Sec. 3, we organize this section around two core questions: **RQ1** evaluates whether MLLMs can establish effective task mappings from demonstrations, and **RQ2** studies whether MLLMs can transfer the learned task mappings from demonstrations to the query.

4.1 RQ1: Can multimodal ICL Establish Effective Task Mappings?

First, we examine whether the failures occur during demonstration-level task mapping construction.

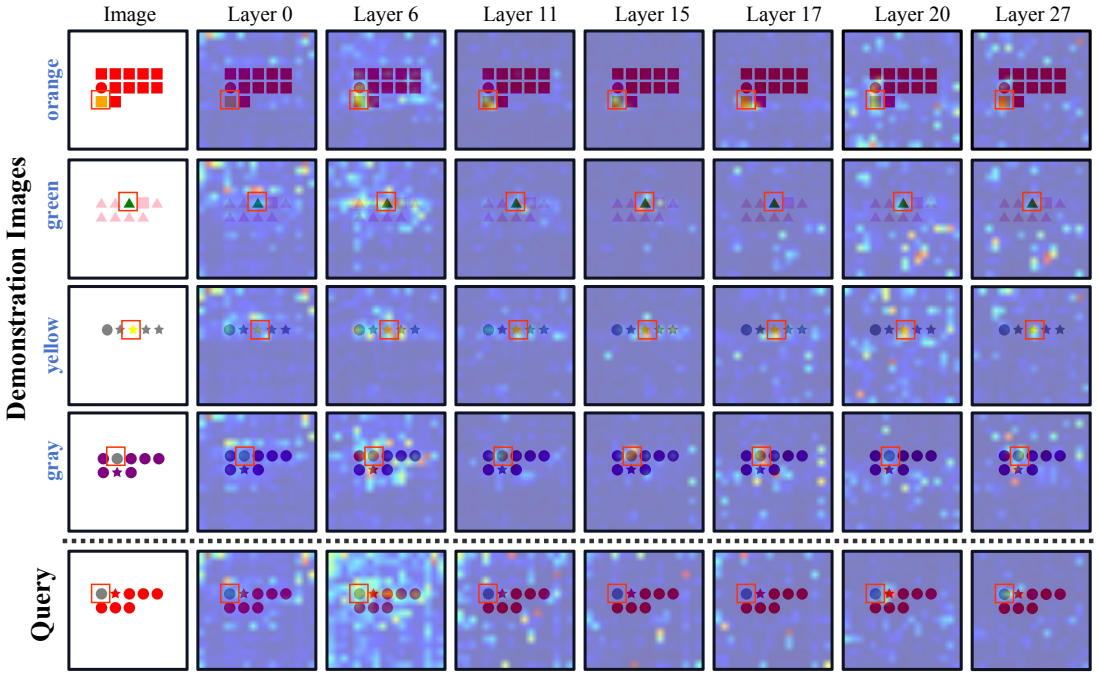


Figure 4: **Layer-wise visualization of attention from demonstration's labels tokens (or the last token) to image tokens in a multimodal ICL example.** The four demonstrations are labeled by the color outlier, but the model exhibits the *False Task Recognition* on the query and incorrectly predicts `star` instead of `gray`. Red bounding boxes denote ground-truth evidence regions. Demonstration label tokens form clear object-level grounding only in mid layers, with early layers being diffuse and deeper layers becoming noisy again. In contrast, the query token shows no meaningful grounding until the final several layers, where localization finally appears.

Mid-layer Grounding as Evidence of Task Mapping Construction. To investigate how task mappings are formed, we begin with a qualitative case study that visualizes attention from demonstration label tokens to image regions across layers (Fig. 4). In this example, constructing a task mapping requires grounding each demonstration label to the color attribute of the outlier object. When the attention from a demonstration's label aligns with the ground truth evidence region in its corresponding image (red bounding boxes), this indicates that the model correctly grounds the label and successfully builds the underlying task mapping.

As shown in Fig. 4, despite the incorrect final prediction, for *demonstrations*, early layers exhibit scattered and unstructured patterns, but the mid layers develop a clear focus on the correct object, indicating the emergence of visual grounding. In deeper layers, this focus becomes less stable and the attention spreads again.

Quantifying Demonstration-Level Visual Grounding. To move beyond individual examples, we quantitatively analyze attention allocation from demonstration label tokens to different image regions across layers, and present the results in Fig. 5. We divide each image into three evidence types: the correct outlier region as *correct evidence*, the incorrect outlier region as *false evidence*, and all remaining areas as *irrelevant evidence*. For each layer, we compute the relative attention assigned to these regions by the demonstration's label tokens.

Across models and across both correct and incorrect samples, we observe a pronounced peak of attention to correct evidence regions in intermediate layers, while attention to false and irrelevant regions remains consistently low. This indicates that the model effectively achieves label-evidence alignment within the demonstrations and are able to construct task mappings in intermediate layers, regardless of whether the final prediction on the query is correct.

Causal Intervention Study. While attention alone does not constitute a complete explanation of model behavior, we complement attention analysis with targeted interventions to establish a causal link. In particular, we conduct an intervention that disrupts visual grounding by replacing attention over image tokens with a uniform distribution. We then compare the model's multimodal ICL perfor-

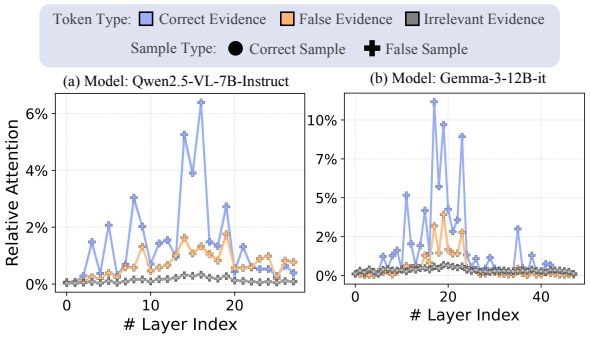


Figure 5: **Layer-wise attention ratios over different image regions for correct (•) vs. incorrect (+) predictions.** Relative attention from demonstration label tokens to correct, false, and irrelevant evidence regions. Both models show a strong peak on correct evidence at the midlayer, indicating stable visual grounding within the demonstrations.

mance before and after the intervention on different model families (Tab. 2). The results show uniform attention causes a precipitous performance collapse across all model families. Notably, the accuracy drops to levels near or below the zero-shot baseline, demonstrating that the precise attention to visual evidence in intermediate layers is a causal prerequisite for the performance gains observed in multimodal ICL.

