ICL-R: RECONSTRUCTING THE COMPUTATIONAL EFFECTS OF DEMONSTRATIONS FOR IMPLICIT MULTIMODAL IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal in-context learning (ICL) enables multimodal large language models (MLLMs) to adapt to downstream tasks through interleaved image and text demonstrations. However, retaining these demonstrations not only increases the sequence length and latency at inference time but also makes predictions sensitive to how demonstrations are selected and presented. Although existing implicit ICL methods approximate demonstration effects through representation shifts or attention routing, they still suffer from coarse or partial reconstruction of these effects. To address these limitations, we first decompose the change in attention output caused by demonstrations into the attention routing shift, the value shift, their interaction, and information contributed directly by demonstration tokens. Based on this decomposition, we propose ***I**n-**C**ontext **L**earning via **R**econstruction (**ICL-R**)*, a lightweight framework for jointly reconstructing these computational effects. Aiming at incomplete reconstruction under demonstration removal, we propose *mass-restoring joint attention-value calibration* which jointly calibrates attention routing and value representations while rescaling the attention output. To further extend supervision beyond isolated state matching, we introduce a *Riemannian state transition objective* that aligns both incoming states and the transitions induced by attention. Extensive experiments across multiple MLLMs and benchmarks show that ICL-R improves performance while maintaining an inference cost close to that of standard zero-shot inference.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Li et al., 2023b; Liu et al., 2023a) provide a general interface for tasks including visual question answering (Shao et al., 2023; Goyal et al., 2017), image captioning (Chen et al., 2015), and visual reasoning (Xu et al., 2026). A key approach to adapting these models to downstream tasks is multimodal in-context learning (ICL), which provides an MLLM with interleaved image and text demonstrations of inputs followed by a target query. The main advantage of this mechanism lies in its ability to support rapid and flexible adaptation without any updates to model parameters (Zong et al., 2025). However, such explicit ICL is costly with each demonstration substantially adding visual and textual tokens that increase memory use and inference latency (Ma et al., 2026). The performance of ICL also strongly depends on demonstration selection and format, which determine the task-specific cues and how the outputs are presented (Li et al., 2024b; Qin et al., 2024). Even with the same demonstrations, changing their order can lead to different predictions (Zhang et al., 2023; Baldassini et al., 2024). Therefore, demonstrations may provide limited gains or distract the model from the target query (Gupta et al., 2026). To address these limitations, implicit multimodal ICL captures demonstration effects within the model, allowing it to answer queries without prefixing demonstrations to the input.

Early studies on task and function vectors (Hendel et al., 2023; Todd et al., 2024) show that tasks conveyed through ICL can be encoded as compact vectors within language models. These vectors are extracted from internal activations and injected into hidden states to elicit the corresponding task behavior on new queries. ICV (Liu et al., 2023b) develops this perspective into a steering approach by extracting a shared direction from differences between demonstration input and output representations. Following this idea, LIVE (Peng et al., 2024) and MimIC (Jiang et al., 2025) learn additive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

representation shifts. M^2IV (Li et al., 2025b) similarly approximates demonstration effects through separate learnable shifts to the outputs of attention and feedforward modules at each layer. HiFICL (Li et al., 2026) introduces virtual key-value pairs to approximate demonstration effects. However, these designs do not explicitly distinguish interventions on attention scores from changes to the information carried by value vectors, which may limit their ability to adjust the relative attention scores among query tokens. In contrast, attention routing methods such as ICR (Li et al., 2025a) and Hyper-ICL (Talemi et al., 2026) focus on adjusting the attention weights assigned to query tokens, but their reconstruction of demonstrations remains incomplete. Therefore, the remaining challenge is to identify and reconstruct the complementary computational effects induced by demonstrations.

In this paper, we first compare attention outputs with and without demonstrations, decomposing the difference into changes in attention routing and value representations, their interaction, and the direct contribution of demonstration tokens. This decomposition reveals that demonstrations reshape information selection and content propagation in a coupled manner while contributing additional information directly. Motivated by both the limitations of existing methods and this empirical decomposition, we propose ***I**n-**C**ontext **L**earning via **R**econstruction (**ICL-R**)* , a query-only framework that reconstructs the computational effects of demonstrations from a frozen teacher to a student with lightweight modules. Specifically, we propose joint attention and value calibration to reconstruct the coupled changes in information selection and content propagation, with query-adaptive corrections accounting for variation across inputs. The accompanying missing mass estimation compensates for the attention normalization mismatch caused by removing demonstrations, rescaling the query contribution. Our Riemannian state transition objective extends supervision beyond isolated state matching by aligning incoming states on the unit hypersphere and the direction and magnitude of attention-induced transitions in the corresponding tangent space. We evaluate ICL-R across different model architectures on six datasets covering visual question answering, image captioning, and multimodal perception and reasoning. Additionally, we analyze the complementary roles of attention and value calibration and evaluate the contributions of the individual components in ICL-R. We further investigate the data efficiency, inference efficiency, robustness to varying demonstration counts, sensitivity to rank configurations, and object hallucination.

In summary, our main contributions in this work are as follows:

- We formulate implicit ICL as the problem of the reconstruction of demonstration effects under demonstration removal and establish an exact mathematical decomposition of these effects in attention output. This view reveals that demonstration effects arise from more than a single representation shift, providing the basis for our explicit reconstruction.
- We propose ICL-R, a lightweight framework that reconstructs information selection and content propagation through joint attention and value calibration with missing mass estimation correcting the output scale change caused by removing demonstrations.
- We introduce a Riemannian objective that aligns the states entering attention on the unit hypersphere and the direction and magnitude of their attention-induced transitions in the corresponding tangent spaces to more accurately capture the effects of demonstrations.

2 Related Work

Multimodal In-Context Learning. Multimodal Large Language Models provide general capabilities for vision and language understanding (Alayrac et al., 2022; Li et al., 2023b). Building on these capabilities, multimodal in-context learning adapts models to new tasks by inputting interleaved image and text demonstrations and a target query without parameter updates (Tsimpoukelli et al., 2021). However, visual demonstrations substantially increase input length and inference cost (Jiang et al., 2025; Li et al., 2025b; Chen et al., 2024), while performance also depends on demonstration selection and configuration (Li et al., 2024b; Sun et al., 2025). Selecting useful examples is particularly challenging because similarity in one modality may overlook information required in the other, allowing similar questions to induce answer copying that conflicts with the query image (Li et al., 2024b). To improve demonstration quality, prior studies investigate prompt retrieval in visual ICL (Zhang et al., 2023), sequence configuration for visual question answering (Li et al., 2024b), and multimodal context retrieval and refinement (Fu et al., 2026b). Although these strategies improve the use of demonstrations, they continue to process explicit examples during inference and therefore retain the associated computational burden and memory overhead.

Implicit In-Context Learning. To address the inference overhead, implicit ICL seeks to preserve the benefits of demonstrations through representation engineering (Snell et al., 2022; Li et al., 2025c). Early studies like task and function vector (Hendel et al., 2023; Todd et al., 2024; Liu et al., 2023b) extract task representations from the activations and inject them at empirically selected layers to guide responses to new queries. Building on this idea of vector-based intervention, LIVE (Peng et al., 2024) learns an in-context vector for each layer, while MimIC (Jiang et al., 2025) introduces separate shifts after each attention head and adjusts their strength according to the query. M^{2}IV (Li et al., 2025b) further applies separate learnable shifts to the outputs of attention and feedforward modules at each layer. Moving beyond shift vectors, HiFICL (Li et al., 2026) learns virtual key and value pairs, allowing each query to dynamically combine the original attention output with learned context values. Although effective, these designs conflate changes in attention distribution with changes in the information carried by value vectors (Talemi et al., 2026). Attention routing methods instead directly adjust the attention weights assigned to tokens, providing direct calibration over how information is selected during aggregation. ICR (Li et al., 2025a) learns to combine reusable attention directions, while Hyper-ICL (Talemi et al., 2026) applies low-rank attention-logit corrections with query-dependent strength. However, they capture only part of the computation induced by demonstrations. To deal with these limitations, our ICL-R explicitly separates and reconstructs the effects of demonstrations in attention routing and value content with missing mass correction accounting for the change in attention normalization after demonstration removal.

Geometric Intermediate Representation Distillation. Intermediate representation distillation transfers knowledge by supervising internal features beyond task predictions. Early work such as FitNets (Romero et al., 2015) introduced feature regression to guide student learning with teacher representations. Subsequent work extended supervision to relationships between layers, aiming to transfer the teacher's solution process (Yim et al., 2017). Within implicit multimodal ICL, MimIC (Jiang et al., 2025) combines task supervision with squared ℓ_2 alignment between teacher and student hidden states. HiFICL (Li et al., 2026) relies on cross-entropy supervision alone, while Hyper-ICL (Talemi et al., 2026) uses Lorentz geodesic distances for intermediate representation alignment. However, task supervision constrains output predictions, while the intermediate alignment objectives above compare absolute teacher and student hidden states at individual layers. To account for both the state transitions before and after the attention, we draw on Riemannian geometry, which provides geodesic distances for comparing states and logarithmic maps for describing local displacements in tangent spaces (Pennec, 2006; Lin et al., 2021). We represent normalized hidden states on the unit hypersphere and jointly supervise pre-attention states and attention-induced transitions in their corresponding tangent spaces. This objective extends pointwise hidden-state matching by jointly aligning pre-attention states and transitions represented as tangent vectors.

3 METHOD

In this section, we first decompose demonstration effects to guide their modeling in Section 3.1. We then present mass-restoring joint attention–value calibration, integrating routing and value corrections with missing mass estimation, in Section 3.2. Finally, we introduce the Riemannian state transition loss for training the query-only student in Section 3.3.

3.1 MOTIVATION AND PROBLEM FORMULATION

Following prior work on transferring demonstration effects to query-only models (Jiang et al., 2025; Talemi et al., 2026), we consider a training episode containing a set of demonstrations $\mathcal{D}$ and a target query X. The teacher receives the concatenated context $[\mathcal{D}; X]$, while the student receives only X. For one self-attention head with projected dimension d_h, let $H_{T,X}$ and $H_{T,D}$ denote the teacher hidden states for query and demonstration tokens. The teacher projections are:

$$Q_{T,X} = H_{T,X}W_Q, \qquad K_T = [H_{T,D}; H_{T,X}]W_K, \qquad V_T = [H_{T,D}; H_{T,X}]W_V. \tag{1}$$

We partition the resulting attention matrix and output into demonstration and query components to distinguish information retrieved from demonstrations from that aggregated over query tokens:

$$A_T = \mathrm{softmax}\left(\frac{Q_{T,X}K_T^\top}{\sqrt{d_h}}\right) = [A_{T,D}, A_{T,X}], \qquad O_T = A_{T,D}V_{T,D} + A_{T,X}V_{T,X}, \tag{2}$$

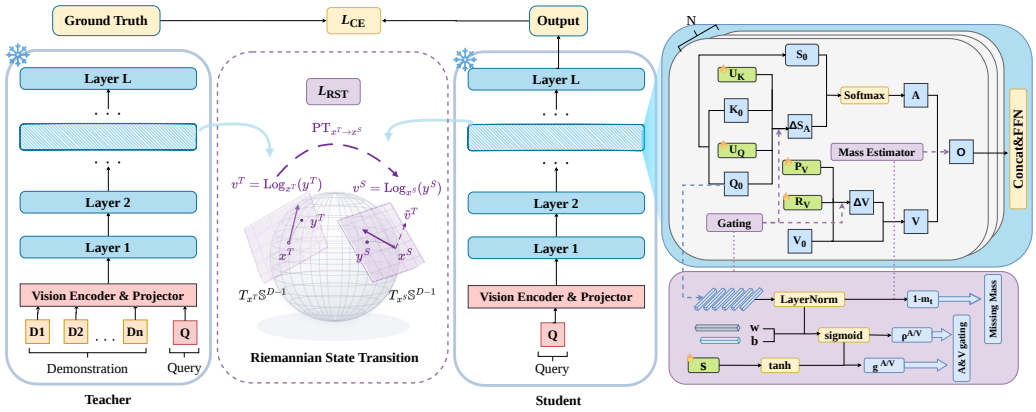


Figure 1: **Overview of ICL-R.** Our method transfers the effects of demonstrations from a frozen teacher to a query-only student through mass-restoring joint attention-value calibration and Riemannian state transition alignment. The student receives only the query without demonstrations.

where $A_{T,D} \in \mathbb{R}^{n_X \times n_D}$ captures attention from query tokens to demonstrations and $A_{T,X} \in \mathbb{R}^{n_X \times n_X}$ captures attention among query tokens, with n_X and n_D denoting the numbers of query and demonstration tokens. For the frozen query-only model, let $Q_0 = H_0 W_Q$, $K_0 = H_0 W_K$, and $V_0 = H_0 W_V$ denote the query, key, and value projections of the query-only hidden states H_0 respectively. The attention matrix and corresponding output are computed as follows:

$$A_0 = \mathrm{softmax}\left(\frac{Q_0 K_0^\top}{\sqrt{d_h}}\right), \quad O_0 = A_0 V_0. \tag{3}$$

To expose the effect of demonstrations, define the query-side changes $\Delta A_T = A_{T,X} - A_0$ and $\Delta V_T = V_{T,X} - V_0$. Substituting them into Eq. (2) and subtracting the base output in Eq. (3) yields:

$$O_T - O_0 = \underbrace{\Delta A_T V_0}_{\text{routing shift}} + \underbrace{A_0 \Delta V_T}_{\text{value shift}} + \underbrace{\Delta A_T \Delta V_T}_{\text{interaction}} + \underbrace{A_{T,D} V_{T,D}}_{\text{demonstration information}}. \tag{4}$$

The four terms in Eq. (4) respectively capture changes in routing, value content, their interaction, and information retrieved directly from demonstrations. On this basis, our goal is to learn a student with lightweight modules whose output $\widetilde{O}$ approximates O_T, as summarized in Figure 1.

3.2 Mass-Restoring Joint Attention-Value Calibration

The decomposition in Eq. (4) motivates reconstructing demonstration effects by jointly calibrating attention and value content. This requires calibrating not only where information is selected and what content is propagated, but also how much attention mass is assigned to query content. Specifically, for a query token indexed by t, the teacher distributes its attention over both demonstrations and query tokens, with m_t^T denoting the attention mass assigned to demonstrations. Consequently, the total mass assigned to query tokens is reduced as expressed below:

$$m_t^T = \sum_{s \in \mathcal{D}} (A_{T,D})_{t,s}, \qquad \sum_{s \in X} (A_{T,X})_{t,s} = 1 - m_t^T. \tag{5}$$

Removing demonstrations forces a query-only softmax to assign unit mass to query tokens, changing the scale of their contribution. We incorporate mass correction into the attention weights that aggregate calibrated values. Let q_t be the student query projection, $\bar{q}_t = \mathrm{LN}(q_t)$ its normalized form, $S_0 = Q_0 K_0^\top / \sqrt{d_h}$ the base logits before calibration at the current layer and ΔS^A adjusts these logits to redistribute attention among query tokens. The resulting calibrated output is:

$$\widetilde{O}_t = \sum_s \widehat{A}_{t,s} \widetilde{V}_{t,s}, \qquad \widehat{A}_{t,s} = \left[1 - s_m \, \sigma\left(\frac{\bar{q}_t^\top w_m}{\sqrt{d_h}} + b_m\right)\right] \left[\mathrm{softmax}(S_0 + \Delta S^A)\right]_{t,s}. \tag{6}$$

Here, $\widehat{A}$ combines calibrated routing with missing mass estimation, and $\widetilde{V}_{t,s} = V_{0,s} + \Delta V_{t,s}$ denotes the calibrated value content. The trainable $w_m \in \mathbb{R}^{d_h}$ and $b_m \in \mathbb{R}$ parameterize a learned proxy

for the demonstration attention mass, while s_m sets its maximum value. Within this unified output, we parameterize the routing and value corrections jointly through query-adaptive low-rank updates. The scalar gates g_t^A and g_t^V control the overall strengths of the two corrections for query token t, while ρ_t^A and ρ_t^V modulate the contributions of their individual low-rank directions:

$$\Delta S_{t,s}^A = g_t^A \left[\rho_t^A \odot (Q_0 U_Q)_{t,:}\right] (K_0 U_K)_{s,:}^\top, \qquad \Delta V_{t,s} = g_t^V \left[\left(\rho_t^V \odot V_{0,s} P_V\right) R_V\right]. \tag{7}$$

Here, $U_Q, U_K \in \mathbb{R}^{d_h \times r_A}$ map the uncalibrated queries and keys into a low-rank attention subspace, while $P_V \in \mathbb{R}^{d_h \times r_V}$ and $R_V \in \mathbb{R}^{r_V \times d_h}$ project values into and out of a low-rank value subspace. The score correction changes the relative allocation within $\widehat{A}$, while the value correction changes the content aggregated under these mass-aware weights. The ranks satisfy $r_A, r_V \ll d_h$, keeping the updates lightweight. Both low-rank updates use the same query-adaptive design. For component $c \in \{A, V\}$, let $u_t^c = \bar{q}_t^\top w_c + b_c$ be the scalar score derived from the current query token. We use this score to define gates for correction strength and the contribution of each low-rank direction:

$$g_t^c = 2\tanh(s_c)\,\sigma(u_t^c), \qquad \rho_{t,j}^c = 2\sigma\left(u_t^c \gamma_j^c + \eta_j^c\right), \qquad j = 1, \ldots, r_c. \tag{8}$$

where $w_c \in \mathbb{R}^{d_h}$, $b_c, s_c \in \mathbb{R}$, and $\gamma^c, \eta^c \in \mathbb{R}^{r_c}$ are trainable parameters. The scalar gate $g_t^c \in (-2, 2)$ sets the signed correction strength, while $\rho_t^c \in (0, 2)^{r_c}$ selects the active low-rank directions.