Model	Qwen2.5-VL		Gemma-3	
Size	7B	32B	12B	27B
Zero-shot	45.20	43.60	44.60	48.00
Original (4-shot)	69.09	78.45	72.93	78.62
After UAS (4-shot)	45.44	52.19	50.56	0.0

Table 2: **Causal Intervention Study.** We apply our *Uniform Attention Suppression* (UAS) intervention, which replaces the model's attention over image tokens with a uniform distribution, leading to significant performance degradation.

Task Mapping Construction Is Necessary but Not Sufficient. Within the scope of our controlled setup, these findings show that MLLMs are capable of constructing task mappings from multimodal demonstrations by grounding labels to the correct visual evidence in intermediate layers. However, since this grounding occurs even when final predictions fail, demonstration-level task mapping construction alone cannot explain the error patterns observed in Sec. 3. This suggests that the failure of multimodal ICL arises not solely from constructing task mappings, but from transferring or applying these mappings when reasoning over the query—an issue we investigate next.

Finding 1

During multimodal ICL, models can form task mappings within the demonstrations by grounding demonstration labels more to the correct visual evidence in intermediate layers.
This grounding emerges regardless of whether the final prediction on the query is correct or not.

4.2 RQ2: Can Task Mappings Be Transferred from Demonstrations to the Query?

Successful multimodal ICL requires not only inferring the task structure from demonstrations, but also reusing this structure when reasoning over the query. In this section, we analyze how demonstration-induced task information influences the model's final decision, and identify the mechanisms that lead to failures in task mapping transfer.

Failure to Access Task Mappings in Intermediate Layers. To examine whether task mappings are transferred to the query, we measure the attention from the query's last token to both the demonstration labels and the label-relevant image regions across layers. In text-only ICL, such attention patterns are known to reflect how LLMs incorporate demonstration-induced information when answering the query (Cho et al., 2025; Xie et al., 2021; Dai et al., 2023; Wang et al., 2023; Han et al., 2023a; Jeon et al., 2024; Zheng et al., 2024). If multimodal task mappings are successfully transferred, we would expect the query's final token to attend to the demonstration labels and their grounded visual regions, particularly in layers where task mappings are constructed.

As shown in Fig. 6, we find that attention from the last token to demonstration labels and their grounded visual evidence is nearly zero in the middle layers, despite the presence of demonstration-level grounding at these layers (recall Sec. 4.1). Meanwhile, it subsequently rises in later layers to drive the final prediction. This implies that task information is encoded early but not immediately utilized. In other words, *task mapping construction and task application are decoupled across different layers of the model.*

Perception-Reasoning Misalignment as the Core Bottleneck. Interestingly, we also observe in Fig. 6 that attention from query's last token to the

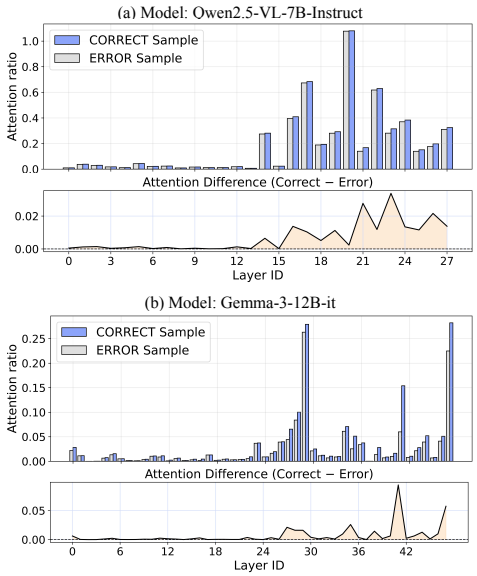


Figure 6: Comparison of Correct and Error Samples in Layer-wise Last-Token Attention to Evidence Regions.

demonstration labels and their grounded visual evidence increases sharply in later layers. However, at these later layers, the task mappings induced by the demonstrations are no longer accurate or reliably represented. As a result, the final prediction is driven primarily by late-layer perceptual cues rather than by demonstration-induced task mappings, causing multimodal demonstrations to provide little meaningful guidance for the decision. Specifically, we observe that correct samples maintain significantly higher attention to evidence regions in late layers compared to error samples (see the difference plots in Fig. 6), implying that strong late-layer perceptual cues are a prerequisite for accurate predictions. This finding also clarifies the error patterns in Sec. 3: when task mappings fail to influence late-layer reasoning, the model may either behave as if the task was not inferred or attend to correct visual evidence without applying the correct decision rule.

Finding 2

Task mappings inferred from multimodal demonstrations fail to reliably influence query-time reasoning. This is primarily due to the *cross-modal misalignment between visual perception and textual reasoning*: task mappings involving visual contexts are constructed in mid layers, but fail to propagate across the modality gap to the final reasoning process. In other words, *task mapping construction and task application are decoupled across different layers and the cross-modal barrier.*

5 Mapping-Guided Inference: Bridging Task Mapping and Transfer

The analyses in Section 4 reveal a clear mechanistic bottleneck in multimodal in-context learning. In this section, we ask a pragmatic follow-up question: can insights from this analysis be leveraged to improve multimodal ICL behavior in practice?

To this end, we introduce Mapping-Guided Inference (MGI), an inference-time intervention that extracts the latent task mapping encoded within the demonstrations in mid layers and explicitly injects it into the query's attention mechanism. By guiding the query token to attend to evidence regions grounded by demonstration labels in later layers, MGI is designed to bridge the misalignment between perception and reasoning in multimodal ICL. We view this approach as an exploratory, analysis-driven intervention that probes the practical utility of the mechanisms identified above, rather than as a fully optimized method.

5.1 Estimating Task Mapping from Demonstrations

As shown in Section 4.1, task mappings in multimodal ICL are constructed within demonstrations via mid-layer visual grounding. In MGI, we explicitly extract this latent task mapping from the model's internal attention patterns.

Constructing the Attention Candidate Set. We first collect the attention weights from each demonstration's label token to its corresponding all image tokens. For n-shot setup at layer ℓ, we define the attention set $\mathcal{A}^{\ell} = \{\mathbf{a}^{(\ell,h)}_{\text{lbl}_i \to \text{img}_i}\}_{i,h}$, where $\mathbf{a}^{(\ell,h)}_{\text{lbl}_i \to \text{img}_i} \in \mathbb{R}^{L_{\text{img}_i}}$ denotes the attention distribution from the label token of demonstration i to its image tokens at layer ℓ and head h. Here, $\ell \in \{1, \ldots, L\}$ and $h \in \{1, \ldots, H\}$ index the Transformer layers and attention heads, respectively, while L_{img_i} denotes the number of image tokens in the i-th demonstration.

Identifying the Peak Grounding Layer. We aim to identify the layer ℓ^* that exhibits the strongest visual grounding between demonstration labels and image regions. We quantify grounding strength using the entropy of the attention distribution, where lower entropy indicates a more focused attention over specific visual evidence.

For each layer ℓ, we compute the average entropy of the attention from demonstration label tokens to their corresponding image tokens, aggregated

across all demonstrations and attention heads. The peak grounding layer is defined as:

$$\ell^* = \arg\min_{\ell} \sum_{i=1}^{n} \sum_{h=1}^{H} \mathcal{H}\left(\mathbf{a}^{(\ell,h)}_{\text{lbl}_i \to \text{img}_i}\right), \quad (1)$$

where $\mathcal{H}(\mathbf{p}) = -\sum_j p_j \log p_j$ is the entropy of a probability distribution $\mathbf{p}$. Note that the attention weights $\mathbf{a}$ are normalized by their sum (i.e., $\mathbf{p} = \mathbf{a}/\|\mathbf{a}\|_1$) to form a valid probability distribution over the image tokens. The attention patterns at this peak layer ℓ^* are then used as our estimate of the task mapping:

$$\hat{\mathcal{M}} = \mathcal{A}^{(\ell^*)}. \quad (2)$$

5.2 Intervening on Query Attention

After obtaining the estimated task mapping $\hat{\mathcal{M}}$, we guide the model to answer the query by reusing this mapping. Specifically, when the query's label token attends to the image tokens of demonstration i, we encourage it to mimic the attention pattern established by the label of demonstration i at the peak layer.

Injecting the Task Mapping into Query Attention. We intervene on the attention mechanism in layers deeper than a starting layer L_{start}. For a given layer $\ell > L_{\text{start}}$ and head h, we modify the attention from the query token (q) to the image tokens of demonstration i as follows:

$$\tilde{\mathbf{a}}^{(\ell,h)}_{\text{q}\to\text{img}_i} = \mathbf{a}^{(\ell,h)}_{\text{q}\to\text{img}_i} + \lambda \cdot \mathbf{a}^{(\ell^*,h)}_{\text{lbl}_i\to\text{img}_i}, \quad (3)$$

where $\mathbf{a}^{(\ell,h)}_{\text{q}\to\text{img}_i}$ is the original attention from the query to the image tokens of demonstration i, $\mathbf{a}^{(\ell^*,h)}_{\text{lbl}_i\to\text{img}_i} \in \hat{\mathcal{M}}$ denotes the label-to-image attention extracted at the peak grounding layer l^*, and $\lambda > 0$ controls the strength of the intervention.

After the intervention, we re-normalize the attention scores $\hat{\mathbf{a}}$ to ensure they form a valid probability distribution: $\tilde{\mathbf{a}} = \hat{\mathbf{a}}/\sum \hat{\mathbf{a}}$. This process explicitly encourages the query token to "look at" the demonstration images in the same way that defined the task in the demonstrations, thereby promoting consistent visual grounding between demonstrations and query inference.

5.3 Experiments

Dataset and Models. We evaluate MGI on the TrueMICL benchmark (Chen et al., 2025a) (Outlier Detection, Clock Math, Operator Induction) and a natural-image datasets (OK-VQA (Marino et al., 2019)). For evaluation, we use exact matching for numerical answers and keyword matching for text. Experiments are conducted on Qwen2.5-VL (Bai et al., 2025) and Gemma3 (Team et al., 2025) across different model scales to ensure robustness. More details refer to Appendix A.

	Size	Method	Outlier	Clock	Operator	OK-VQA
Qwen2.5-VL	7B	Vanilla	69.09$_{\pm 0.81}$	63.45$_{\pm 0.00}$	77.59$_{\pm 0.00}$	48.13$_{\pm 0.11}$
		Ours	**70.17**$_{\pm 0.00}$	**64.49**$_{\pm 0.00}$	**77.93**$_{\pm 0.00}$	**48.17**$_{\pm 0.27}$
	32B	Vanilla	**78.45**$_{\pm 0.00}$	74.14$_{\pm 0.00}$	97.93$_{\pm 0.00}$	52.67$_{\pm 0.64}$
		Ours	**78.45**$_{\pm 0.00}$	**74.48**$_{\pm 0.00}$	**98.28**$_{\pm 0.00}$	**53.13**$_{\pm 0.10}$
Gemma3	12B	Vanilla	72.93$_{\pm 0.00}$	**66.90**$_{\pm 0.00}$	47.93$_{\pm 0.00}$	48.39$_{\pm 0.75}$
		Ours	**73.85**$_{\pm 0.14}$	**66.90**$_{\pm 0.00}$	**48.28**$_{\pm 0.00}$	**49.51**$_{\pm 0.26}$
	27B	Vanilla	78.62$_{\pm 0.00}$	76.55$_{\pm 0.00}$	70.00$_{\pm 0.00}$	48.37$_{\pm 0.63}$
		Ours	**79.54**$_{\pm 0.16}$	**76.90**$_{\pm 0.00}$	**70.34**$_{\pm 0.00}$	**49.29**$_{\pm 0.17}$

Table 3: Results of 4-shot multimodal ICL. Experiments are averaged over 3 different seeds, and the performance is in percentage with the standard deviations. Rows shaded in gray correspond to MGI (ours).