3.3 Riemannian State Transition Loss

Existing implicit ICL methods use cross-entropy supervision either alone (Li et al., 2026) or together with intermediate representation alignment based on Euclidean or hyperbolic distances (Jiang et al., 2025; Talemi et al., 2026). Cross-entropy loss supervises only the final prediction, while directly matching hidden states with a Euclidean distance constrains only their endpoints. Neither objective directly constrains the transition from the pre-attention state to the post-attention state. We therefore model self-attention as a state transition in the tangent space on the hypersphere and distill both its starting point and its demonstration-induced motion. For each aligned query token indexed by $i = (b, \ell, t)$, let $h_i^a \in \mathbb{R}^D$ be the residual state before self-attention and $o_i^a \in \mathbb{R}^D$ its attention update where $a \in \{T, S\}$ denotes the teacher or student and D is the hidden dimension of the language model. We map the states before and after attention onto the unit hypersphere:

$$x_i^a = \pi(h_i^a), \qquad y_i^a = \pi(h_i^a + o_i^a), \qquad \pi(z) = \frac{z}{\|z\|_2}. \tag{9}$$

Here, x_i^a represents the normalized state entering self-attention, while y_i^a represents the normalized state after applying the attention update o_i^a. Both lie on $\mathbb{S}^{D-1}$, and the transition $x_i^a \to y_i^a$ captures the effect of self-attention on that token. We encode this transition as a tangent vector through the spherical logarithmic map. For two points $x, y \in \mathbb{S}^{D-1}$ with $y \neq \pm x$, we define:

$$\mathrm{Log}_x(y) = \frac{\theta}{\|u\|_2} u. \tag{10}$$

Here, $u = y - \langle x, y\rangle x$ and $\theta = \mathrm{atan2}(\|u\|_2, \langle x, y\rangle)$. Using the logarithmic map in Eq. (10), we represent each attention-induced transition as a tangent vector $v_i^a = \mathrm{Log}_{x_i^a}(y_i^a) \in T_{x_i^a}\mathbb{S}^{D-1}$. This tangent vector represents the state change induced by self-attention. Its direction specifies where the state moves, while its norm $\|v_i^a\|_2 = \theta$ quantifies the magnitude of the change. Since v_i^T and v_i^S lie in different tangent spaces, they cannot be compared directly. We therefore parallel-transport the teacher transition from the tangent space at x_i^T to that at x_i^S along the shortest spherical geodesic:

$$\widetilde{v}_i^T = v_i^T - \frac{\langle v_i^T, x_i^S\rangle}{1 + \langle x_i^T, x_i^S\rangle}(x_i^T + x_i^S). \tag{11}$$

We then align both the pre-attention state and the attention-induced transition. Let $\mathcal{I}$ be the set of aligned batch, layer, and query-token indices. The point and fiber losses are:

$$\mathcal{L}_{\text{point}} = \frac{4}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \arcsin^2\left(\frac{\|x_i^S - x_i^T\|_2}{2}\right), \qquad \mathcal{L}_{\text{fiber}} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \|v_i^S - \widetilde{v}_i^T\|_2^2. \tag{12}$$

The point term $\mathcal{L}_{\text{point}}$ in Eq. (12) uses the squared spherical geodesic distance to align the states entering self-attention. The fiber term $\mathcal{L}_{\text{fiber}}$ compares the student transition with the transported teacher transition from Eq. (11). The complete training objective is:

$$\mathcal{L}_{\text{RST}} = \mathcal{L}_{\text{point}} + \mathcal{L}_{\text{fiber}}, \qquad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{RST}}. \tag{13}$$

Here, $\lambda \geq 0$ is a hyperparameter that balances task supervision and geometric alignment. The complete algorithmic workflow is summarized in Appendix A.3.

4 EXPERIMENTS

Implementation details. To assess the effectiveness and generalization of our method, we evaluate it on two fully autoregressive models, Idefics2-8B-Base (Laurençon et al., 2024) and LLaVA-Interleave-7B (Li et al., 2024a) across six multimodal benchmarks. Specifically, VQAv2 (Goyal et al., 2017) and OK-VQA (Marino et al., 2019) evaluate general and knowledge-based visual question answering. COCO Caption (Chen et al., 2015) and Flickr30K (Young et al., 2014) evaluate image captioning, while MME (Fu et al., 2026a) and SEED-Bench (Li et al., 2023a) assess broader multimodal perception and reasoning. We report answer accuracy for VQAv2, OK-VQA, MME, and SEED, and CIDEr (Vedantam et al., 2015) for COCO Caption and Flickr30K. During training, the frozen teacher receives demonstrations together with the query, while the student receives only the query. We keep the vision encoder, projector, and language model frozen, updating only our modules. For each benchmark, we adopt eight sampled demonstrations per episode, set both calibration ranks to 8 and optimize our modules with AdamW (Loshchilov & Hutter, 2017). Evaluation subsets, sampling procedures, and additional implementation settings are detailed in Appendix A.2.

Baselines and main results. We compare ICL-R with zero-shot inference, 8-shot ICL and representative trainable baselines. Specifically, LoRA (Hu et al., 2021) adapts the attention projection matrices through low-rank updates. LIVE (Peng et al., 2024) and MimIC (Jiang et al., 2025) approximate demonstration effects through additive shifts while HiFICL (Li et al., 2026) replaces demonstrations with virtual key-value pairs. Table 1 shows that 8-shot ICL consistently improves over zero-shot inference with Idefics2. With LLaVA-Interleave, it improves performance on VQAv2, OK-VQA, and COCO, but remains below zero-shot inference on Flickr30K, MME, and SEED, highlighting that the benefits of explicit demonstrations vary across tasks and backbones. LoRA and LIVE fall below vanilla ICL on OK-VQA with Idefics2, while LoRA underperforms zero-shot inference on COCO with LLaVA-Interleave. In contrast, our method achieves the best results on every benchmark with Idefics2, leading both the average accuracy and average CIDEr scores. On LLaVA-Interleave, our ICL-R leads on VQAv2, OK-VQA, COCO and SEED and achieves the highest average accuracy and CIDEr, while ranking second on MME and Flickr30K. Overall, ICL-R delivers strong and consistent performance across diverse multimodal tasks while avoiding the inference-time overhead of retaining explicit demonstrations. Additional results on other model backbones and task-specific analyses are provided in Appendices A.8 and A.13.

Table 1: Results on six benchmarks with Average denoting the mean of accuracy or of CIDEr. Bold and underlined numbers denote the best and second-best results.

Model	Method	Trainable	Accuracy ↑					CIDEr ↑		
			VQAv2	OKVQA	MME	SEED	Average	COCO	Flickr30K	Average
Idefics2	Zero-shot	×	55.39	43.08	74.82	44.77	54.52	0.6828	0.5327	0.6078
	8-shot ICL	×	68.25	57.68	78.17	48.34	63.11	1.1923	<u>0.8457</u>	1.0190
	LoRA (Hu et al., 2021)	✓	69.14	55.05	79.26	<u>63.14</u>	66.65	1.2665	0.8054	1.0360
	LIVE (Peng et al., 2024)	✓	67.60	54.86	<u>85.52</u>	61.71	67.42	1.2764	0.8280	<u>1.0522</u>
	MimIC (Jiang et al., 2025)	✓	69.29	58.74	82.61	60.80	<u>67.86</u>	1.2827	0.7879	1.0353
	HiFICL (Li et al., 2026)	✓	<u>72.08</u>	<u>59.56</u>	77.07	46.12	63.71	<u>1.2951</u>	0.7831	1.0391
	ICL-R (Ours)	✓	**73.38**	**62.49**	**86.83**	**63.95**	**71.66**	**1.3757**	**0.8982**	**1.1370**
LLaVA-Interleave	Zero-shot	×	13.02	5.10	81.88	70.07	42.52	1.1516	0.7879	0.9698
	8-shot ICL	×	68.19	43.84	79.48	62.67	63.55	1.2085	0.7411	0.9748
	LoRA (Hu et al., 2021)	✓	70.12	48.19	**85.30**	72.29	68.98	1.0665	0.7529	0.9097
	LIVE (Peng et al., 2024)	✓	74.17	51.77	84.35	<u>72.55</u>	70.71	1.2743	0.8815	1.0779
	MimIC (Jiang et al., 2025)	✓	74.40	52.29	84.43	71.26	70.60	1.3169	**0.9137**	<u>1.1153</u>
	HiFICL (Li et al., 2026)	✓	<u>74.66</u>	<u>54.19</u>	84.64	71.03	<u>71.13</u>	<u>1.3315</u>	0.7064	1.0190
	ICL-R (Ours)	✓	**76.59**	**57.38**	<u>84.72</u>	**72.75**	**72.86**	**1.3386**	<u>0.8926</u>	**1.1156**

Attention value effect analysis. To determine whether attention and value calibration recover complementary components of the effects of demonstrations, we conduct a head-wise recovery analysis, with the results shown in Figure 2. All three metrics calculate cosine similarity between changes relative to the frozen model without demonstrations. M_A compares the attention redistribution produced by attention calibration with the demonstration-induced redistribution in the teacher. M_V compares the attention-output changes attributable to value calibration with the teacher's query-side value shifts aggregated under the frozen model's attention weights. M_{AV} compares the total attention-output changes produced by joint calibration with the teacher's full demonstration-induced changes. The left panel of Figure 2 shows highly head-specific recovery, with its directional alignment varying substantially across heads. The middle panel shows weaker recovery at individual heads but a broader distribution across layers. The right panel shows more consistent positive alignment across layers and heads, indicating that the two mechanisms capture complementary aspects of the teacher's computation. Together, these results support jointly reconstructing information selection and content propagation to better recover demonstration effects within attention. Appendix A.5 provides an additional layer-wise and question-level analysis.

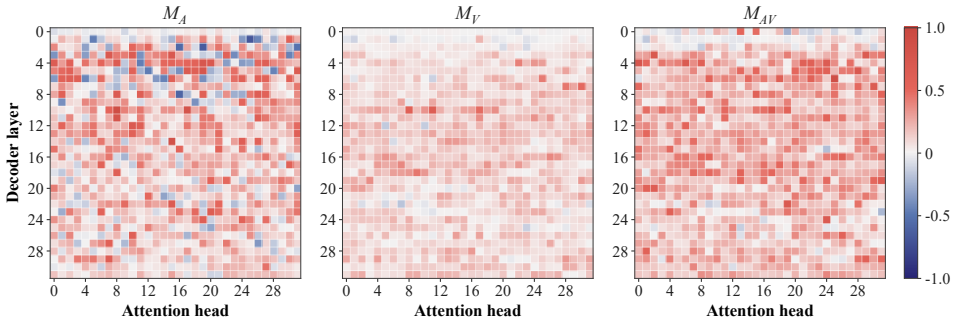


Figure 2: Head-wise cosine similarities between the student and teacher for attention routing, value content, and joint output recovery. Joint calibration produces more consistent positive alignment, demonstrating the complementary roles of attention and value calibration.

Core component ablations. To assess the contributions of our key designs, we compare our method with routing-only and static calibration variants in Figure 3. Routing only removes value calibration and allocates the full rank budget to attention routing, keeping the total rank equal to ICL-R for a fair comparison. Static calibration ablates query-adaptive modulation while retaining joint attention value calibration. Compared with routing-only calibration, ICL-R improves accuracy by 5.62 percentage points on OK-VQA from 56.87 to 62.49 and 4.44 points on VQAv2 from 68.94 to 73.38, while increasing COCO Caption CIDEr by 0.0358 from 1.3399 to 1.3757. Our method consistently outperforms the routing-only variant, indicating that increasing routing capacity alone cannot compensate for the absence of value calibration. Its advantage over static calibration further suggests that query-adaptive modulation provides useful flexibility beyond fixed correction.

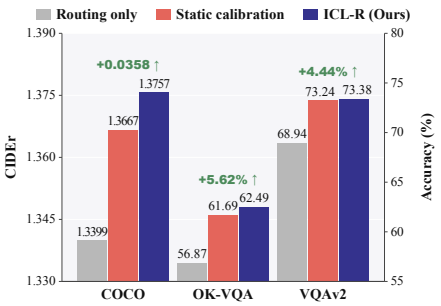


Figure 3: Ablation of value calibration and query-adaptive modulation in ICL-R on Idefics2 across three benchmarks.

Afterwards, to assess the effectiveness of Riemannian state transition objective, we compare our method with variants using cross-entropy alone or augmented with Euclidean ℓ_2 alignment, the point loss, or the fiber loss in Table 2. Euclidean alignment underperforms the CE baseline, suggesting that directly matching hidden states does not necessarily provide useful supervision for transferring demonstration effects. The point and fiber losses yield only small and mixed changes when applied separately, whereas their combination achieves the best performance. This contrast suggests that the incoming state and the transition provide complementary supervision.

Table 2: Ablation of the Riemannian state transition objective and its components.

Dataset	CE	$+\ell_2$	+Point	+Fiber	**ICL-R(Ours)**
OK-VQA	61.64	60.93	61.66	61.77	**62.49**
VQAv2	73.20	73.19	73.31	73.16	**73.38**

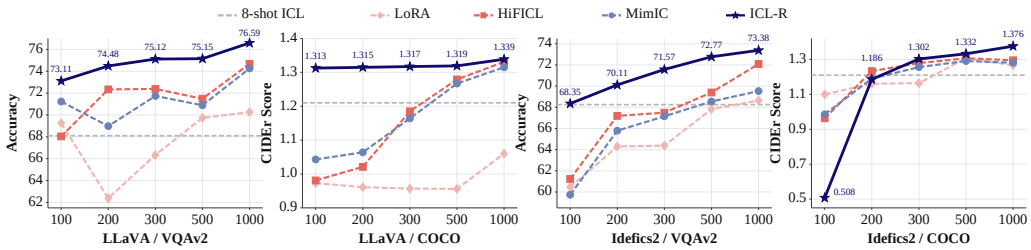


Figure 4: Performance under different training set sizes on VQAv2 and COCO Caption. ICL-R generally provides stronger data efficiency and more consistent scaling than the compared methods.

Data efficiency analysis. Having established the effectiveness of its individual components, we next assess the data efficiency and scaling behavior of our method by varying the training set size from 100 to 1,000 instances and comparing it with LoRA, MimIC, and HiFICL in Figure 4. Our method achieves the strongest performance across most training sizes on both VQAv2 and COCO settings with Idefics2 and LLaVA-Interleave. In particular, its performance on LLaVA-Interleave remains high even with only 100 examples, whereas the competing methods require substantially more data to approach the same level. On COCO with Idefics2, ICL-R is less effective in the extreme 100-example regime but improves rapidly as the training set grows and outperforms all baselines from 300 examples onward. Overall, these results show that our method can reconstruct useful demonstration effects from limited supervision while continuing to benefit from larger training sets. This pattern holds across tasks and model backbones, showing that ICL-R remains effective across data regimes and that its data efficiency is not limited to a particular benchmark or model architecture.

Attention-value rank allocation. To investigate how capacity allocation between attention and value calibration affects performance, we vary their respective ranks while fixing the total rank at $r_A + r_V = 16$ to keep the total number of trainable parameters unchanged. Table 3 reports results on OK-VQA and VQAv2 with LLaVA-Interleave and the reported deltas are measured relative to the balanced 8/8 allocation. The best allocation differs across tasks, with 14/2 achieving the highest accuracy on OK-VQA and 4/12 performing best on VQAv2, while gains over 8/8 remain small. Under the fixed parameter budget, these results suggest that the preferred capacity allocation depends on the task, with the balanced configuration providing competitive performance across both benchmarks.

Table 3: Rank allocation between attention and value calibration under a fixed total rank with Δ denoting the change in percentage points relative to the 8/8 allocation.

r_A/r_V	OK-VQA	Δ	VQAv2	Δ
2/14	56.76	-0.62	76.538	-0.047
4/12	57.40	$+0.02$	**76.644**	$+0.059$
8/8	57.38	–	76.585	–
12/4	56.18	-1.20	76.359	-0.226
14/2	**57.50**	$+0.12$	76.367	-0.218

Total rank sensitivity and gating behavior. To study calibration capacity, we vary rank configurations of 1/1, 2/2, 4/4, and 8/8 on Idefics2 with OK-VQA and COCO in Figure 5. Increasing rank generally improves performance: OK-VQA and COCO continue to improve with relatively weak gains from 4/4 to 8/8. To determine whether these gains merely reflect larger corrections, we analyze the learned static output scales. Each branch yields a 32×32 matrix of s_A or s_V, with one entry per layer–head pair. The center heatmaps show normalized task-averaged $\tanh(s_A)$ and $\tanh(s_V)$ at rank 4/4, revealing distinct layer–head structures, while the right panels show their absolute-value distributions across tasks and rank settings. Median magnitudes remain stable or decrease slightly with rank, indicating that additional rank enriches the low-rank representation and improves calibration capacity without increasing static gate scales.

Effect of the number of demonstrations. We further examine the robustness of our method to the amount of demonstration context used during training. We vary the number of demonstrations from 4 to 48 and report the results in Figure 6 (left). On VQAv2, explicit ICL degrades sharply beyond 16 demonstrations, whereas our method varies by only 0.125 points across all settings and exceeds explicit ICL by more than 22 points at 32 demonstrations. On COCO Caption, explicit ICL also fluctuates substantially and declines at 48 demonstrations. In contrast, ICL-R remains within a narrow range of 0.011 CIDEr, with its best result obtained using eight demonstrations. These small variations suggest that ICL-R remains robust to the growing number of redundant visual tokens introduced by additional demonstrations. Thus, ICL-R preserves the benefits of demonstrations across context sizes, even when longer contexts degrade explicit ICL.