Main Results. As shown in Tab. 3, MGI consistently enhances multimodal in-context learning performance across various architectures and datasets. The method is particularly effective on tasks requiring strict visual reasoning, contributing to gains in OK-VQA and Outlier. We present detailed hyperparameter analysis in Appendix C.3. Overall, these findings highlight that bridging the gap between mid-layer mapping construction and query inference serves as a promising enhancement.

6 Conclusion

In this work, we presented a systematic mechanistic analysis of multimodal ICL. By decomposing multimodal ICL into task mapping construction and task mapping transfer, we identified a core bottleneck: while current multimodal models can construct task mappings by grounding labels to relevant visual evidence, these mappings are not reliably transferred to guide query-time reasoning, due to a misalignment between perception and reasoning. Meanwhile, we introduced a simple inference-time intervention that reinforces task mapping transfer. This lightweight intervention provides empirical evidence for the causal role of task mapping transfer and yields consistent improvements in multimodal ICL performance. More broadly, our findings suggest that effective multimodal ICL requires mechanisms that preserve task structure across layers, and point to future directions in model design and training aimed at better aligning perception and reasoning in MLLM.

Acknowledgments

We thank Changdae Oh, Seongheon Park, Samuel Yeh for their valuable comments on the manuscript. This work is supported in part by the AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation under awards IIS-2237037 and IIS-2331669, Office of Naval Research under grant number N00014-23-1-2643, Schmidt Sciences Foundation, Open Philanthropy, Alfred P. Sloan Fellowship, and gifts from Google and Amazon.

Ethical Considerations

This work focuses on analyzing the internal mechanisms of multimodal large language models in in-context learning settings. Our study is empirical and diagnostic in nature, aiming to improve the understanding of cross-modal task mapping rather than training new foundation models. All datasets used in our experiments are publicly available benchmarks for visual question answering and do not contain personally identifiable or sensitive information. In addition, we do not collect new data or involve human subjects in this work.

Limitation

Despite these improvements, our work points to important avenues for future research. First, as an inference-time intervention, MGI operates by guiding the model to utilize its existing mid-layer representations; however, it cannot fundamentally rectify the inherent architectural misalignment between perception and reasoning modules in MLLMs. We hope that our analysis and findings provide a solid foundation for understanding multimodal ICL dynamics and inspire more effective approaches for multimodal adaptation. Second, the method introduces additional hyperparameters (e.g., intervention strength) that may require tuning. To facilitate practical application, we provide comprehensive ablation studies and recommended settings in the Appendix C.3.

References

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

Folco Bertini Baldassini, Mustafa Shukor, Matthieu Cord, Laure Soulier, and Benjamin Piwowarski. 2024. What makes multimodal in-context learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1539–1550.

Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Qin Yao, Philip Torr, Volker Tresp, and Jindong Gu. 2023. Can multimodal large language models truly perform multimodal in-context learning? *arXiv preprint arXiv:2311.18021*.

Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. 2024a. Can multimodal large language models truly perform multimodal in-context learning? *Preprint*, arXiv:2311.18021.

Shuo Chen, Jianzhe Liu, Zhen Han, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. 2025a. True multimodal in-context learning needs attention to the visual context. *Preprint*, arXiv:2507.15807.

Song Chen, Xinyu Guo, Yadong Li, Tao Zhang, Mingan Lin, Dongdong Kuang, Youwei Zhang, Lingfeng Ming, Fengyu Zhang, Yuran Wang, and 1 others. 2025b. Ocean-ocr: Towards general ocr application via a vision-language model. *arXiv preprint arXiv:2501.15558*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. 2025. Revisiting in-context learning inference circuit in large language models. *Preprint*, arXiv:2410.04468.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

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

Jun Gao, Qian Qiao, Ziqiang Cao, Zili Wang, and Wenjie Li. 2024. Aim: Let any multi-modal large language models embrace efficient in-context learning. *Preprint*, arXiv:2406.07588.

Jun Gao, Qian Qiao, Tianxiang Wu, Zili Wang, Ziqiang Cao, and Wenjie Li. 2025. Aim: Let any multimodal large language models embrace efficient in-context

learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, number 3 in 39, pages 3077–3085.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. 2023a. Explaining emergent in-context learning as kernel regression. *arXiv preprint arXiv:2305.12766*.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023b. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12660–12673.

Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. 2024. Multimodal task vectors enable many-shot multimodal in-context learning. *Advances in Neural Information Processing Systems*, 37:22124–22153.

Chengyue Huang, Yuchen Zhu, Sichen Zhu, Jingyun Xiao, Moises Andrade, Shivang Chopra, and Zsolt Kira. 2025. Mimicking or reasoning: Rethinking multi-modal in-context learning in vision-language models. *arXiv preprint arXiv:2506.07936*.

Hong Jun Jeon, Jason D Lee, Qi Lei, and Benjamin Van Roy. 2024. An information-theoretic analysis of in-context learning. In *Forty-first International Conference on Machine Learning*.

Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization. *arXiv preprint arXiv:2411.11909*.

Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023b. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. 2023c. How to configure good in-context sequence for visual question answering. *arXiv preprint arXiv:2312.01571*.

Yanshu Li. 2025. Advancing multimodal in-context learning in large vision-language models with task-aware demonstrations. *Preprint*, arXiv:2503.04839.

Yanshu Li, Jianjiang Yang, Zhennan Shen, Ligong Han, Haoyan Xu, and Ruixiang Tang. 2025a. Catp: Contextually adaptive token pruning for efficient and enhanced multimodal in-context learning. *Preprint*, arXiv:2508.07871.

Yanshu Li, Jianjiang Yang, Ziteng Yang, Bozheng Li, Hongyang He, Zhengtao Yao, Ligong Han, Yingjie Victor Chen, Songlin Fei, Dongfang Liu, and Ruixiang Tang. 2025b. Cama: Enhancing multimodal in-context learning with context-aware modulated attention. *Preprint*, arXiv:2505.17097.

Yanshu Li, Jianjiang Yang, Tian Yun, Pinyuan Feng, Jinfa Huang, and Ruixiang Tang. 2025c. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. *Preprint*, arXiv:2505.17098.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Joy Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Benoit Dumoulin, and Hanghang Tong. 2025. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. *Preprint*, arXiv:2510.17771.

Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. 2024. How does the textual information affect the retrieval of multimodal in-context learning? *arXiv preprint arXiv:2404.12866*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *Preprint*, arXiv:2202.12837.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. 2025. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *Preprint*, arXiv:2506.09047.

Jane Pan. 2023. What in-context learning "learns" in-context: Disentangling task recognition and task learning. Master's thesis, Princeton University.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. 2024a. What factors affect multi-modal in-context learning? an in-depth exploration. *Preprint*, arXiv:2410.20482.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. 2024b. What factors affect multi-modal in-context learning? an in-depth exploration. *arXiv preprint arXiv:2410.20482*.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. 2024a. The transient nature of emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 36.

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie CY Chan, and Andrew M Saxe. 2024b. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation. In *Forty-first International Conference on Machine Learning*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9840–9855.

Daiqing Wu, Dongbao Yang, Sicheng Zhao, Can Ma, and Yu Zhou. 2025. An empirical study on configuring in-context learning demonstrations for unleashing mllms' sentimental perception capability. *arXiv preprint arXiv:2505.16193*.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Nan Xu, Fei Wang, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. From introspection to best practices: Principled analysis of demonstrations in multimodal in-context learning. *arXiv preprint arXiv:2407.00902*.

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

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. 2023a. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023b. What makes good examples for visual in-context learning? *Advances in Neural Information Processing Systems*, 36:17773–17794.

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

Bowen Zheng, Ming Ma, Zhongqiao Lin, and Tianming Yang. 2024. Distributed rule vectors is a key mechanism in large language models' in-context learning. *arXiv preprint arXiv:2406.16007*.

Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2024. The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14365–14378, Miami, Florida, USA. Association for Computational Linguistics.

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. 2024. Vl-icl bench: The devil in the details of benchmarking multimodal in-context learning. *arXiv preprint arXiv:2403.13164*.

A Dataset and Evaluation Details

A.1 Controlled Outlier Detection Dataset.

To provide a systematic comparison between multimodal and text-only ICL under controlled conditions, we construct a controlled outlier detection task (Chen et al., 2025a; Nikankin et al., 2025). The core objective is to identify a single minority instance—the outlier—that differs from the majority group based on a designated attribute, specifically *shape* or *color*, as illustrated in Fig. 1 (a-b). The label space is defined by 4 distinct shape categories (circle, triangle, square, star) and 10 color categories (e.g., yellow, blue, red, etc.). To isolate the effect of modality, we instantiate each problem in this task in two distinct formats: text-only and multimodal, as illustrated in Fig. 1 (a-b).

Dataset Statistics. In total, the dataset contains 2,000 samples, consisting of 1,000 color-based and 1,000 shape-based instances. For evaluation purposes, we randomly select 500 samples per category to form the query set. The remaining samples constitute the support pool. In our experiments, we adopt an n-shot setting where each query is paired with n demonstrations sampled from the corresponding support pool to facilitate in-context learning.

A.2 Multimodal In-Context Learning Dataset

To validate the effectiveness of MGI, we evaluate it on the TrueMICL benchmark (Chen et al., 2025a) (comprising Outlier Detection, Clock Math, and Operator Induction) and a natural-image datasets (OK-VQA (Marino et al., 2019)). An overview of all datasets used in this work is presented in Fig. 7:

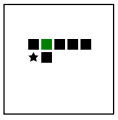
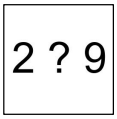
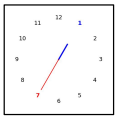
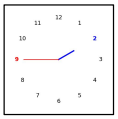
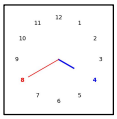
Dataset	Demo1	Demo2	Query	Label	Explanation
Outlier Detection	 green	 red		black	Find the outlier by color feature in the image.
Operator Induction	 1	 8		9	Induce the operation (**add**) in demonstration, calculate the result of the two numbers in the query image.
Clock Math	 7	 18		32	Induce the operation (**multiply**) in demonstration, identify the numbers in the clock; calculate the result of two numbers.
OK-VQA				goalie	Sports and Recreation Question

Figure 7: **An overview of datasets in our papers.** For the datasets sourced from TrueMICL (Chen et al., 2025a) (Outlier Detection, Operator Induction, Clock Math), the label to the query requires the model to learn the relationship between images and text in the demos. Meanwhile, for the natural-image datasets (OK-VQA (Marino et al., 2019)), the model can leverage query-relevant demonstrations to enhance inference on the query instance.

- *Outlier Detection* is designed to assess the model's variable binding capability. Each image depicts varying quantities of objects with two distinct shapes and colors. The model is required to identify the target outlier attribute based on the textual demonstration. The label space is defined by 4 distinct shape categories (circle, triangle, square, star) and 10 color categories (yellow, blue, green, red, black, orange, purple, pink, brown, gray).
- *Operator Induction* consists of simple arithmetic equations formed using three basic operators: addition, subtraction, and multiplication. As the specific operator for each image is not explicitly stated, the model must infer the underlying arithmetic rule by analyzing the relationship between the image content and the target answer.
- *Clock Math* is a task inspired by arithmetic pattern recognition. To correctly answer the query, the model must first decipher the time displayed on the clock face and subsequently deduce the implicit relationship (e.g., addition) between the time and the answer.
- *OK-VQA* includes 9,055 training samples and 5,000 validation samples. This dataset challenges models to integrate external knowledge beyond the immediate image and context to generate accurate answers.

Dataset Statistics. Given the limited number of samples per task in the TrueMICL benchmark (Chen et al., 2025a), we adopted its format to generate new questions and images, thereby enlarging the dataset scale. Consistent with TrueMICL (Chen et al., 2025a), we partitioned the samples in each dataset into a support set of 50 samples and a test set of 290 queries. For the natural-image datasets (OK-VQA (Marino et al., 2019)), we randomly sampled 2,048 instances from the validation set to serve as the test set. Furthermore, following standard multimodal ICL evaluation protocols, we utilized Retrieval-based In-Context Example Selection (RICES) to retrieve relevant instances from the training dataset as demonstrations for each query.