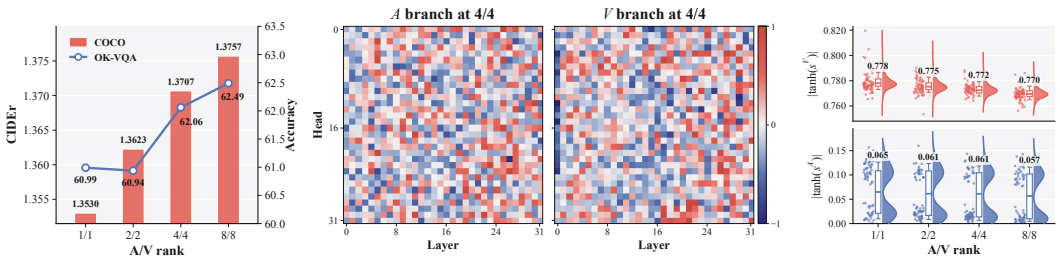


Figure 5: Rank sensitivity and learned calibration scales on Idefics2. Left: COCO CIDEr and OK-VQA accuracy across ranks. Center: normalized attention and value scales at rank 4/4. Right: distributions of their absolute magnitudes across both tasks and rank configurations.

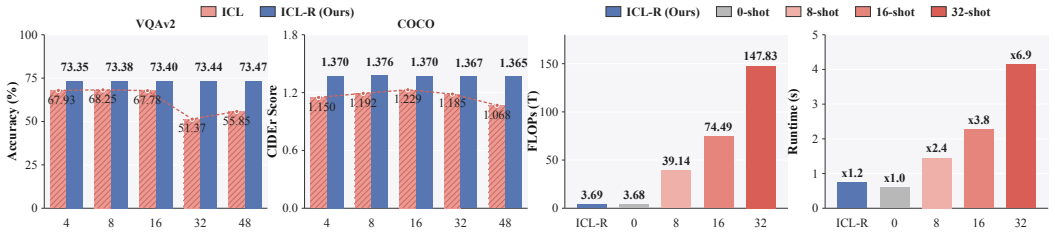


Figure 6: Performance and inference efficiency comparisons. Left: VQAv2 accuracy and COCO CIDEr for ICL-R and explicit ICL with varying demonstration counts. Right: FLOPs and runtime of ICL-R, zero-shot inference, and explicit ICL with different context sizes.

Inference efficiency. We evaluate inference efficiency on VQAv2 with Idefics2-8B-Base using an NVIDIA RTX A6000 GPU, comparing the FLOPs and runtime of ICL-R with zero-shot inference and explicit ICL at different demonstration counts in Figure 6 (right). ICL-R remains close to zero-shot inference in computational cost, with approximately 98% fewer FLOPs and 82% lower runtime than 32-shot ICL. The runtime reduction confirms that the savings in arithmetic operations translate into a substantial practical speedup in our evaluation setting. Overall, these results show that ICL-R combines stable performance across demonstration counts with inference cost close to that of zero-shot inference, offering an effective balance between accuracy and efficiency.

Hallucination analysis. To assess whether ICL-R reduces object hallucination while preserving coverage, we compare CHAIR (Rohrbach et al., 2018) scores and object recall with zero-shot inference, explicit ICL, and trainable baselines in Table 4. ICL-R achieves the lowest instance-level CHAIR of 2.0 and the highest object recall of 46.0 among all compared methods. Compared with 8-shot ICL, it reduces the instance-level score from 2.3 to 2.0 while maintaining the same sentence-level score of 3.1 and increasing recall from 40.3 to 46.0. These results demonstrate that ICL-R provides a favorable balance between reducing object hallucination and preserving object coverage.

Table 4: Hallucination performance measured by CHAIR and object recall on Idefics2. Our method achieves the lowest CHAIR_I and highest object recall among compared approaches.

	0-shot	4-shot	8-shot	16-shot	32-shot	LoRA	LIVE	MimIC	HiFICL	**ICL-R(Ours)**
$\text{CHAIR}_\text{S} \downarrow$	**2.8**	3.4	3.1	3.2	3.9	3.3	4.4	4.0	3.2	3.1
$\text{CHAIR}_\text{I} \downarrow$	3.3	2.5	2.3	2.4	2.4	2.4	3.1	2.9	2.2	**2.0**
Recall $\uparrow$	27.2	41.3	40.3	41.3	42.1	44.3	44.8	45.4	45.7	**46.0**

5 Conclusion

In this work, we introduced ICL-R, a lightweight framework for reconstructing demonstration-induced computation in frozen MLLMs. The framework jointly models attention routing and value transformations, while using internal state transitions to provide supervision beyond endpoint matching. By preserving the benefits of few-shot prompting while reducing context length and inference latency, ICL-R offers a practical route to efficient multimodal adaptation. Future work will assess its generality across a broader range of models and datasets and investigate how the framework can be extended to video understanding to reconstruct demonstration effects on temporal reasoning.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, LaTeX formatting, figure-layout code, and the organization across the paper. The research ideas, method design, experimental data, and quantitative results were produced and verified by the authors. All AI-assisted content was manually reviewed and revised, and the authors take full responsibility for the final text, claims, equations, citations, and reported results.

REPRODUCIBILITY STATEMENT

We specify the complete formulation, model components, and training objective in Section 3 and report the evaluated backbones, datasets, metrics, trainable modules and optimization setup in Section 4. Appendix A.1 gives the complete derivation of the formulation, and Appendix A.3 shows the general procedure of the algorithm. Appendix A.11 describes the memory-efficient fused SDPA implementation. Overall, these details support independent reproduction of our work.

REFERENCES

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

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

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Honghao Fu, Yuan Ouyang, Kai-Wei Chang, Yiwei Wang, Zi Huang, and Yujun Cai. Contextnav: Towards agentic multimodal in-context learning. In *International Conference on Learning Representations*, volume 2026, pp. 74934–74959, 2026b.

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

Akash Gupta, Amos Storkey, and Mirella Lapata. Meta-adaptive prompt distillation for few-shot visual question answering. In *International Conference on Learning Representations*, volume 2026, pp. 120800–120837, 2026.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yuchu Jiang, Jiale Fu, Chenduo Hao, Xinting Hu, Yingzhe Peng, Xin Geng, and Xu Yang. Mimic in-context learning for multimodal tasks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29825–29835. IEEE, 2025.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024a.

Jiaqian Li, Yanshu Li, Ligong Han, Ruixiang Tang, and Wenya Wang. Train once, reuse everywhere: Generalizable implicit in-context learning by routing attention. *arXiv preprint arXiv:2509.22854*, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023b.

Li Li, Jiawei Peng, Huiyi Chen, Chongyang Gao, and Xu Yang. How to configure good in-context sequence for visual question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26700–26710. IEEE, 2024b.

Xiaoyu Li, Yuhang Liu, Xuanshuo Kang, Zheng Luo, Fangqi Lou, Xiaohua Wu, and Zihan Xiong. Hificl: High-fidelity in-context learning for multimodal tasks. *arXiv preprint arXiv:2603.12760*, 2026.

Yanshu Li, Yi Cao, Hongyang He, Qisen Cheng, Xiang Fu, Xi Xiao, Tianyang Wang, and Ruixiang Tang. M^2IV: Towards efficient and fine-grained multimodal in-context learning via representation engineering, 2025b.

Zhuowei Li, Zihao Xu, Ligong Han, Yunhe Gao, Song Wen, Di Liu, Hao Wang, and Dimitris Metaxas. Implicit in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 101644–101666, 2025c.

Ya-Wei Eileen Lin, Yuval Kluger, and Ronen Talmon. Hyperbolic procrustes analysis using riemannian geometry. *Advances in Neural Information Processing Systems*, 34:5959–5971, 2021.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. *arXiv preprint arXiv:2311.06668*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ziyu Ma, Chenhui Gou, Yiming Hu, Yong Wang, Bohan Zhuang, and Jianfei Cai. Where and what matters: Sensitivity-aware task vectors for many-shot multimodal in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7892–7900, 2026.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 3190–3199. IEEE, 2019.

Yingzhe Peng, Chenduo Hao, Xinting Hu, Jiawei Peng, Xin Geng, and Xu Yang. Live: Learnable in-context vector for visual question answering. *Advances in Neural Information Processing Systems*, 37:9773–9800, 2024.

Xavier Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25(1):127–154, 2006.

Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. What factors affect multimodal in-context learning? an in-depth exploration. *Advances in Neural Information Processing Systems*, 37:123207–123236, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14974–14983. IEEE, 2023.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. *arXiv preprint arXiv:2209.15189*, 2022.

Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. Exploring effective factors for improving visual in-context learning. *IEEE Transactions on Image Processing*, 2025.

Niloufar Alipour Talemi, Hossein Kashiani, and Fatemeh Afghah. Hyper-icl: Attention calibration with hyperbolic anchor distillation for multimodal in-context learning. *arXiv preprint arXiv:2606.04434*, 2026.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In *International conference on learning representations*, volume 2024, pp. 17282–17333, 2024.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

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

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 25966–26003, 2026.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 7130–7138. IEEE, 2017.

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

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

Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. Vl-icl bench: The devil in the details of multimodal in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 100058–100100, 2025.

APPENDIX CONTENTS

A Appendix **14**

A.1 Complete Derivation of the Formulation 14

A.2 Implementation Details 15

A.3 Algorithmic Procedure of ICL-R 16

A.4 Decomposing Demonstration-Induced Attention-Output Shifts 16

A.4.1 Relative Magnitudes of the Four Decomposition Terms 17

A.4.2 Layer-wise Dynamics and Alignment with the Total Shift 18

A.4.3 Term Coupling and Missing Attention Mass 18

A.5 Layer-Wise and Question-level Recovery Analysis 19

A.6 Additional Gating Ablations 19

A.7 Geometric Analysis of RST 20

A.7.1 Complementarity of Point and Fiber Supervision 20

A.7.2 Geometric Alignment on Unseen Queries 21

A.7.3 Geometric Consistency of Parallel Transport 22

A.8 Generalization to Other MLLMs 24

A.9 Dataset Transfer 25

A.10 Trainable Parameter Analysis 25

A.11 Single-Pass Fused SDPA Implementation 25

A.12 Peak GPU memory usage 26

A.13 Error Correction Across Categories and Backbones 27

A.14 Paired discrimination on MME 29

A.15 Residual failures and qualitative limitations. 31

A.16 Qualitative Examples Analysis 32

A Appendix

A.1 Complete Derivation of the Formulation

We provide the complete derivation of our decomposition in Section 3.1. The derivation is exact for each self-attention head. Let n_X and n_D denote the numbers of query and demonstration tokens, respectively, as in the main text. Then $A_{T,D} \in \mathbb{R}^{n_X \times n_D}$, $A_{T,X}, A_0 \in \mathbb{R}^{n_X \times n_X}$, $V_{T,D} \in \mathbb{R}^{n_D \times d_h}$, and $V_{T,X}, V_0 \in \mathbb{R}^{n_X \times d_h}$.

Step 1: Partition the teacher output. For the query-aligned attention rows, the teacher attends to both demonstration and query tokens, whereas the query-only model attends only to query tokens. Their outputs are:

$$O_T = A_{T,D}V_{T,D} + A_{T,X}V_{T,X}, \qquad O_0 = A_0V_0. \tag{14}$$

Step 2: Express the query-side changes. Using the definitions from Section 3.1:

$$\begin{aligned} \Delta A_T = A_{T,X} - A_0 \quad &\Longleftrightarrow \quad A_{T,X} = A_0 + \Delta A_T, \\ \Delta V_T = V_{T,X} - V_0 \quad &\Longleftrightarrow \quad V_{T,X} = V_0 + \Delta V_T. \end{aligned} \tag{15}$$

Step 3: Expand the output difference. Substituting these identities into the teacher output gives:

$$\begin{aligned} O_T - O_0 &= A_{T,D}V_{T,D} + (A_0 + \Delta A_T)(V_0 + \Delta V_T) - A_0V_0 \\ &= A_{T,D}V_{T,D} + A_0V_0 + A_0\Delta V_T + \Delta A_T V_0 + \Delta A_T \Delta V_T - A_0V_0. \end{aligned} \tag{16}$$

After canceling the two A_0V_0 terms and arranging the remaining terms by their source, we obtain:

$$O_T - O_0 = \underbrace{\Delta A_T V_0}_{\text{routing shift}} + \underbrace{A_0 \Delta V_T}_{\text{value shift}} + \underbrace{\Delta A_T \Delta V_T}_{\text{interaction}} + \underbrace{A_{T,D}V_{T,D}}_{\text{demonstration information}}. \tag{17}$$

The first term changes attention routing while holding the base values fixed. The second changes value content under the base routing. The third is the coupled effect that appears only when routing and values change together. The final term is information read directly from demonstration tokens. Because the query-only student cannot form this final term explicitly, its effect must be absorbed into the calibrated query-side output.

Step 4: Derive the missing-mass factorization. For query token t, define the teacher attention mass assigned to demonstrations as:

$$m_t^T = \sum_{s \in \mathcal{D}} (A_{T,D})_{t,s}, \qquad \sum_{s \in X} (A_{T,X})_{t,s} = 1 - m_t^T. \tag{18}$$

For $m_t^T \in (0, 1)$, normalize the teacher attention within the query block as

$$(\overline{A}_{T,X})_{t,s} = \frac{(A_{T,X})_{t,s}}{1 - m_t^T}, \qquad s \in X, \tag{19}$$

and define the attention-weighted demonstration information conditioned on attending to demonstrations as:

$$E_t^T = \frac{1}{m_t^T} \sum_{s \in \mathcal{D}} (A_{T,D})_{t,s} (V_{T,D})_s. \tag{20}$$

The teacher output for token t can then be rewritten exactly as:

$$\begin{aligned} O_{T,t} &= \sum_{s \in X} (A_{T,X})_{t,s} (V_{T,X})_s + \sum_{s \in \mathcal{D}} (A_{T,D})_{t,s} (V_{T,D})_s \\ &= (1 - m_t^T) \sum_{s \in X} (\overline{A}_{T,X})_{t,s} (V_{T,X})_s + m_t^T E_t^T. \end{aligned} \tag{21}$$

The first term is the normalized query-side aggregation multiplied by its true total mass, while the second is the conditional content obtained from demonstration tokens. A query-only softmax instead forces its attention mass over query tokens to sum to one and therefore omits the coefficient $1 - m_t^T$. The missing-mass estimator introduces a learned coefficient motivated by this factor, and the joint attention and value calibration reconstructs the routing and content needed to approximate the full teacher output. The limiting cases $m_t^T = 0$ and $m_t^T = 1$ follow directly by removing the absent component. For multi-head attention, the same factorization applies independently to each head, and the output projection preserves the equality by linearity.

A.2 Implementation Details

Training and evaluation protocol. For VQAv2, OK-VQA, COCO Caption, and Flickr30K, we randomly sample 1,000 instances from the corresponding training split to form the support/training set. Because MME and SEED-Bench do not provide an equivalent training split for our setting, we instead construct disjoint support and evaluation partitions from their released data. Each training episode contains eight demonstrations sampled from the support set and an independent query. The frozen teacher processes the demonstrations together with the query, whereas the student receives only the query. At inference, all trainable methods receive the query alone; only vanilla ICL retains eight demonstrations in the input context.

All parameters of the vision encoder, multimodal projector, and language model remain frozen. Only our proposed modules are optimized using the task loss and our Riemannian State Transition (RST) loss. By default, the ranks of the attention and value calibration modules are both set to 8. We use 16-bit mixed precision (`16-mixed`) and AdamW (Loshchilov & Hutter, 2017), with cosine learning-rate decay and linear warmup over the first 10% of the optimization steps.

We use 10,000 examples from the VQAv2 validation set and the complete validation splits of OK-VQA (5,046 examples), COCO Caption under the Karpathy split (Karpathy & Fei-Fei, 2015) (5,000 examples), and Flickr30K under the Karpathy split (1,014 examples). For MME, we deterministically use the first 1,000 question–answer pairs (500 images) as support and remaining 1,374 pairs (687 images) for evaluation; the two partitions have no image overlap. For the image subset of SEED-Bench, we use the first 1,000 image questions as support and a fixed sample of 10,000 questions from the remaining image questions for evaluation. The exact partitions and metrics are summarized in Table 5.

Dataset	K	N_{eval}	Metric
VQAv2	1,000	10,000	VQA Acc.
OK-VQA	1,000	5,046	VQA Acc.
COCO Caption	1,000	5,000	CIDEr
MME	1,000	1,374	Accuracy
SEED-Bench	1,000	10,000	Exact match
Flickr30K	1,000	1,014	CIDEr

Table 5: Training and evaluation partitions. K is the size of the training set, and N_{eval} is the number of evaluated queries.

For VQAv2 and OK-VQA, the first annotated answer is used as the training target, while evaluation uses all human-annotated answers and the standard VQA accuracy. For COCO Caption and Flickr30K, the first reference caption is used as the training target, while evaluation computes CIDEr against all five references. MME is evaluated using question-level yes/no accuracy and SEED-Bench is evaluated by exact matching of the predicted option in $\{\text{A, B, C, D}\}$.

Benchmarks. VQAv2 (Goyal et al., 2017) evaluates general visual question answering on MS COCO images, whereas OK-VQA (Marino et al., 2019) additionally requires external knowledge beyond the visible image. COCO Caption (Chen et al., 2015) and Flickr30K (Young et al., 2014) evaluate open-ended image description. MME (Fu et al., 2026a) covers a broad collection of perception and cognition abilities using paired yes/no questions, and SEED-Bench (Li et al., 2023a) evaluates multimodal perception and reasoning in a multiple-choice format. We report VQA accuracy for VQAv2 and OK-VQA, question-level accuracy for MME, exact match for SEED-Bench, and CIDEr (Vedantam et al., 2015) for COCO Caption and Flickr30K.