A.3 Evaluation Metrics

For evaluation, we use exact matching for numerical answers and keyword matching for text. For OK-VQA (Marino et al., 2019), we report VQA accuracy, i.e., exact match accuracy over

a set of ground truth answers. Experiments are conducted on Qwen2.5-VL (Bai et al., 2025) and Gemma3 (Team et al., 2025) across different model scales to ensure robustness.

B Models and Implementation Details

MLLMs. In this paper, we evaluate the effectiveness and generality of our MGI using recent models from two representative VLM families: Qwen2.5-VL (Bai et al., 2025) and Gemma 3 (Team et al., 2025). For reproducibility, we list all models with their publicly available checkpoints on Hugging Face[1]. We implement all experiments using the `transformers` and `PyTorch` libraries on two NVIDIA A100 (80GB) GPUs under `bfloat16 mixed precision`. For our attention analysis, we use the setting `attn_implementation='eager'` to retrieve raw attention outputs.

We also attempted to include LLaVA-Next (Liu et al., 2024) and InternVL 3.5 (Chen et al., 2024b); however, these models exhibited significant memory leakage when extracting raw attention maps during inference. Specifically, even the 7B/8B parameter variants caused out-of-memory (OOM) errors on an 80GB A100 GPU under bfloat16 precision. Consequently, we exclude them from the current analysis and leave the resolution of these implementation constraints to future work.

Implement Details of Our MGI. In the main text, we formulated the attention intervention as an additive process (Eq. 3) to provide an intuitive understanding of the mechanism. However, empirically, we observed that directly adding the attention maps yielded suboptimal performance. In our actual implementation, we adopt a selective scaling strategy. Guided by the estimated task mapping $\hat{\mathcal{M}}$ in Eq. 2, we identify the specific image tokens that serve as visual evidence for the demonstration label—the critical features the model should attend to when answering the query.

Specifically, we first calculate the mean attention value μ of the label-to-image attention map $\mathbf{a}^{(\ell^*,h)}_{\text{lbl}_i \to \text{img}_i}$. We then generate a set of salient indices $\mathcal{S}$ containing only the tokens with attention scores exceeding k times this average:

$$\mathcal{S} = \left\{ j \mid \mathbf{a}^{(\ell^*,h)}_{\text{lbl}_i \to \text{img}_i}[j] > k \cdot \mu \right\}, \quad \text{where } \mu = \frac{1}{N} \sum_{j=1}^{N} \mathbf{a}^{(\ell^*,h)}_{\text{lbl}_i \to \text{img}_i}[j] \tag{4}$$

where N is the number of image tokens. Finally, we amplify the query's attention towards these evidence tokens by a factor λ while leaving others unchanged:

$$\tilde{\mathbf{a}}^{(\ell,h)}_{\text{q} \to \text{img}_i}[j] = \begin{cases} \lambda \cdot \mathbf{a}^{(\ell,h)}_{\text{q} \to \text{img}_i}[j] & \text{if } j \in \mathcal{S} \\ \mathbf{a}^{(\ell,h)}_{\text{q} \to \text{img}_i}[j] & \text{otherwise,} \end{cases} \tag{5}$$

Where the strength of the intervention $\lambda > 1$.

After this intervention, we re-normalize $\tilde{\mathbf{a}}$ to ensure it forms a valid probability distribution. In our experiments, we typically set $k = 1.5$.

C Additional Analysis Study

C.1 Generality beyond controlled synthetic settings.

Setup. To evaluate whether our observations extend beyond controlled synthetic settings, we conduct additional experiments on a more realistic multimodal reasoning benchmark derived from MM-Vet. Specifically, we sample 50 VQA instances and manually annotate ground-truth evidence regions for each question-answer pair. We analyze layer-wise attention patterns on both Qwen2.5-VL (7B/32B) (Bai et al., 2025) and Gemma-3 (12B/27B) (Team et al., 2025). For each layer, we measure (i) attention from demonstration label tokens to ground truth evidence regions versus other regions, and (ii) attention from the query's final token to demonstration labels and their corresponding visual evidence (refer to Sec. 4.1).

Mid-layer construction of grounded mappings. We first examine how attention to ground truth evidence regions evolves across layers. Across all models, we observe a consistent stage-wise pattern. Early layers exhibit uniformly low and comparable attention to both correct and irrelevant regions. In contrast, middle layers show a sharp concentration on grounded regions, where attention to correct regions becomes significantly higher (typically 3–9×) than to other regions.

However, this separation diminishes again in later layers, where attention to grounded regions drops and becomes comparable to or even lower

[1] **Qwen2.5 VL 7B**https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct
https://huggingface.co/google/gemma-3-12b-it
https://huggingface.co/google/gemma-3-27b-it

than that to irrelevant regions. For example, in Gemma-3-12B, attention to correct regions peaks in middle layers (e.g., L11–L23: ∼0.02–0.05 vs. ∼0.003–0.006), but declines in later layers (e.g., L30+: ∼0.0015 vs. ∼0.0032).

This suggests that multimodal task mappings are strongly constructed in mid-layers but are not reliably preserved during the later stages of inference.

Late-layer consumption and temporal mismatch. We further analyze how the query's final token attends to demonstration labels and their associated visual evidence. If task mappings were successfully transferred, the query token would attend to these elements within the same layers where mappings are formed.

Instead, we observe a temporal mismatch. In early and middle layers, attention from the query token to demonstrated evidence remains weak and diffuse. This attention increases sharply only in much later layers, where the query begins to strongly attend to labels and grounded regions.

For instance, in Gemma-3-12B, attention remains low in middle layers (e.g., L10–L20: $\sim 10^{-4}$), but rises significantly in later layers (e.g., L29+: $\sim 10^{-3}$). Notably, this surge occurs after the layers identified as responsible for constructing task mappings.

Summary. These results consistently reveal a *mid-layer construction vs. late-layer transfer failure* pattern across all evaluated architectures. While multimodal models successfully construct grounded task mappings in middle layers, these mappings are not effectively preserved or utilized during the final stages of inference. This phenomenon generalizes beyond synthetic settings and holds across both Qwen2.5-VL and Gemma-3 models.

C.2 Attention-Based Analysis of Text and Visual Token Contributions

Setup. Recent work has shown that MLLMs (Chen et al., 2025a; Li et al., 2025a), on average, assign significantly less attention to visual tokens compared to textual ones in demonstrations. Here, we take a step further by examining how this imbalance evolves across layers. For each layer, we compute the Relative Attention per Token (Liu et al., 2025) (RAT), defined as the ratio between the average attention from the last token (query) to each image or text token in the demonstrations (key). This metric reflects how much attention each modality receives on a per-token basis, rather than describing the overall distribution of attention. Here, we compute the RAT on Qwen2.5-VL-7B and Gemma-3-12B under the 4-shot setup, and present the results in Fig. 8.

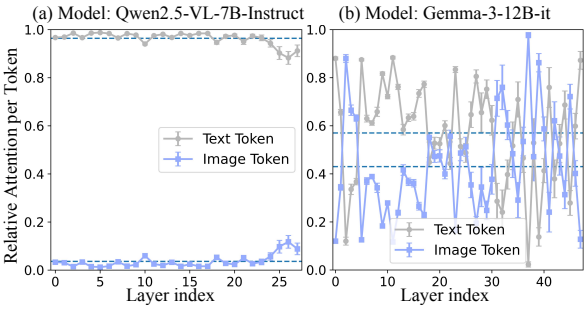


Figure 8: **Layer-wise Relative Attention per Token of demonstration text and image tokens for different MLLMs families under the 4-shot setup.** Qwen2.5-VL-7B shows a consistently text-dominant pattern across all layers, while Gemma-3-12B displays a modality-switching pattern.

Layer-wise Attention Allocation to Image and Text Demonstrations. The results in Fig. 8 reveal clear imbalances in how the model allocates attention to different modalities within the demonstrations. Across models, we observe two distinct patterns in how attention is allocated to text versus image demonstration tokens: (a) *Text-Dominant Pattern.* Qwen2.5-VL-7B exhibits a stable text-dominant pattern across all layers, with visual tokens receiving only negligible attention; (b) *Modality-Switching Pattern.* Gemma-3-12B displays a modality-switching pattern, where attention alternates between text-dominated and image-dominated layers rather than integrating both modalities jointly. Despite their differences, both behaviors reflect the same underlying limitation: the model does not integrate textual and visual information from the demonstrations in a balanced or parallel manner. Instead, attention is assigned to one modality at a time, resulting in sequential rather than fused Multimodal processing.

Layer-wise Attention Head Behavior on Demonstrations. To further examine how the model uses the demonstrations, we analyze how each attention head behaves under the 4-shot setup (see Fig. 9). By inspecting a randomly selected example, we track the attention from the last token to the demonstration text and image tokens across all layers and heads. The heatmaps show that only a few heads exhibit meaningful attention to demonstra-

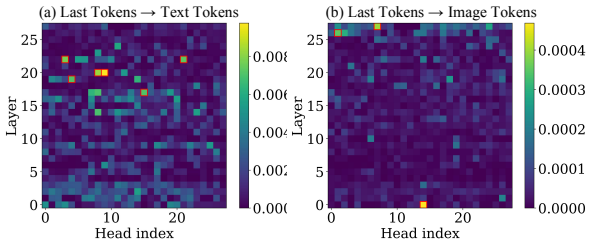


Figure 9: **Layer-wise Attention Head Activation Patterns on Image and Text Demonstrations under the 4-shot setup.** Only a few heads show noticeable attention, while most heads remain inactive.

tion text, while attention to demonstration image tokens is even sparser. Most heads remain nearly inactive for both modalities. ***These results indicate that multimodal ICL relies on only a small subset of specialized heads rather than broad distributed attention.***

C.3 Hyper-Parameter Analysis

To ensure the robustness and optimal performance of MGI, we conduct a comprehensive sensitivity analysis on its two key hyperparameters: the start layer L_{start} and the intervention strength λ. All experiments in this section are performed using the Qwen2.5-VL-7B model on the Outlet Detection dataset.

λ	1	1.5	2	2.5	3	3.5	4	4.5	5
Acc.	68.28	68.28	**68.97**	68.28	68.28	68.62	68.62	**68.97**	67.93

Table 4: Effect of the strength of the intervention λ.

Effect of the strength of the intervention λ. We first analyze the impact of the steering vector's magnitude ($\lambda > 0$). We vary λ within the range of $\{1.5, \ldots, 5\}$ while we fix L_{start} to a constant value (e.g., $L_{start} = 14$). As shown in Table 4, we observe a trade-off: a small λ may not provide enough guidance to correct the model's behavior, whereas an excessively large λ risks dominating the original features and degrading generation quality. The empirical results indicate that $\lambda = 2$ yields the best balance between effectiveness and stability. Based on these empirical results, we adopt $\lambda = 2$ for all datasets in the TrueMICL benchmark. However, for the two VQA datasets, we find that a stronger intervention is necessary to effectively steer the model, thus, we set $\lambda = 6$ for those tasks.

Effect of the start layer L_{start}. Next, we evaluate the influence of the layer where the intervention begins. Guided by the hypothesis that intermediate layers play a crucial role in forming the target concept, we focus our search on the middle block of the model, specifically exploring $L_{start} \in \{11, 12, \ldots, 19\}$. We keep the strength of the intervention fixed at the optimal value identified previously, *i.e.* $\lambda = 2$. The results, summarized in Table 5, demonstrate that the performance is sensitive to the insertion depth, with the optimal performance achieved at layer $L_{start} = 14$. This suggests that intervening too early may disrupt low-level feature extraction, while intervening too late may fail to sufficiently steer the model's high-level reasoning. Meanwhile, the optimal insertion depth varies depending on the total depth of the model architecture. Specifically, for the *Qwen2.5-VL* family, we set $L_{start} = 16$ for the 7B model and $L_{start} = 40$ for the 32B model. Similarly, for the *Gemma-3* family, we utilize $L_{start} = 17$ for the 12B model and $L_{start} = 37$ for the 27B model.