Numerical treatment of RST. We evaluate the geometric operations in float32 with a numerical tolerance of $\varepsilon = 10^{-6}$. State normalization is implemented as $\pi_\varepsilon(z) = z/\max(\|z\|_2, \varepsilon)$. For the spherical logarithmic map, we compute $c = \text{clip}(\langle x, y\rangle, -1+\varepsilon, 1)$, form $u = y - cx$, and obtain the angle using $\theta = \text{atan2}(\|u\|_2, c)$. When $\|u\|_2 \le \varepsilon$, the scaling factor is set to one; otherwise, it is $\theta/\|u\|_2$. Near coincident points, this implements the first-order approximation $\text{Log}_x(y) \approx u$, recovering the zero transition at coincidence. For parallel transport, we lower-bound $1 + \langle x^T, x^S\rangle$ by ε. At exactly antipodal pairs, the spherical logarithmic map and shortest-geodesic parallel transport are not uniquely defined. When the clipping or denominator bound is activated near these singularities,

the resulting quantities serve as numerical surrogates. The half-chord argument in the point-distance calculation is capped at one before $\arcsin$.

A.3 Algorithmic Procedure of ICL-R

To clarify how the proposed components are integrated into a complete workflow, we provide pseudocode for the training and inference procedures below. Algorithm 1 summarizes the overall training and inference procedure of ICL-R. During training, the frozen teacher and query-only student are evaluated under teacher forcing, after which the student is optimized by combining task supervision with Riemannian State Transition (RST) alignment while the backbone remains fixed; at inference, the trained student generates answers from the query alone. Algorithm 2 further details the student attention computation, including query-adaptive gating, low-rank corrections to attention logits and values, missing-mass estimation, and aggregation with the resulting calibrated attention weights.

Algorithm 1 Training and inference procedure of ICL-R.

Require: Frozen backbone parameters θ; training episodes containing demonstrations $\mathcal{D}$, query X, and target answer Y; loss weight λ
Ensure: Trained calibration parameters ϕ

Freeze θ and initialize ϕ, including the attention and value calibration modules and the missing-mass estimator
for all training episodes $(\mathcal{D}, X, Y)$ **do**
  Run the frozen teacher on $[\mathcal{D}; X]$ with teacher forcing using Y; record h_i^T and o_i^T
  Run the student on X with teacher forcing using Y and CalibratedAttention$(Q_0, K_0, V_0; \phi)$; record h_i^S and o_i^S
  for all aligned token–layer indices $i \in \mathcal{I}$ **do**
    for $a \in \{T, S\}$ **do**
      $x_i^a \leftarrow \pi(h_i^a)$ and $y_i^a \leftarrow \pi(h_i^a + o_i^a)$
      $v_i^a \leftarrow \mathrm{Log}_{x_i^a}(y_i^a)$
    end for
    $\widetilde{v}_i^T \leftarrow \mathrm{PT}_{x_i^T \to x_i^S}(v_i^T)$
  end for
  $\mathcal{L}_{\mathrm{RST}} \leftarrow \mathcal{L}_{\mathrm{point}} + \mathcal{L}_{\mathrm{fiber}}$
  $\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{CE}}(Y) + \lambda \mathcal{L}_{\mathrm{RST}}$
  Update only ϕ to minimize $\mathcal{L}_{\mathrm{total}}$; keep θ frozen
end for

Inference: generate the answer with the trained student using only query X

Algorithm 2 Jointly calibrated attention used by the ICL-R student.

function CalibratedAttention$(Q_0, K_0, V_0; \phi)$
  for all query tokens t **do**
    Compute $\bar{q}_t = \mathrm{LN}(q_t)$ and predict $g_t^A, \rho_t^A, g_t^V, \rho_t^V$ from the current query
    $\Delta S_{t,s}^A \leftarrow g_t^A \left[\rho_t^A \odot (Q_0 U_Q)_{t,:}\right] (K_0 U_K)_{s,:}^\top$ for all allowed keys s
    $\Delta V_{t,s} \leftarrow g_t^V \left[\left(\rho_t^V \odot V_{0,s} P_V\right) R_V\right]$ for all allowed keys s
    Predict $m_t \leftarrow s_m \sigma\left(\bar{q}_t^\top w_m / \sqrt{d_h} + b_m\right)$
    $\widehat{A}_{t,:} \leftarrow (1 - m_t)\,\mathrm{softmax}\left((Q_0 K_0^\top / \sqrt{d_h})_{t,:} + \Delta S_{t,:}^A\right)$
    $\widetilde{O}_t \leftarrow \sum_s \widehat{A}_{t,s}\,(V_{0,s} + \Delta V_{t,s})$
  end for
  return $\widetilde{O}$
end function

A.4 Decomposing Demonstration-Induced Attention-Output Shifts

To understand what internal behavior ICL-R needs to recover beyond task-level predictions, we analyze how in-context demonstrations change the attention outputs of the teacher model. Specifically,

we compare the attention output produced with demonstrations O_T against that of the query-only model O_0 on 64 teacher-forced VQAv2 samples with eight demonstrations. After aligning the query tokens in the two input sequences, we exactly decompose the resulting output shift into routing, value, interaction, and direct-demonstration terms. We then investigate three complementary questions: how large each term is across layers and heads, whether each term aligns with or counteracts the total output shift, and how interactions between the terms relate to the attention probability mass assigned to demonstrations. This analysis provides a quantitative characterization of the internal effects that must be approximated when demonstrations are removed at inference time.

A.4.1 Relative Magnitudes of the Four Decomposition Terms

We first examine whether demonstration-induced attention-output shifts are dominated by a single mechanism such as attention rerouting. Let $\Delta A = A_{T,X} - A_0$ and $\Delta V = V_{T,X} - V_0$, where X denotes the aligned query tokens and D denotes the demonstration-only tokens. The total shift admits the exact decomposition:

$$O_T - O_0 = \underbrace{\Delta A V_0}_{T_R} + \underbrace{A_0 \Delta V}_{T_V} + \underbrace{\Delta A \Delta V}_{T_I} + \underbrace{A_{T,D} V_{T,D}}_{T_D}, \qquad \rho_k = \frac{\|T_k\|_F}{\sum_j \|T_j\|_F}. \tag{22}$$

Here, T_R, T_V, T_I, and T_D denote the routing, value, interaction, and direct-demonstration terms, respectively, and ρ_k measures the normalized relative magnitude of each term for each layer and attention head. As shown in Figure 7, routing has the largest average relative magnitude at 29.2%, followed by value at 25.7%, direct demonstration retrieval at 25.1%, and interaction at 20.0%. The decomposition is numerically exact, with a mean relative reconstruction residual of 2.09×10^{-7} and a maximum residual of 9.81×10^{-7}. Although routing is the largest individual term, it accounts for less than one third of the total normalized magnitude, while the other three terms remain substantial and exhibit considerable layer- and head-level heterogeneity. These results show that the effect of demonstrations cannot be reduced to attention reweighting alone: demonstrations simultaneously alter information routing, value representations, their interaction, and the direct evidence available to the model.

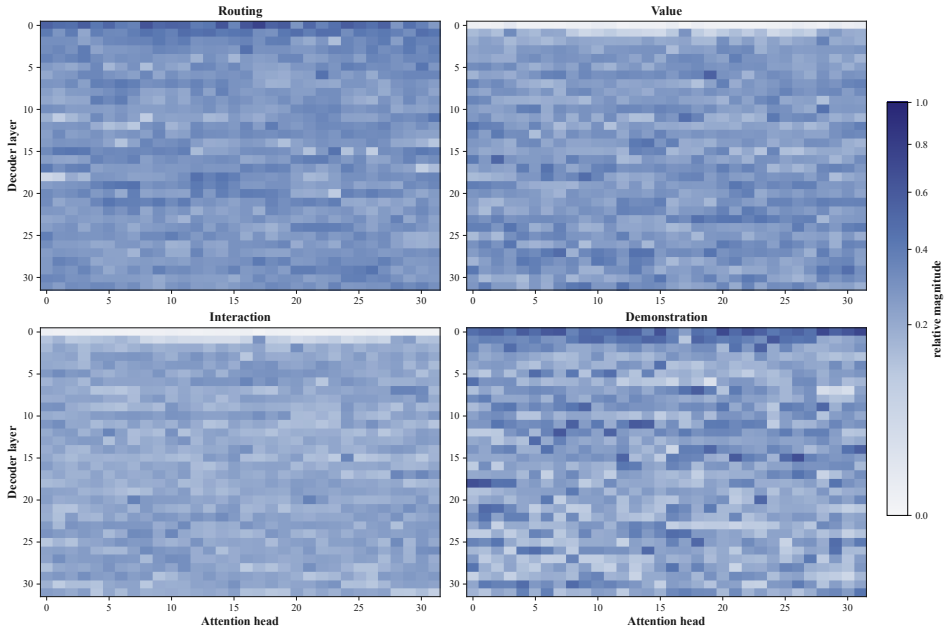


Figure 7: Relative magnitudes of the four terms in the demonstration-induced attention-output decomposition across model layers and attention heads.

A.4.2 Layer-wise Dynamics and Alignment with the Total Shift

We next study how the four mechanisms evolve with model depth and distinguish term magnitude from directional relevance. In addition to the layer-wise average relative magnitude $\rho_{k,l}$, we calculate the cosine similarity between each term and the total attention-output shift,

$$c_{k,l} = \mathbb{E}_h \left[\frac{\langle T_{k,l,h},\, O_{T,l,h} - O_{0,l,h} \rangle}{\|T_{k,l,h}\|_F \|O_{T,l,h} - O_{0,l,h}\|_F} \right]. \tag{23}$$

A positive cosine indicates that a term generally moves the output in the direction of the total teacher-induced shift, whereas a negative cosine indicates a compensatory or cancelling effect. Figure 8 reveals a clear depth-dependent transition. From early to late layers, the relative magnitude of routing decreases from 31.9% to 27.8%, while that of value increases from 21.8% to 28.4%. The direct-demonstration term decreases from 28.2% to 22.1%, whereas the interaction term increases from 18.1% to 21.7%. Directionally, routing is most strongly aligned with the total shift, with an average cosine of 0.574, followed by the direct-demonstration and value terms at 0.477 and 0.437 respectively. In contrast, the interaction term has a substantial magnitude but a negative average cosine of -0.158. These results suggest that early layers primarily modify attention routing and retrieve demonstration information, whereas deeper layers increasingly transform the representations being aggregated. They also show that a large term is not necessarily a direct positive contributor for the interaction term acts predominantly as a compensatory correction to the other components.

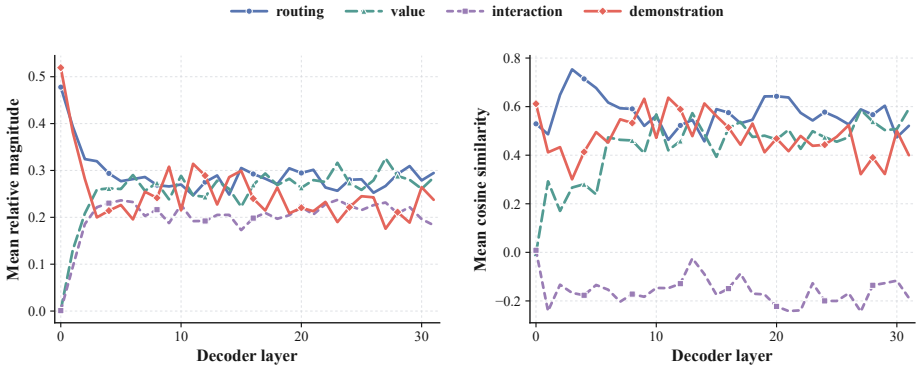


Figure 8: Layer-wise dynamics of the four decomposition terms and their alignment with the total demonstration-induced attention-output shift.

A.4.3 Term Coupling and Missing Attention Mass

Finally, we investigate whether the four decomposition terms operate independently and quantify the attention normalization mass that disappears when demonstrations are removed. We compute the pairwise cosine similarity and the teacher attention mass assigned to demonstration tokens as:

$$C_{ij} = \frac{\langle T_i, T_j \rangle}{\|T_i\|_F \|T_j\|_F}, \qquad m^T_{l,h,i} = \sum_{j \in D} A^T_{l,h,i,j}. \tag{24}$$

Negative values of C_{ij} indicate systematic cancellation. We average $m^T_{l,h,i}$ over aligned query tokens and samples for each layer and head. Figure 9 shows that the strongest coupling occurs between the value and interaction terms, whose cosine similarity is -0.729. Routing is also negatively aligned with interaction and direct demonstration retrieval, with pairwise cosines of -0.250 and -0.198 respectively, while routing and value exhibit only a weak positive alignment of 0.175. This confirms that the four terms are strongly coupled and explains why their normalized magnitudes cannot be treated as independent additive contributions. Meanwhile, the teacher assigns an average attention mass of 0.299 to demonstrations. The missing mass is particularly large in early layers, decreasing from 0.535 in Layer 0 and 0.473 in Layer 1 to approximately 0.250 by Layer 7, with substantial variation across attention heads. Thus, removing demonstrations causes not only the loss of their value evidence but a considerable, layer-dependent change in attention normalization. This observation provides the mechanistic motivation for explicitly accounting for missing attention mass.

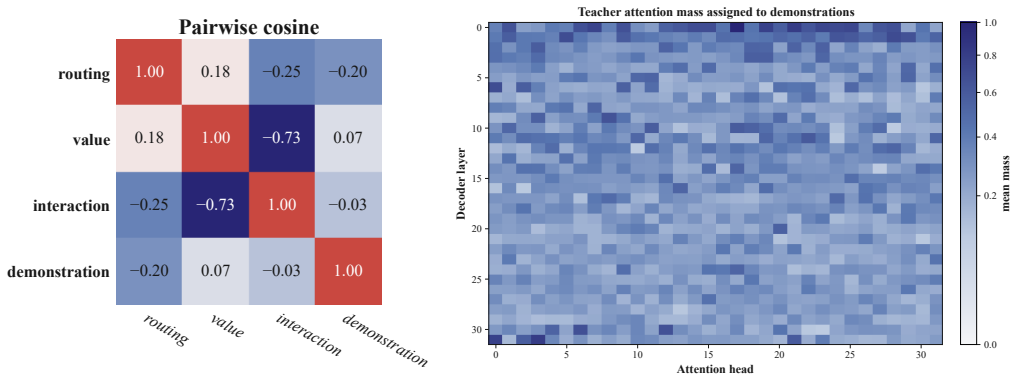


Figure 9: Term coupling and missing attention mass. Left: pairwise relationships among the four decomposition terms. Right: attention mass assigned to demonstration tokens.

A.5 Layer-Wise and Question-Level Recovery Analysis

To assess whether joint calibration improves recovery across model depth and individual questions, we compare attention-only calibration, value-only calibration, and joint calibration against the teacher output shift. For each variant, we subtract the output of the frozen query-only model, denoted by $O_{\mathrm{B}}^{(\ell)}$, to obtain $\Delta O_X^{(\ell)} = O_X^{(\ell)} - O_{\mathrm{B}}^{(\ell)}$. We measure recovery at layer ℓ using the cosine similarity between the vectorized student and teacher output residuals:

$$R_X^{(\ell)} = \cos\Big(\mathrm{vec}(\Delta O_X^{(\ell)}), \mathrm{vec}(\Delta O_{\mathrm{T}}^{(\ell)})\Big), \qquad X \in \{\mathrm{A}, \mathrm{V}, \mathrm{AV}\}. \tag{25}$$

Higher values indicate closer directional alignment with the teacher output shift. We also average recovery across layers for each question and compare the resulting distributions. The left panel of Figure 10 shows that recovery rises after the initial layers, remains relatively high through the middle layers, and generally weakens in later layers. The relative performance of A-only and V-only varies with depth, while joint calibration achieves higher recovery at many early and middle layers without dominating every layer. In the right panel, joint calibration shifts the distribution toward higher recovery, achieving a mean of 0.2000, compared with 0.1810 for A-only and 0.1756 for V-only. Together, these results support the complementary roles of attention routing and value content that their joint calibration improves overall alignment with the teacher output shift.

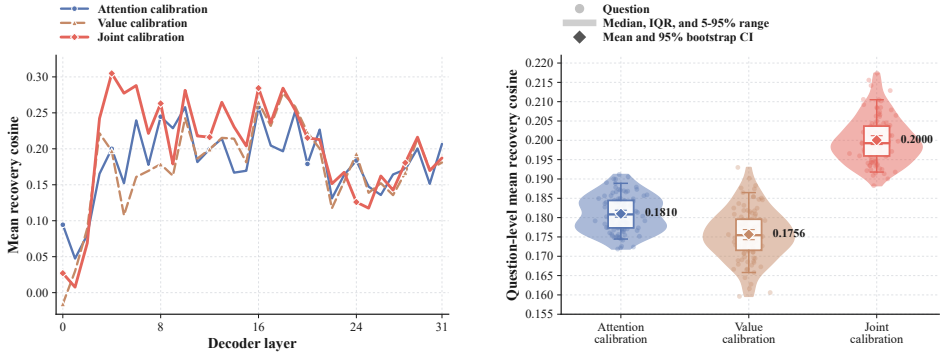


Figure 10: Recovery of the teacher induced output shift by attention only, value only, and joint attention and value calibration. Left: layer wise recovery trajectories. Right: distributions of question level mean recovery. Joint calibration provides the strongest and most stable recovery.

A.6 Additional Gating Ablations

To assess the contributions of query-adaptive gating and missing-mass correction, we ablate the gating components in Eq. 8 and the mass estimator in Eq. 6. We compare the full design with variants that separately remove the $\tanh$ scaling, sigmoid gating, channel gating, or missing-mass estimation, evaluating accuracy on OK-VQA and CIDEr on COCO Caption. As shown in Table 6,

the full design achieves the best results on both benchmarks, while removing the $\tanh$ scaling causes the largest performance drop. Removing the sigmoid or channel gating also lowers both scores, indicating that query-dependent correction strength and low-rank direction selection are beneficial. Without missing-mass estimation, OK-VQA accuracy decreases from 62.49 to 61.86 and COCO CIDEr decreases from 1.3757 to 1.3705, supporting the utility of the learned query-side scaling factor.

Table 6: Ablation of query-adaptive gating components and missing-mass estimation. Bold numbers denote the best scores.

Variant	OK-VQA	COCO CIDEr
ICL-R	**62.49**	**1.3757**
w/o $\tanh$ scaling	59.94	1.3589
w/o sigmoid gating	61.72	1.3671
w/o channel gating	61.86	1.3709
w/o missing mass	61.86	1.3705

A.7 Geometric Analysis of RST

A.7.1 Complementarity of Point and Fiber Supervision

We investigate whether the point and fiber terms provide complementary supervision for operator alignment. Using captured teacher states from VQAv2 and OKVQA as anchors, we construct three controlled settings that preserve the endpoint, preserve the base point, or preserve the transition under parallel transport. We compare the responses of endpoint matching, the individual point and fiber terms, their RST sum, and normalized state pair matching. As shown in Figure 11, endpoint matching remains insensitive to changes in the base and transition when the endpoint is fixed. Likewise, the point term cannot detect transition changes at a fixed base, whereas the fiber term remains near zero when the base moves but the transported transition remains equivalent. RST responds to the discrepancies left undetected by either component alone. These results support the complementary roles of point and fiber supervision: the former aligns the state from which an operator acts, while the latter aligns the update performed from that state.

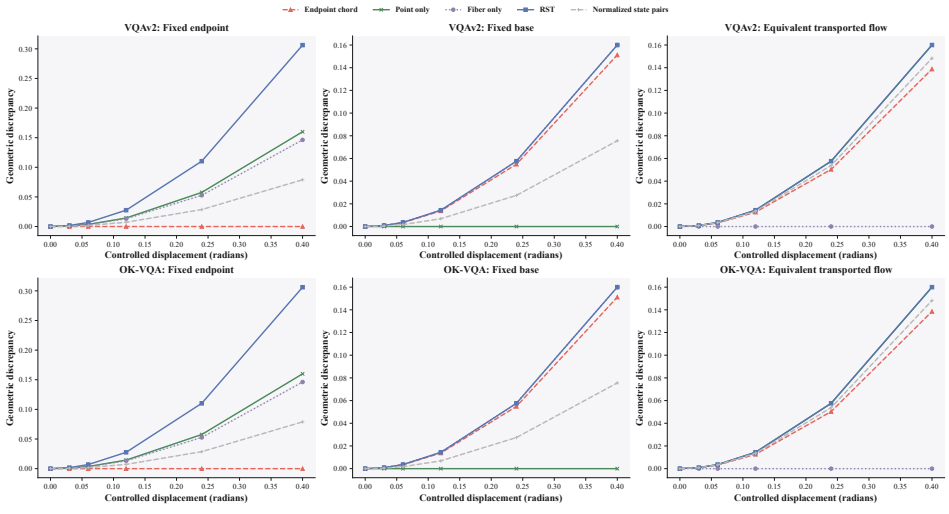


Figure 11: Responses of geometric objectives to controlled displacements anchored at teacher states from VQAv2 (top) and OK-VQA (bottom). Columns show a fixed endpoint, a fixed base point, and an equivalent transition under parallel transport. Curves compare endpoint matching, point and fiber supervision, their RST sum, and normalized state pair matching.

A.7.2 GEOMETRIC ALIGNMENT ON UNSEEN QUERIES

We next examine whether the geometry encouraged by RST is reflected in representations of unseen queries. We compare L2 and RST on 128 distinct images from each of VQAv2, OKVQA, MME and SEED using identical query inputs and shared teacher states. We measure point and transported fiber distances across all decoder layers and eight selected query prompt positions, including boundary and task markers. Figure 12 shows that RST shifts both distance distributions toward lower discrepancy, with improvements in their central regions and upper tails. Figure 13 localizes these improvements across layers and query positions, revealing broadly improved alignment alongside some localized regressions. Together, these observations provide evidence that the RST achieves better geometric fidelity on unseen queries in both their states and transitions.

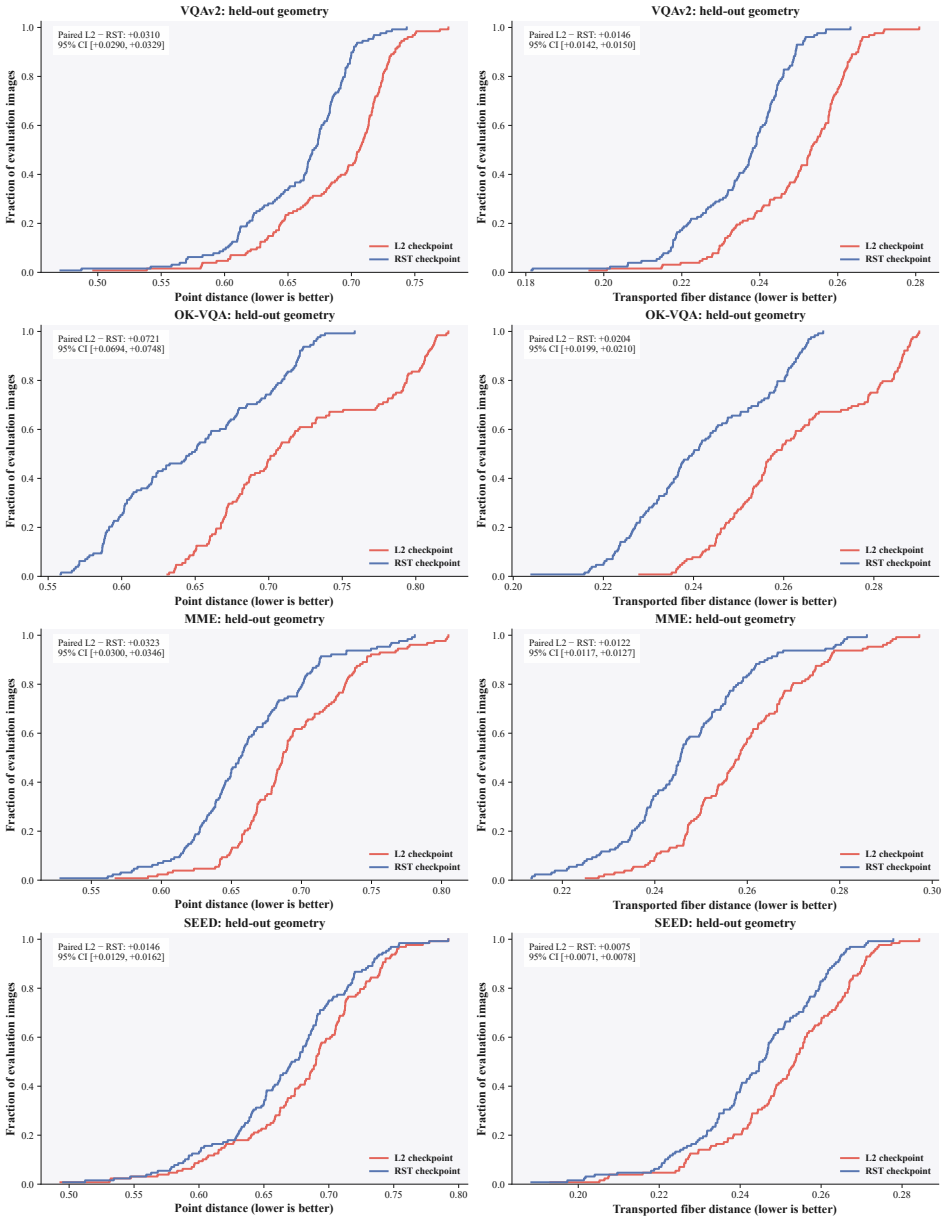


Figure 12: Cumulative distributions of point distance (left) and transported fiber distance (right) for L2 and RST checkpoints on unseen VQAv2, OK-VQA, MME and SEED images. Annotations report paired mean differences (L2 minus RST) and 95% confidence intervals over images. Lower distances indicate closer geometric alignment.

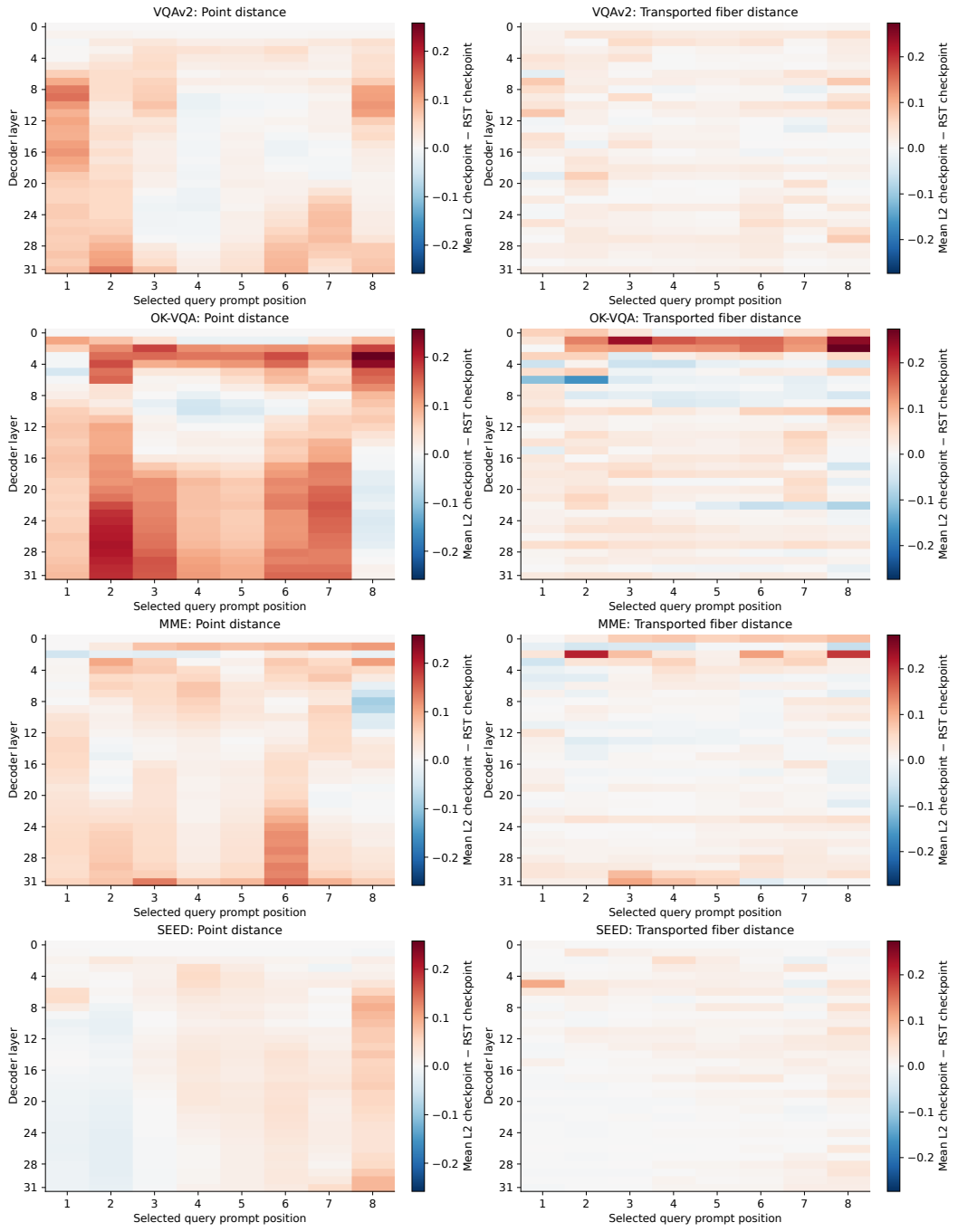


Figure 13: Mean geometric distance differences (L2 minus RST) across decoder layers and eight selected query prompt positions on VQAv2, OK-VQA, MME and SEED. The left and right columns show point and transported fiber distances, respectively. Positive values indicate closer alignment under RST, while negative values indicate localized regressions.

A.7.3 Geometric Consistency of Parallel Transport

To examine whether parallel transport prevents spurious discrepancies between geometrically equivalent transitions, we conduct controlled tests using states collected from existing Idefics2 checkpoints on 128 evaluation images from each of VQAv2 and OK-VQA. At four decoder depths, we

use the final query prompt token and move the teacher base point along either the observed direction toward the student base or the teacher transition direction. An independent plane rotation constructs the corresponding equivalent transition at each target point. We measure the discrepancy relative to this reference, normalized by the teacher transition norm, and additionally evaluate tangency residuals, norm preservation, and agreement with the rotation reference. Figure 14 shows that direct subtraction introduces discrepancies that increase with base displacement and vary substantially across directions. Parallel transport remains close to the reference in both settings, while the numerical consistency checks exhibit small residuals. These results support parallel transport as a geometrically consistent component of the RST fiber term, allowing equivalent transitions at different base points to be compared without introducing an additional penalty solely from their base displacement.

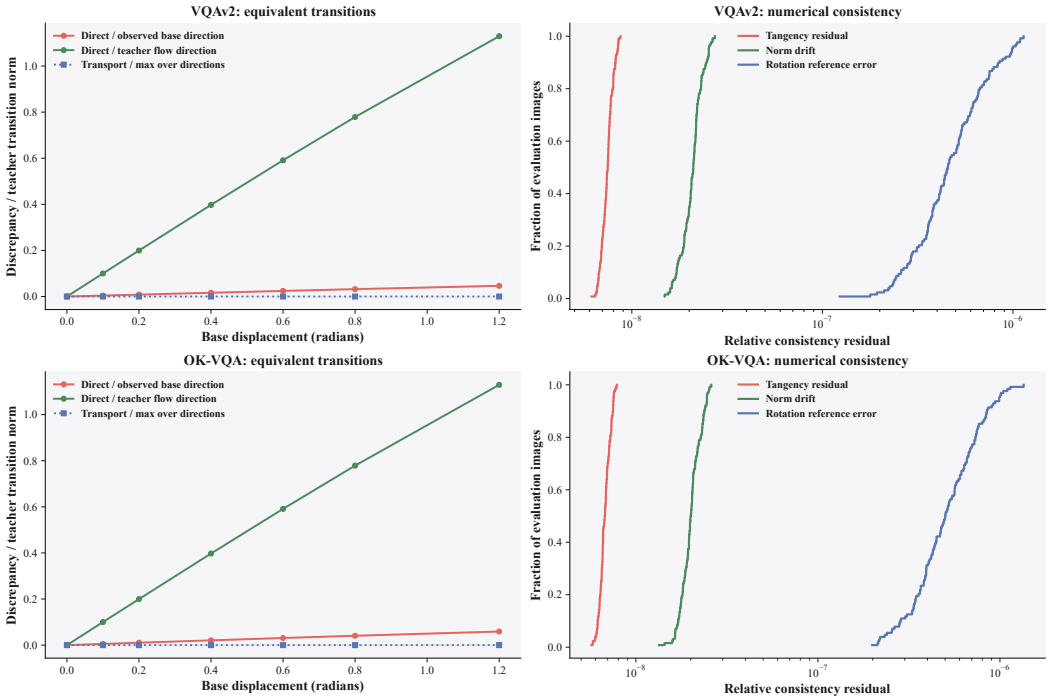


Figure 14: Geometric consistency of parallel transport on VQAv2 (top) and OK-VQA (bottom). Left: normalized discrepancies between equivalent transitions as the base point moves along the observed base direction or the teacher transition direction; the transport curve reports the maximum over both directions. Right: cumulative distributions of relative tangency residuals, norm drift, and errors against the independent rotation reference.

To determine whether this geometric correction preserves sensitivity to actual transition differences, we further evaluate the recovery of controlled perturbations. Using the same datasets, images, and four decoder depths, we displace the teacher base along the observed direction toward the student base and add a perturbation to the equivalent rotated transition. The perturbation is orthogonal to the plane of base displacement, and its magnitude is specified relative to the teacher transition norm. We vary both base displacement and perturbation strength, then measure the absolute deviation between the estimated normalized squared discrepancy and its known value. Figure 15 shows that direct subtraction incurs an additional error that grows with base displacement, whereas parallel transport accurately recovers the injected discrepancy throughout the tested grid. The near-zero values in the transport panels therefore indicate accurate measurement, including under nonzero perturbations. Together, these results show that the RST fiber comparison can exclude discrepancies caused by changing base points while retaining sensitivity to genuine differences between transitions.

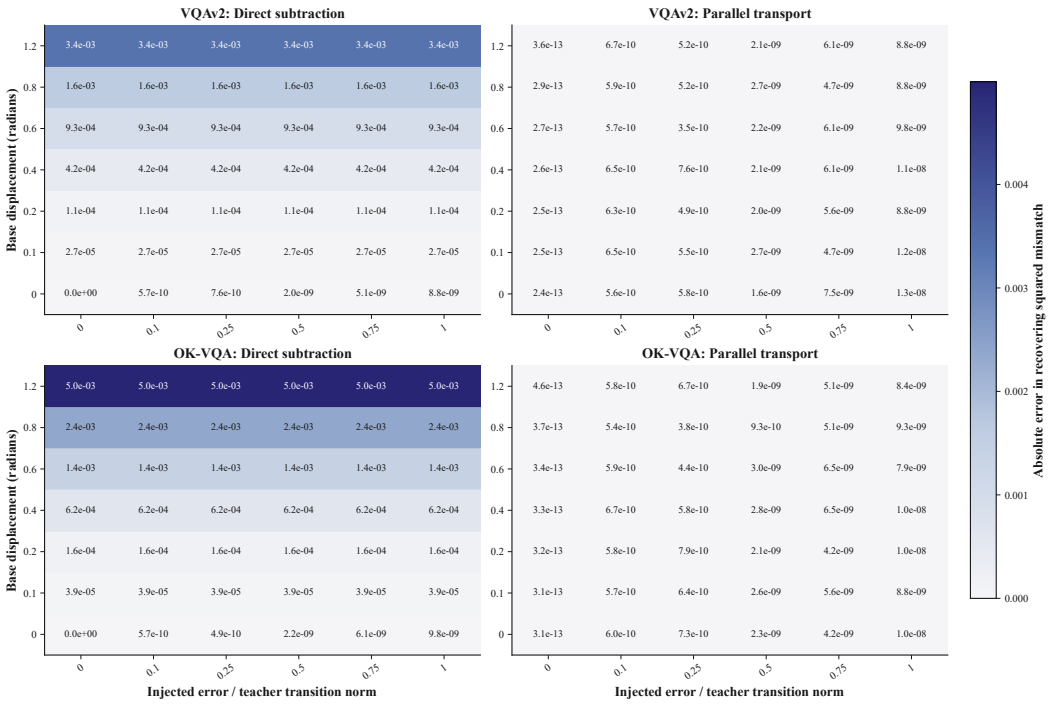


Figure 15: Recovery of controlled transition perturbations on VQAv2 (top) and OK-VQA (bottom), comparing direct subtraction (left) with parallel transport (right). Heatmaps report the absolute error in recovering the known normalized squared discrepancy across base displacements and perturbation strengths. Near-zero errors indicate accurate recovery of the injected discrepancy, not the absence of a transition difference.