λ	11	12	13	14	15	16	17	18	19
Acc.	68.28	67.93	68.02	**68.97**	68.28	68.28	67.93	67.93	67.93

Table 5: Effect of the start layer L_{start}.

Setup	1-shot	2-shot	3-shot	4-shot
Baseline	65.52%	66.55%	67.59%	69.09%
Ours (MGI)	66.90%	67.59%	67.93%	70.17%

Table 6: Effect of the number of demonstrations (k) on Outlier Detection performance with Qwen2.5-VL-7B.

Effect of varying the number of demonstrations. We conduct experiments to analyze how the number of demonstrations (k) affects model performance. Specifically, we compare our MGI method with a baseline on the Outlier Detection dataset using Qwen2.5-VL-7B, where $k \in [1, 2, 3, 4]$.

Notably, while performance improves as the number of demonstrations increases, the gains remain relatively modest, indicating that vanilla multimodal in-context learning (MM-ICL) struggles to fully utilize additional examples. In contrast, MGI consistently enhances demonstration utilization, yielding stable improvements across all values of k.

C.4 Inference Time Cost Analysis

We evaluated the inference latency of MGI using Qwen2.5-VL-7B on *Outlier Detection* dataset. The results indicate that our method maintains high efficiency during the generation phase, with an average inference time of **842.41 ms**, which is comparable

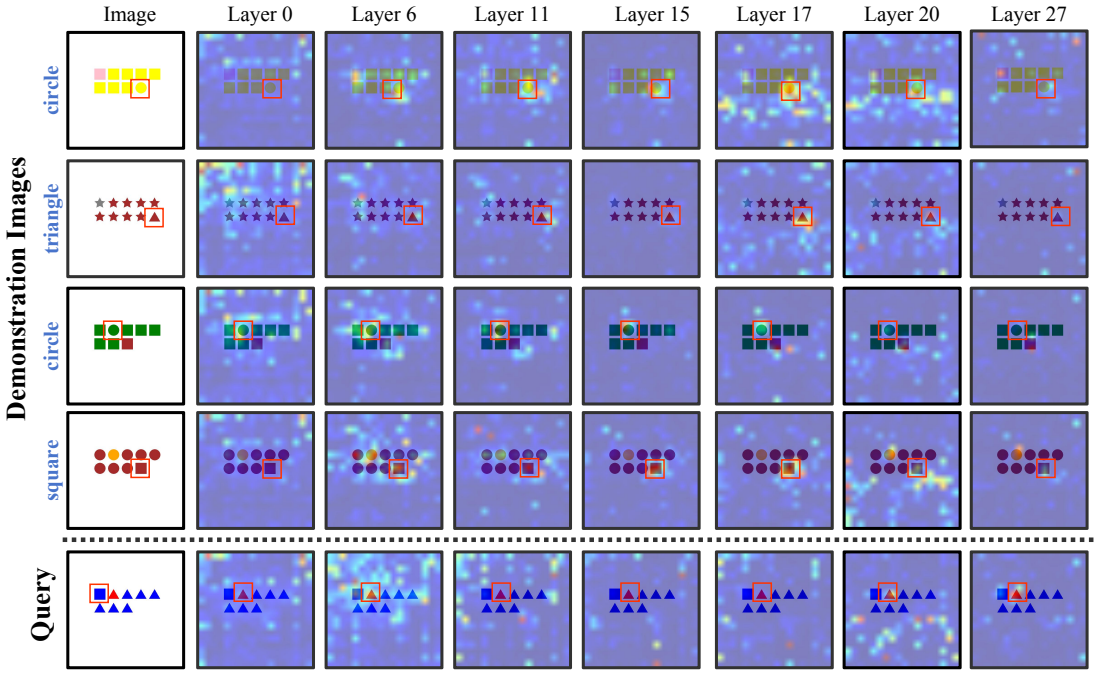


Figure 10: **Additional layer-wise visualization of attention from demonstration label tokens (or the last token) to image tokens in a multimodal ICL example.** The four demonstrations are labeled by the shape outlier, yet the model exhibits *Correct Task Recognition but False Answer* on the query, incorrectly predicting `triangle` instead of `square`. Red bounding boxes denote ground-truth evidence regions. Demonstration label tokens form clear object-level grounding only in mid layers, with early layers being diffuse and deeper layers becoming noisy again. In contrast, the query token shows no meaningful grounding until the final several layers, where localization finally appears.

to the baseline's **800.11 ms**. This negligible overhead (∼5%) demonstrates that the proposed attention intervention (Eq. 3) is lightweight and does not hinder the token decoding speed.

However, the primary increase in total latency stems from the preparation phase, specifically identifying the peak grounding layer for each query, which incurs a setup cost of **876.34 ms**. This step involves aggregating full attention distributions to compute entropy. Crucially, this is a **per-query pre-computation** that depends solely on the input context and does not accumulate with the number of generated tokens, ensuring the method remains efficient even when generating long responses.

C.5 Additional Layer-wise Visualization of Attention

In this section, we present an additional case study visualizing the attention from demonstration label tokens to image regions across layers. In contrast to the *False Task Recognition* case discussed in the main text, Figure 4 illustrates a different failure mode: *Correct Task Recognition but False Answer*.

As shown in Figure 10, we observe that although the model appears to correctly recognize the task (identifying the outlier based on shape features), it fails to execute it correctly on the query instance. Specifically, the attention maps reveal that the model focuses on an incorrect candidate object (`triangle`) in the query image rather than the true outlier (`square`). Consequently, despite the correct task formulation, this misdirected attention leads to an erroneous final prediction.

D Additional Discussion

Usage of Artifacts and AI Assistants. We utilized publicly available models and datasets from Hugging Face, strictly following their licenses for non-commercial research. These models and datasets have been reviewed by their developers/creators to minimize the inclusion of personally identifiable information or offensive content and are widely adopted by the research community. We used AI tools to assist with language refinement during the writing process, the paper contains no AI-generated paragraphs. All material has been carefully reviewed to ensure accuracy and adherence to ethical standards.