A.8 Generalization to Other MLLMs

To examine the applicability of ICL-R beyond the backbones used in the main experiments, we extend our evaluation to Idefics1-9B and LLaVA-v1.6(Mistral-7B). The evaluation covers VQAv2 and OK-VQA for visual question answering and COCO for image captioning. We compare ICL-R with zero-shot inference, explicit ICL, MimIC, and HiFICL, as summarized in Table 7. For both backbones, all trainable methods use 8 demonstrations per training episode. On Idefics1-9B, ICL-R achieves the best performance on all three benchmarks, outperforming both explicit ICL and the implicit ICL baselines. On LLaVA-v1.6, it leads on VQAv2 and COCO and remains competitive on OK-VQA, trailing MimIC and HiFICL by only 0.42 and 0.25 points, respectively.

Table 7: Generalization results on additional MLLM backbones across visual question answering and image captioning benchmarks. Bold numbers denote the best result.

Model	Method	VQAv2	OK-VQA	COCO
Idefics1-9B	Zero-shot	42.77	29.08	0.6111
	8-shot ICL	54.75	47.30	0.9866
	MimIC	57.85	49.94	1.0895
	HiFICL	53.20	49.20	0.8283
	ICL-R	**57.91**	**52.84**	**1.1761**
LLaVA-v1.6(Mistral-7B)	Zero-shot	70.00	26.21	0.7157
	8-shot ICL	68.00	20.64	0.6678
	MimIC	71.24	**63.34**	1.2857
	HiFICL	74.71	63.17	1.3192
	ICL-R	**75.36**	62.92	**1.3203**

A.9 Dataset Transfer

To assess whether the demonstration effects learned by ICL-R generalize beyond the training dataset, we evaluate transfer between benchmarks within the same task family on Idefics2-8B-Base. We consider two settings: training the calibration modules on COCO Caption and evaluating on Flickr30K for image captioning, and training on VQAv2 and evaluating on OK-VQA for visual question answering. The learned modules are applied to the target benchmark without further training, and we compare against zero-shot inference and 8-shot ICL using CIDEr for captioning and answer accuracy for question answering. For 8-shot ICL, the demonstrations are sampled from the source benchmark. As shown in Table 8, ICL-R outperforms both baselines in both transfer settings, indicating that its gains extend to benchmarks not used to train the calibration modules. These results suggest that joint attention and value calibration captures demonstration effects that can transfer across related datasets, supporting its applicability to new benchmarks within the same task family.

Table 8: Cross-dataset transfer performance. Train and Test indicate the source and target datasets, respectively. Bold numbers denote the best scores.

Train	Test	Zero-shot	8-shot ICL	ICL-R
COCO	Flickr30K	0.5327	0.8364	**0.8802**
VQAv2	OK-VQA	43.08	51.43	**54.49**

A.10 Trainable Parameter Analysis

To assess the parameter cost of learning demonstration effects, we compare the number of trainable parameters across methods on Idefics2 and LLaVA-Interleave. Table 9 reports parameter counts in millions and their ratios relative to ICL-R; zero-shot inference and explicit 8-shot ICL require no trainable parameters. ICL-R uses 4.5M trainable parameters on both backbones, compared with 17.6M and 19.7M for LoRA, respectively. Its parameter count is also lower than that of MimIC on LLaVA-Interleave, but higher than those of LIVE and HiFICL on both backbones and MimIC on Idefics2. The larger parameter count of MimIC on LLaVA-Interleave stems from its dense query-dependent linear mapping $f(q)$, whose parameter requirements grow with the hidden dimension, bringing the total to 17.0M on this backbone. Thus, ICL-R does not minimize parameter count among all implicit ICL methods. Instead, its joint attention and value calibration provides the strong task performance reported in Table 1 with substantially fewer trainable parameters than LoRA and without updating the backbone.

Table 9: Trainable parameters in millions (M). Parentheses give the ratio to ICL-R on the same backbone; a dash denotes no trainable parameters.

Model	Zero-shot	8-shot ICL	LoRA	LIVE	MimIC	HiFICL	**ICL-R**
Idefics2	–	–	17.6 (×3.91)	0.13 (×0.03)	0.26 (×0.06)	2.2 (×0.49)	4.5 (×1.00)
LLaVA-Interleave	–	–	19.7 (×4.38)	0.13 (×0.03)	17.0 (×3.78)	2.2 (×0.49)	4.5 (×1.00)

A.11 Single-Pass Fused SDPA Implementation

We express the joint attention and value calibration in Eq. 7 using a single scaled dot product attention (SDPA) call followed by value decoding and mass scaling. The construction does not require explicitly forming the query-dependent value tensor, and a compatible memory-efficient SDPA backend also avoids materializing the full attention score matrix. We first derive the equivalence with attention dropout disabled. Batch and head indices are omitted in the per-token equations, and $Q_{0,t}$, $K_{0,s}$, and $V_{0,s}$ denote row vectors. The query used for gating in Section 3.2 is the corresponding column vector $q_t = Q_{0,t}^{\top}$; the channel gates ρ_t^A and ρ_t^V are represented as row vectors below. Let $s_0 = d_h^{-1/2}$ be the original attention scale, and define

$$\widehat{Q}_t^A = \frac{g_t^A}{s_0} \left[\rho_t^A \odot (Q_{0,t} U_Q)\right]. \tag{26}$$

We concatenate the base and calibration features as

$$Q^+ = Q_0 \,\|\, \widehat{Q}^A, \qquad K^+ = K_0 \,\|\, (K_0 U_K), \qquad V^+ = V_0 \,\|\, Z, \quad Z = V_0 P_V, \tag{27}$$

where $\|$ denotes concatenation along the feature dimension within each head. For batch size N, H heads, and query and key lengths T_q and T_k, the augmented tensors satisfy

$$\begin{aligned} Q^+ &\in \mathbb{R}^{N\times H\times T_q\times (d_h+r_A)}, \\ K^+ &\in \mathbb{R}^{N\times H\times T_k\times (d_h+r_A)}, \\ V^+ &\in \mathbb{R}^{N\times H\times T_k\times (d_h+r_V)}. \end{aligned} \tag{28}$$

The SDPA call explicitly uses the original scale s_0, not the default scale $(d_h + r_A)^{-1/2}$ associated with the augmented query and key dimension. No additional $r_A^{-1/2}$ factor is introduced, consistent with Eq. 7. The augmented score then satisfies

$$\begin{aligned} \left[s_0 Q^+ (K^+)^\top\right]_{t,s} &= s_0 Q_{0,t} K_{0,s}^\top + s_0 \widehat{Q}_t^A (K_{0,s} U_K)^\top \\ &= s_0 Q_{0,t} K_{0,s}^\top + g_t^A \left[\rho_t^A \odot (Q_{0,t} U_Q)\right] (K_{0,s} U_K)^\top \\ &= S_{0,t,s} + \Delta S_{t,s}^A, \end{aligned} \tag{29}$$

which is exactly the calibrated pre-mask score in Eq. 7. We therefore compute

$$Y^+ = \mathrm{SDPA}(Q^+, K^+, V^+; M, s_0) = O^P \,\|\, Z^P, \tag{30}$$

where M is the additive form of the original causal and padding mask, with zero at allowed positions and $-\infty$ at masked positions. Each evaluated query row is assumed to have at least one allowed key. The attention matrix implicitly produced by this call is $\widetilde{A} = \mathrm{softmax}(S_0 + \Delta S^A + M)$, and the two output blocks satisfy

$$O_t^P = \sum_s \widetilde{A}_{t,s} V_{0,s}, \qquad Z_t^P = \sum_s \widetilde{A}_{t,s} (V_{0,s} P_V). \tag{31}$$

To match the mass-aware weights in Eq. 6, define

$$m_t = s_m \, \sigma\!\left(\frac{\bar{q}_t^\top w_m}{\sqrt{d_h}} + b_m\right), \qquad \widehat{A}_{t,s} = (1 - m_t)\widetilde{A}_{t,s}, \tag{32}$$

where $\bar{q}_t = \mathrm{LN}(q_t)$ and the same attention mask is understood in the method formulation. The value correction can be decoded after attention because its query-dependent gates are constant with respect to the key index:

$$\begin{aligned} O_t^N &= O_t^P + g_t^V \left[\left(\rho_t^V \odot Z_t^P\right) R_V\right] \\ &= \sum_s \widetilde{A}_{t,s} \left(V_{0,s} + g_t^V \left[\left(\rho_t^V \odot V_{0,s} P_V\right) R_V\right]\right), \\ \widetilde{O}_t &= (1 - m_t) O_t^N. \end{aligned} \tag{33}$$

Here, $O_t^N = \sum_s \widetilde{A}_{t,s} \widetilde{V}_{t,s}$ denotes the normalized aggregation before applying the total-mass factor. Consequently, $\widetilde{O}_t = \sum_s \widehat{A}_{t,s} \widetilde{V}_{t,s}$, establishing algebraic equivalence to Eq. 6 under the same mask and with attention dropout disabled. If attention dropout is enabled, the same rearrangement holds for a fixed dropout realization when both formulations use identical dropped attention weights; these weights no longer necessarily sum to one. The augmented feature construction, value decoding, and mass scaling occur outside the SDPA call. Softmax, masking, optional dropout, and value aggregation are handled by SDPA. A memory-efficient fused backend can be used when the augmented dimensions, data type, and hardware are supported; a single SDPA call alone does not guarantee fused execution.

A.12 Peak GPU memory usage

To assess the memory overhead of joint attention and value calibration, we compare the peak GPU memory usage of ICL-R, HiFICL, and MimIC on VQAv2, OK-VQA, and COCO Caption. Figure 16 reports relative memory usage, normalized to HiFICL on each benchmark. ICL-R requires

1.13×, 1.08×, and 1.13× the memory of HiFICL, respectively, whereas MimIC requires 1.50×, 1.40×, and 1.50×. Thus, ICL-R incurs an 8%–13% overhead over HiFICL while using less memory than MimIC on all three benchmarks. The fused SDPA implementation described above avoids explicitly materializing calibrated attention logits and query-dependent value tensors, helping limit the intermediate storage required by joint calibration. Overall, these results indicate that joint attention and value calibration can be implemented with a modest memory overhead relative to HiFICL, although ICL-R does not achieve the lowest peak memory usage among the compared methods.

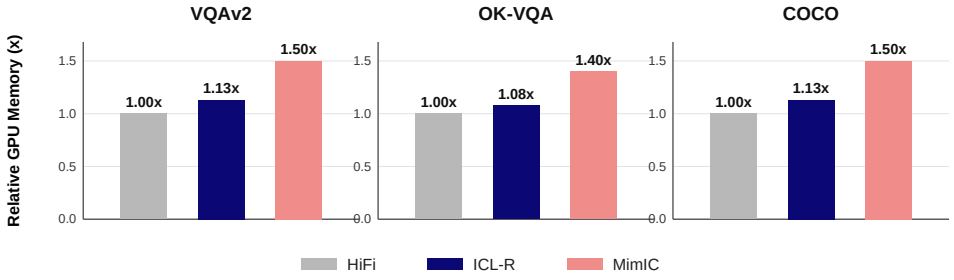


Figure 16: Relative peak GPU memory usage of HiFICL, ICL-R, and MimIC on VQAv2, OK-VQA, and COCO Caption, normalized to HiFICL on each benchmark.

A.13 Error Correction Across Categories and Backbones

To identify which capabilities benefit from ICL-R and how individual prediction changes contribute to its accuracy gains, we analyze paired predictions on MME and SEED-Bench with Idefics2 and LLaVA-Interleave. For each backbone, we compare ICL-R with zero-shot inference, 8-shot ICL, LoRA, LIVE, MimIC, and HiFICL on the same evaluation examples using a common answer parser. We examine category accuracy differences, the balance between corrected and newly introduced errors, and the uncertainty of the resulting overall gains.

Category accuracy differences. To determine whether the improvements span multiple capabilities, we compute the accuracy difference between ICL-R and each baseline in percentage points for all MME categories and SEED-Bench categories. Figure 17 reports these differences, with the number of questions shown for each category. On Idefics2, ICL-R matches or exceeds 8-shot ICL in every evaluated category, improving all categories except MME position, where performance is unchanged. Gains are broadly distributed, with particularly large improvements on SEED-Bench instance location and counting, reaching 18.2 and 23.7 percentage points over 8-shot ICL, respectively. This pattern is consistent with the benefits of joint attention and value calibration. Specifically, adjusting attention can help identify relevant objects, while calibrating values can improve the object information aggregated for localization and counting. However, on SEED-Bench text understanding, ICL-R underperforms all four trainable baselines. A possible explanation is that value calibration introduces interference into representations used for text recognition, altering fine-grained textual information even when attention selects the relevant image regions. On LLaVA-Interleave, ICL-R also improves broadly over explicit ICL, with positive gains in every evaluated category, although its advantages over trainable baselines are less uniform. MME position is a notable weakness. Despite improving over 8-shot ICL, ICL-R falls below zero-shot inference, LoRA, and LIVE. One possible explanation is that the joint corrections favor salient object content but sometimes disturb the relative spatial information needed to distinguish object positions, limiting their benefit for precise positional judgments. These observations suggest that joint calibration benefits object selection and aggregation, while potentially introducing interference in text recognition and spatial reasoning.

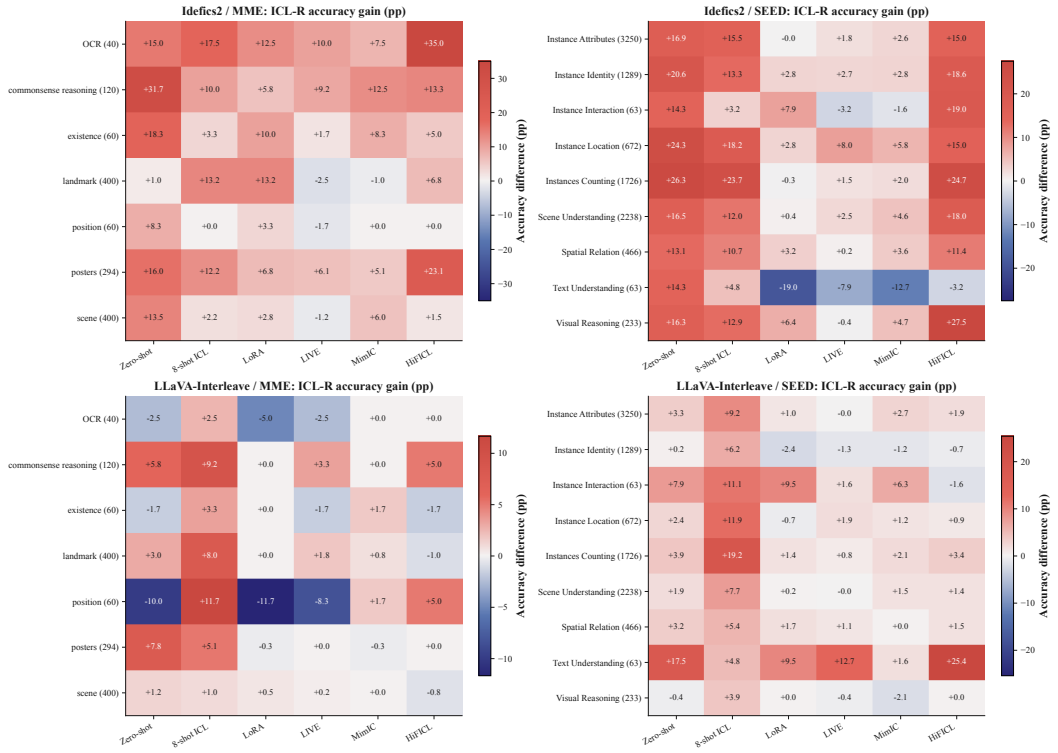


Figure 17: Category accuracy differences between ICL-R and each baseline, measured in percentage points. Rows correspond to Idefics2 (top) and LLaVA-Interleave (bottom); columns correspond to MME (left) and SEED-Bench (right). Category sizes are shown in parentheses. Positive values indicate improvements, while negative values indicate regressions.

Corrected and newly introduced errors. To distinguish beneficial prediction changes from regressions, we count questions that each baseline answers incorrectly but ICL-R answers correctly, and questions for which the reverse holds. Figure 18 expresses both counts as shares of all evaluated questions, so their difference equals the overall accuracy gain. On Idefics2, corrected errors outnumber newly introduced errors for every comparator on both benchmarks. For example, relative to HiFICL on MME, ICL-R corrects 223 errors while introducing 89 new errors, yielding a net gain of 9.75 percentage points. On LLaVA-Interleave, corrections clearly exceed new errors relative to 8-shot ICL, whereas the two are more closely balanced relative to LoRA and LIVE. Thus, the gains over explicit ICL reflect a favorable balance of corrections and regressions on both backbones, while this balance is less consistently favorable against strong trainable baselines on LLaVA-Interleave.

Paired overall gains and uncertainty. To assess the uncertainty of the aggregate improvements, we estimate 95% confidence intervals for paired accuracy differences using 2,000 bootstrap resamples of images, retaining all questions associated with each sampled image. Figure 19 reports the net gains and their intervals for both benchmarks and backbones. Relative to HiFICL, ICL-R gains 9.75 points on MME and 17.83 points on SEED-Bench with Idefics2, and both confidence intervals exclude zero. On LLaVA-Interleave, the gains over 8-shot ICL are 5.24 and 10.08 points, with confidence intervals of [3.35, 7.13] and [8.85, 10.60], respectively. It also gains 1.72 points over HiFICL on SEED-Bench, whereas the smaller differences from LoRA and LIVE have confidence intervals that include zero. These results provide evidence that ICL-R can improve accuracy across the evaluated backbones and benchmarks.

To examine whether the same capabilities benefit across backbones, we pair the accuracy gains for each category and comparator and plot the gains on Idefics2 and LLaVA-Interleave as the two coordinates in Figure 20. This comparison includes all seven MME categories and nine SEED-Bench categories, retaining positive, negligible, and negative differences. Relative to HiFICL, ICL-R improves MME OCR accuracy by 35.00 percentage points on Idefics2, while leaving it unchanged

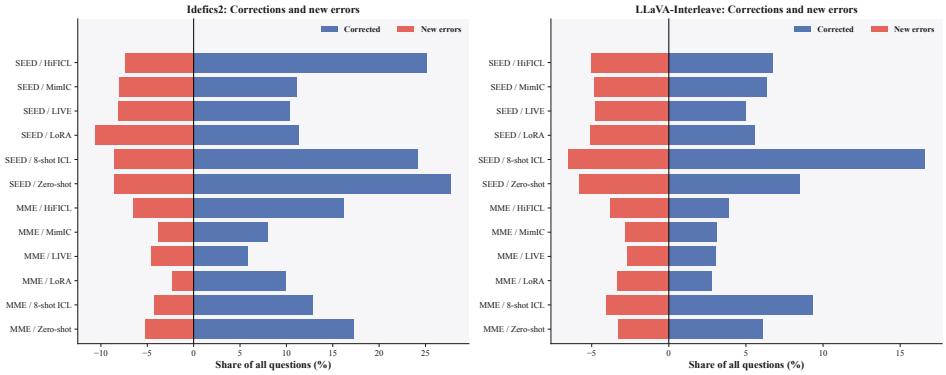


Figure 18: Corrected and newly introduced errors on Idefics2 (left) and LLaVA-Interleave (right). For each benchmark and baseline, bars show the shares of all questions corrected by ICL-R (right of zero) or newly answered incorrectly (left of zero). The difference between their magnitudes gives the net accuracy gain.

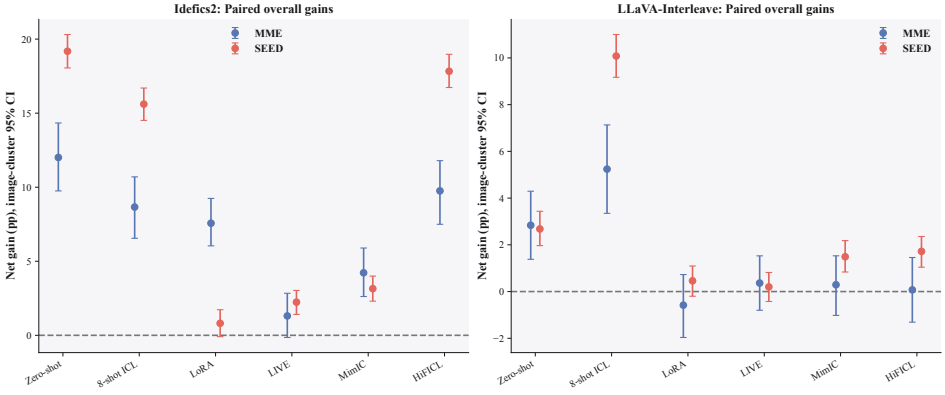


Figure 19: Paired overall accuracy gains of ICL-R over each baseline on Idefics2 (left) and LLaVA-Interleave (right). Markers show net gains in percentage points on MME and SEED-Bench; error bars show 95% confidence intervals from 2,000 paired bootstrap resamples of images. The dashed line denotes no accuracy difference.

on LLaVA-Interleave. For SEED-Bench text understanding, the corresponding changes are -3.17 and +25.40 points. These patterns show that the distribution and magnitude of the gains depend on the backbone. Category analysis thus complements aggregate accuracy by revealing both shared improvements and differences in where ICL-R provides the greatest benefit.

A.14 Paired discrimination on MME

We complement question-level accuracy with the existing MME Accuracy+ metric, which requires both questions associated with an image to be answered correctly. We retain the same 687 evaluation images and all seven methods on both backbones, including invalid outputs as errors. ICL-R increases paired accuracy from 59.68% to 74.53% on Idefics2 and from 60.99% to 72.05% on LLaVA-Interleave, relative to explicit 8-shot ICL. The corresponding paired gains are 14.85 percentage points (95% CI: [11.21,18.49]) and 11.06 points (95% CI: [7.71,14.56]). Confidence intervals use 2,000 bootstrap resamples of complete images. The within-image error distribution and transition matrices show how changes in individual answers translate into more images with two correct judgments. Advantages over trainable baselines are less uniform: Idefics2's interval relative to LIVE includes zero, as do LLaVA-Interleave's intervals relative to all four trainable baselines.

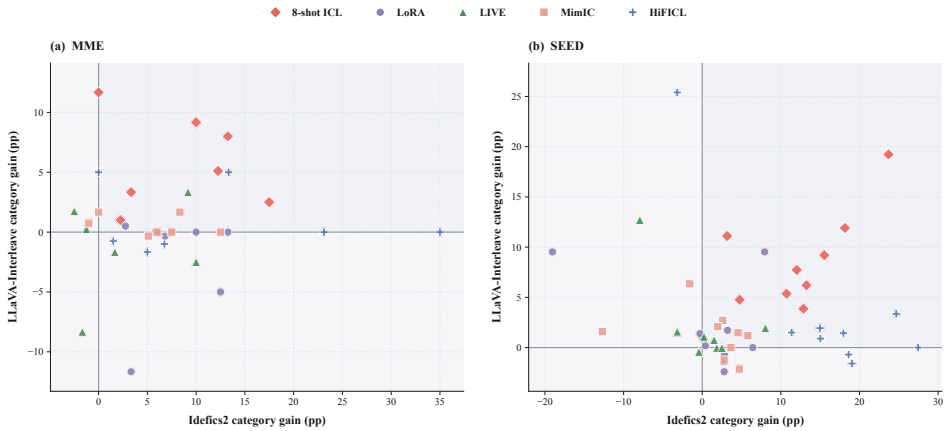


Figure 20: Category accuracy gains across backbones on MME (left) and SEED-Bench (right). Each point represents a category and comparator, with gains on Idefics2 and LLaVA-Interleave shown on the horizontal and vertical axes, respectively. Gains are measured in percentage points.

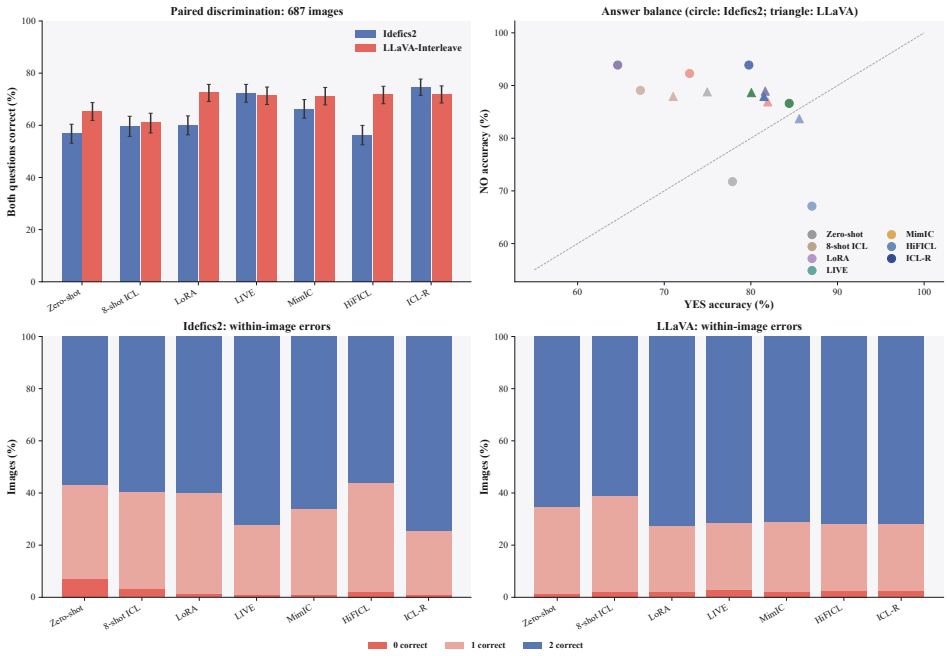


Figure 21: Paired discrimination on the fixed MME evaluation subset. The panels show both-questions-correct accuracy with image-bootstrap 95% intervals, YES/NO accuracy, and the fraction of images with zero, one or two correct answers. All seven methods are retained for both backbones.

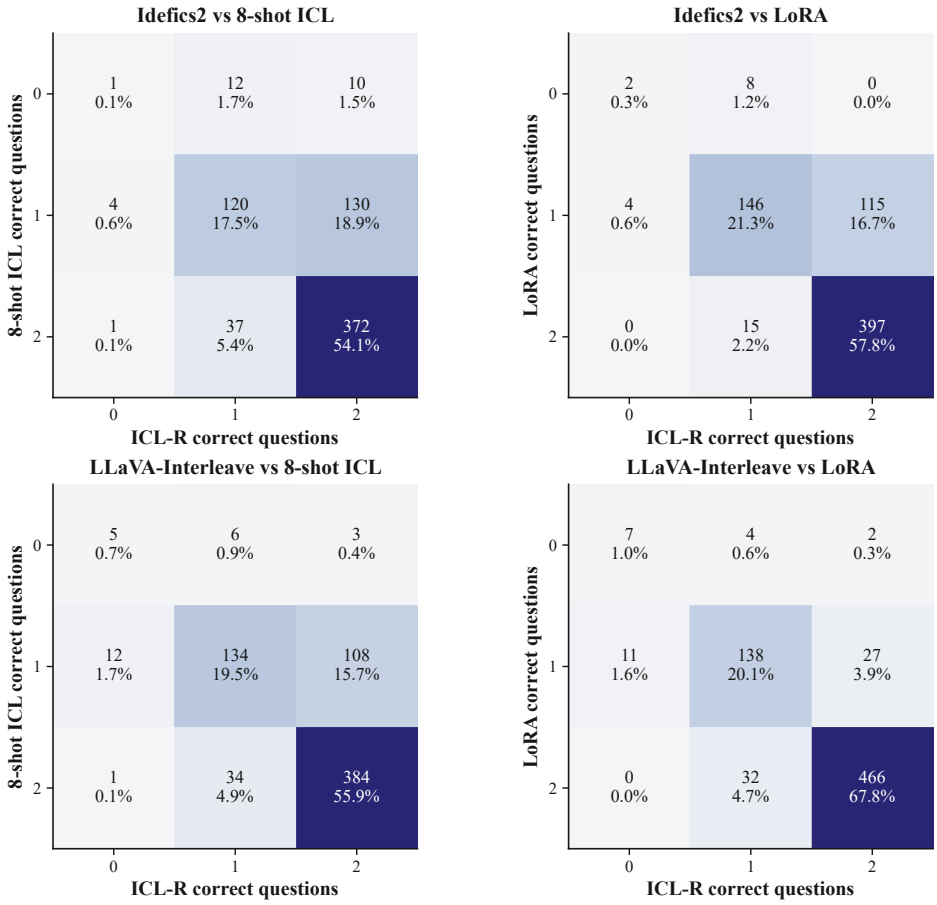


Figure 22: Changes in the number of correctly answered questions per MME image. Rows denote the comparator's correct-question count and columns denote ICL-R's count. Cells show image counts and percentages of all 687 evaluation images. Results are shown against explicit 8-shot ICL and LoRA on both backbones.

A.15 Residual failures and qualitative limitations.

We distinguish persistent errors from newly introduced errors on the complete paired MME and SEED evaluation records. Task categories are taken from benchmark annotations and are not interpreted as causal labels for internal model failure. We additionally prepare a fixed stratified sample of 120 ICL-R errors, with 30 records for each backbone–dataset combination, and inspect ten diagnostic examples selected by declared task and outcome strata. This qualitative review includes text-associated color errors, a failed verification of the visible words "CASA GRECQUE" and "RESTAURANTS", and incorrect spatial judgments, alongside correct controls. Some examples also expose evaluation ambiguity: two text-related controls concern stock-image watermarks, one spatial question calls a visible fire extinguisher a fire hydrant, and a scene with multiple cranes leaves the intended spatial referents unclear. We retain these records in the quantitative evaluation and report their limitations rather than attributing them to a particular calibration branch. The qualitative review is not a double-annotated estimate of causal error-type prevalence.

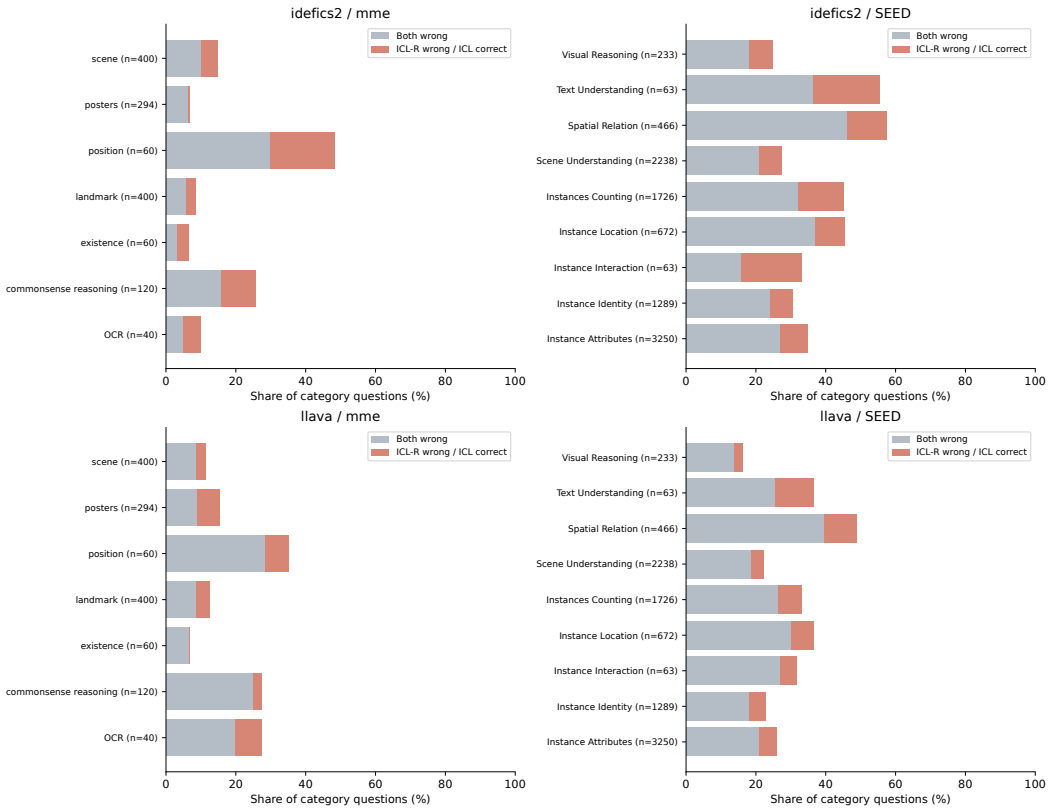


Figure 23: Residual ICL-R errors by benchmark task category. Gray bars represent questions answered incorrectly by both ICL-R and explicit 8-shot ICL; red bars represent questions answered correctly by explicit ICL but incorrectly by ICL-R. Percentages use all questions in each category, with category sample sizes included. The categories describe task requirements rather than proven internal failure mechanisms.

A.16 Qualitative Examples Analysis

We examine 56 curated examples to understand where ICL-R's gains arise and where its predictions remain unreliable. In the MME cases (Figure 24), ICL-R answers every question correctly, while at most two of the six baselines answer any individual question correctly. The examples span exact logo spelling, chart reading, object-related commonsense, and spatial relations, suggesting that the advantage is not confined to a single question type. The SEED-Bench cases in Figures 25 and 26 provide further examples where only ICL-R selects the ground-truth answer. Several require identifying what is absent or recognizing what cannot be inferred from the image: no cat is visible among the cushions, and a monochrome photograph does not reveal a guitar's actual color. Figure 27 provides an important complement: ICL-R also succeeds on cases solved by one or more baselines. The qualitative advantage therefore concerns which visual distinctions are preserved across cases.

The Flickr30k comparisons reveal differences that answer accuracy alone cannot express (Figures 28 and 29). ICL-R identifies the yellow vehicle as a pedicab where several baselines call it a taxi, describes a conductor rather than a trumpet player, and retains details such as a tattooed arm or the number of medals. These are phrase-level comparisons: a baseline caption can contain both useful observations and an unsupported detail. The failure cases expose three remaining challenges. On VQAv2, ICL-R can bind a visible attribute to the wrong object or misread what the question requests, as when it gives the color of a suitcase for a purse or a time instead of the word printed on a clock. On OK-VQA, recognizing the pictured object does not always lead to the requested knowledge: the model answers "dog" when asked for its wild ancestor. On COCO, fluent captions can still introduce unsupported objects, equipment, or actions, including a dog absent from a vehicle

and a tennis racket in place of a frisbee (Figure 30). Together, these examples suggest that efficient query-only adaptation must preserve task-relevant visual evidence while also controlling how that evidence is used in an answer. The selected cases illustrate these behaviors; their prevalence and overall performance are assessed by the quantitative results.

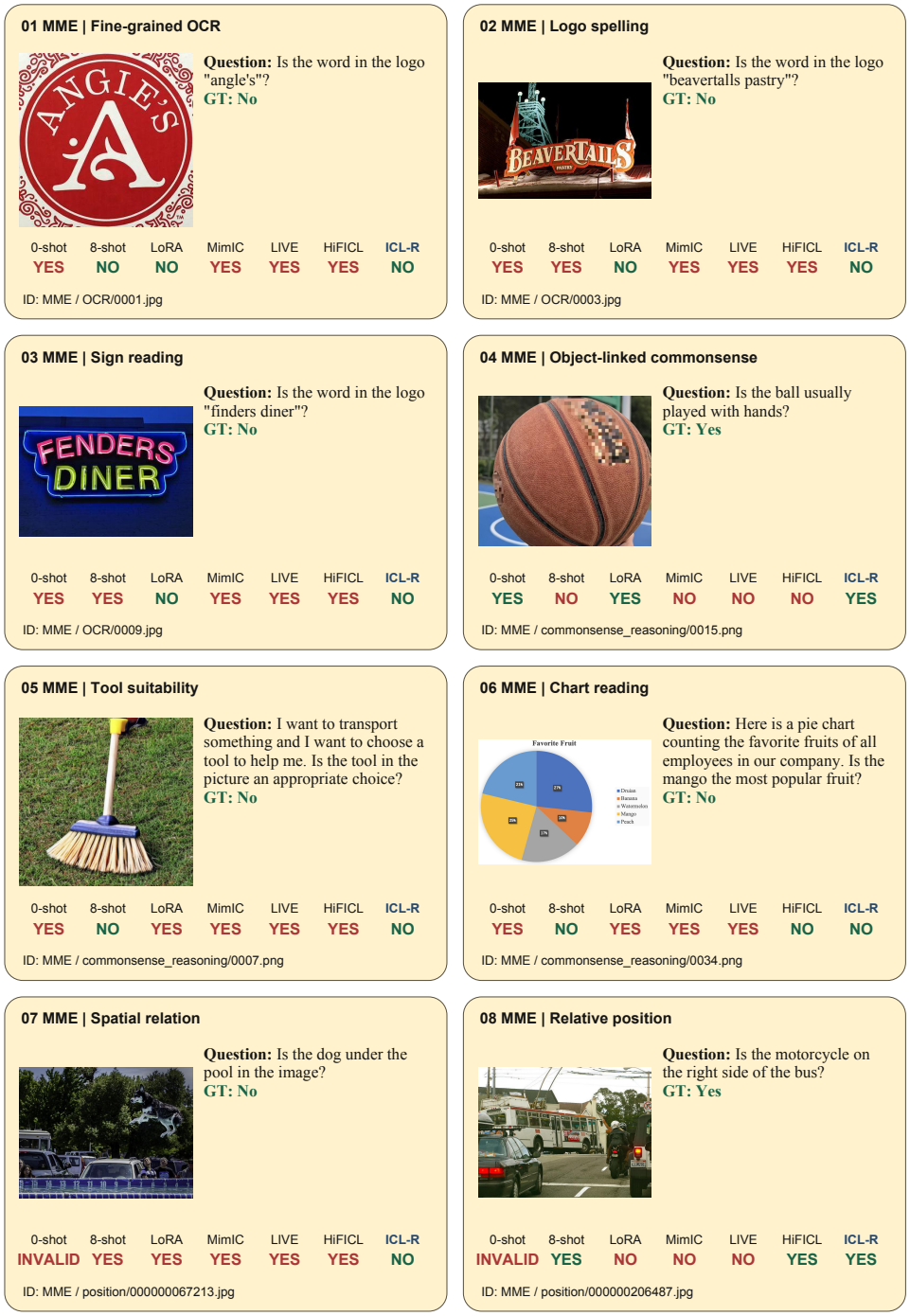


Figure 24: Selected MME examples using Idefics2. ICL-R answers all eight correctly, while no more than two of the six baselines answer any individual case correctly. The examples cover fine-grained OCR, commonsense, chart reading, and spatial relations. Green denotes correct answers; red denotes incorrect or invalid answers.

09 SEED | Absence detection



Question: How many cats are in the image?
A. 1
B. 0
C. 2
D. 3
GT: B

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	A	A	A	A	B

ID: SEED / 105046

10 SEED | Distant-object counting



Question: How many wind turbines are visible in the image?
A. 1
B. 7
C. 5
D. 3
GT: D

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	B	C	B	B	D

ID: SEED / 1358

11 SEED | Visual commonsense



Question: What can be inferred about the green strawberries based on their position?
A. They are smaller in size than the red strawberries
B. They are riper than the red strawberries
C. They are not yet ripe
D. They are overripe
GT: C

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
B	B	A	D	A	A	C

ID: SEED / 15327

12 SEED | People counting



Question: How many people are present in the image?
A. 2
B. 8
C. 6
D. 4
GT: D

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	C	C	C	A	D

ID: SEED / 21641

13 SEED | Limits of visual evidence



Question: What is the color of the guitar in the image?
A. White
B. Black
C. Brown
D. The image is in black and white, so it's impossible to tell
GT: D

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
B	C	B	B	B	B	D

ID: SEED / 34543

14 SEED | Attribute absence



Question: Which of the following attributes is NOT shared by any of the women in the picture?
A. Smiling
B. Wearing a hat
C. Standing
D. Posing
GT: B

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	C	C	C	C	B

ID: SEED / 5755

15 SEED | Object identity



Question: How many clocks are in the image?
A. 1
B. 0
C. 2
D. 3
GT: B

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	A	A	A	A	B

ID: SEED / 62405

16 SEED | Door absence



Question: Where is the door located in the image?
A. To the right of the fireplace
B. To the left of the fireplace
C. Behind the fireplace
D. There is no door in the image
GT: D

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
B	B	A	B	B	A	D

ID: SEED / 102670

Figure 25: SEED-Bench examples answered correctly only by ICL-R among the seven Idefics2 methods. These cases involve object absence, counting, visual commonsense, and recognizing the limits of visual evidence. Green denotes correct answers and the ground-truth choice; red denotes incorrect baseline answers.

17 SEED | Furniture proximity



Question: Which piece of furniture is located closest to the window?
A. Bed
B. Nightstand
C. Chair
D. Table
GT: A

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	C	C	C	B	A

ID: SEED / 46986

18 SEED | Window counting



Question: How many windows are in the room?
A. One
B. Two
C. Three
D. Four
GT: A

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	B	B	B	B	A

ID: SEED / 14057

19 SEED | Missing object



Question: Which of the following objects is NOT present in the image?
A. Balloons
B. Gift boxes
C. Hats
D. None of the above
GT: A

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
D	D	D	D	D	D	A

ID: SEED / 11382

20 SEED | Horseback relation



Question: Where are the people positioned relative to the horses?
A. In front of the horses
B. Behind the horses
C. On top of the horses
D. Next to the horses
GT: C

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
B	B	D	D	D	A	C

ID: SEED / 16103

21 SEED | Ball color



Question: What color is the soccer ball in the image?
A. Yellow
B. White
C. Black and white
D. Red
GT: C

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
B	B	B	B	B	B	C

ID: SEED / 53792

22 SEED | Toy cars vs. suitcases



Question: Are there any suitcases in the image?
A. No
B. Yes, one yellow and one red
C. Yes, one red and one blue
D. Yes, one yellow and one green
GT: A

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	B	B	B	B	B	A

ID: SEED / 56128

23 SEED | Candle absence



Question: How many candles are in the image?
A. None
B. One
C. Three
D. Two
GT: A

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
D	C	B	D	D	B	A

ID: SEED / 1231

24 SEED | Absent flowers



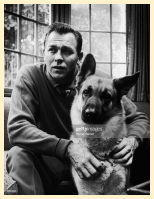
Question: Which of the following is not present in the image?
A. microphone
B. flowers
C. building
D. priest
GT: B

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
D	C	A	D	C	D	B

ID: SEED / 106310

Figure 26: Additional SEED-Bench examples answered correctly only by ICL-R. The cases test object identity, spatial relations, counting, and discrimination between visually similar objects. Green denotes correct answers and the ground-truth choice; red denotes incorrect baseline answers.

25 SEED | Relative position



Question: Where is the dog positioned in the image?
A. Below the man
B. Right of the man
C. Above the man
D. Left of the man
GT: B

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
B	C	D	C	B	A	B

ID: SEED / 100868

26 SEED | Background relation



Question: What is the relation between the boat and the mountain in the image?
A. The boat is sailing towards the mountain
B. The mountain is in the background
C. The boat is sailing away from the mountain
D. There is no relation
GT: B

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	B	A	C	C	A	B

ID: SEED / 102294

27 SEED | Instrument counting



Question: How many men are playing guitars on stage?
A. One
B. Two
C. Three
D. Four
GT: D

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	D	C	C	B	D

ID: SEED / 105913

28 SEED | Repeated pattern



Question: Which object can be found in multiple places in the image?
A. Star
B. Notebook
C. Paper
D. Polka dots
GT: D

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	B	C	D	A	D

ID: SEED / 102021

29 SEED | Uniform color



Question: What is the predominant color of the football team's uniform in the image?
A. Yellow
B. Purple
C. Blue
D. Orange
GT: B

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
B	A	B	A	A	A	B

ID: SEED / 105309

30 SEED | Hat detection



Question: Is the man in the image wearing a hat?
A. Yes
B. No
C. Cannot determine
D. Ambiguous
GT: A

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	B	A	B	B	B	A

ID: SEED / 10225

31 SEED | Race-car counting



Question: How many race cars can be seen in the image?
A. Three
B. Five
C. Four
D. Six
GT: A

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	C	A	C	A	A

ID: SEED / 102534

32 SEED | Chair counting



Question: How many chairs are in the living room?
A. 0
B. 1
C. 2
D. 3
GT: B

0-shot	8-shot	LoRA	MimIC	LIVE	HiFICL	ICL-R
C	C	B	A	B	A	B

ID: SEED / 102868

Figure 27: SEED-Bench examples where ICL-R and at least one baseline answer correctly. The comparisons cover spatial relations, repeated patterns, colors, and counting, showing which capabilities are shared and where predictions differ. Green denotes correct answers; red denotes incorrect answers.

33 Flickr30k | Vehicle identity



ID: Flickr30k / 6046473448.jpg

Zero-shot: A man sits on the back of a rickshaw in London.
8-shot ICL: A man is sitting on the back of a **yellow taxi**. Image: C
LoRA: A man is sitting on the back of a **yellow taxi**.
MimIC: A man is sitting on a **yellow taxi** with a black top and a black seat.
LIVE: A man is sitting on a **yellow taxi cab** with a black canopy.
HiFICL: A man in a black shirt and jeans sits on a **yellow taxi cab**.
ICL-R: A man sits on the back of a **yellow pedicab**.
Reference (1/5): A pedicab driver waiting on his bike

34 Flickr30k | Unsupported action



ID: Flickr30k / 2373370266.jpg

Zero-shot: A child dressed as a Jedi knight, complete with lightsaber and stroller.
8-shot ICL: A young boy in a Jedi robe is holding a toy lightsaber and talking on a
LoRA: A young boy dressed in a black robe with a white belt and a black cape is holding
MimIC: A young boy in a black robe with a sword is **talking on a cell phone** while pushing a
LIVE: A young boy in a black robe is holding a sword and **talking on a cell phone**.
HiFICL: A young boy in a black robe with a sword in his hand is **talking on a cell phone**
ICL-R: A young boy wearing a black robe and **holding a sword** is standing **next to a stroller**.
Reference (1/5): A child standing by a stroller and holding a toy sword.

35 Flickr30k | Grounded description



ID: Flickr30k / 3634828052.jpg

Zero-shot: A photographer takes a picture of a group of people waiting to cross the street.
8-shot ICL: A woman dressed in a pink and white dress is taking a picture of something off-camera.
LoRA: A woman in a blue and white dress is taking a picture of a woman in a black dress and
MimIC: A woman with curly red hair is taking a picture of **a man in a suit and tie**.
LIVE: A woman with curly hair is taking a picture of **a man in a suit and tie**.
HiFICL: A woman with curly hair is taking a picture of **a man in a suit and tie**.
ICL-R: A woman with curly red hair is taking a picture with a **large camera**.
Reference (1/5): A woman taking a picture with a large professional camera.

36 Flickr30k | Relevant detail



ID: Flickr30k / 3306464579.jpg

Zero-shot: A pancake being flipped in the air.
8-shot ICL: A man flips a pancake in a frying pan. Image:
LoRA: A man in a **black shirt** is flipping a pancake in a black pan.
MimIC: A man in a **black shirt** is making pancakes in a pan.
LIVE: A man is making pancakes in a pan.
HiFICL: A man in a **black shirt** is making pancakes in a pan.
ICL-R: A man with a **tattoo on his arm** is flipping a pancake in a frying pan
Reference (1/5): Hands holding a frying pan.

37 VQAv2 | Wrong object / color



ID: VQAv2 / 208761002

Question: What color is the purse?
ICL-R: red
GT: blue (10/10)

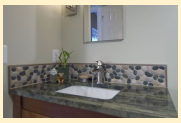
38 VQAv2 | Shape error



ID: VQAv2 / 114303002

Question: Is the mirror square?
ICL-R: yes
GT: no (10/10)

39 VQAv2 | Reflected-content error



ID: VQAv2 / 69121011

Question: What is reflected in the mirror?
ICL-R: sink
GT: door (9/10); doorway (1/10)

40 VQAv2 | Text vs. time



ID: VQAv2 / 64150000

Question: What is the clock written?
ICL-R: 12:00
GT: colgate (10/10)

Figure 28: Flickr30k caption comparisons (top) and ICL-R failures on VQAv2 (bottom). The Flickr30k cases highlight grounded details in ICL-R's descriptions and unsupported details in some baseline captions. The VQAv2 cases show errors in object attributes, mirror interpretation, and reading the question. Green marks supported phrases or ground-truth answers; red marks incorrect or unsupported content.

41 Flickr30k | Crucifix detail



ID: Flickr30k / 4442382200.jpg

Zero-shot: A priest of the Russian Orthodox Church in New York City. The Russian Orthodox Church
8-shot ICL: A man in a black robe is holding a cross in the middle of the street. Image
LoRA: A man with a white beard and white gloves is holding a cross in the middle of a busy street
MimIC: A man with a long white beard and a black hat holding a cross.
LIVE: A man in a black robe and a white beard is holding a cross in the middle of a
HiFICL: A man in a black robe and a cross in his hand is standing in the middle of a
ICL-R: A man in a black robe and **white gloves holds a cross** in the middle of a street.
Reference (1/5): A man with black and white gloves holds a cross with a crucified Jesus on it in a street.

42 Flickr30k | Conductor vs. player



ID: Flickr30k / 432870699.jpg

Zero-shot: This is a picture of a jazz band.
8-shot ICL: A man in a white shirt is conducting an orchestra. Image: C
LoRA: A man in a white shirt is conducting a group of musicians in a concert hall.
MimIC: A man in a white shirt is **playing a trumpet** in front of a group of musicians.
LIVE: A man in a white shirt is **playing a trumpet** in front of a group of musicians.
HiFICL: A man in a white shirt is **playing a trumpet** in front of a group of musicians.
ICL-R: A man in a white shirt is **conducting an orchestra**.
Reference (1/5): Man in white shirt conducting a orchestra, horns and drums.

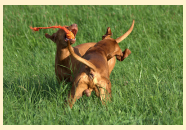
43 Flickr30k | Medal counting



ID: Flickr30k / 3667157255.jpg

Zero-shot: The Italian swimmer Federica Pellegrini won the gold medal in the 20
8-shot ICL: A female swimmer is holding up **three gold medals**. Image: C
LoRA: A woman in a gray jacket is holding up two gold medals and two fingers.
MimIC: A woman in a gray jacket and gold medals around her neck.
LIVE: (empty output)
HiFICL: A young woman with short blonde hair is holding up two gold medals and **two silver medals**
ICL-R: A woman in a gray jacket is holding up **two gold medals**.
Reference (1/5): An Olympic winner wearing her two gold medals and a gray jacket showing a peace sign and a wave

44 Flickr30k | Multiple animals



ID: Flickr30k / 1206506157.jpg

Zero-shot: A pair of **dachshunds** playing in the grass.
8-shot ICL: Two dogs playing in the grass with a toy. Image: Caption:
LoRA: A brown dog is running through the grass with a red toy in its mouth.
MimIC: Two dogs are playing in the grass with a toy.
LIVE: Two dogs are playing in the grass with a toy.
HiFICL: Two brown dogs are playing in the grass with a red toy.
ICL-R: Two brown dogs are playing in the grass.
Reference (1/5): Two large brown dogs are playing in the grass.

45 OK-VQA | Question / knowledge mismatch



ID: OK-VQA / 2626265

Question: What happens when you cross paths with this color of cat?
ICL-R: black
GT: bad luck (8/10); it is bad luck (2/10)

46 OK-VQA | Animal knowledge



ID: OK-VQA / 2369045

Question: What wild animal did this animal evolve from?
ICL-R: dog
GT: wolf (10/10)

47 OK-VQA | Everyday convention



ID: OK-VQA / 1960855

Question: When discussing things over dinner what part of the body is it considered rude to place upon the table?
ICL-R: foot
GT: elbow (10/10)

48 OK-VQA | Brand recognition



ID: OK-VQA / 5242165

Question: What gas station is pictured?
ICL-R: chevron
GT: shell (10/10)

Figure 29: Further Flickr30k caption comparisons (top) and ICL-R failures on OK-VQA (bottom). ICL-R captures details such as a conductor's role and the number of medals, but the OK-VQA cases expose difficulties with external knowledge, everyday conventions, and brand recognition. Green marks supported phrases or ground-truth answers; red marks incorrect or unsupported content.

49 COCO | Equipment confusion



ICL-R: A man on a **snowboard** jumping over another man on a **snowboard**.
Reference (1/5): **Skiers** perform on a lighted half-pipe beside a mountain.

ID: COCO / COCO_val2014_000000076822.jpg

50 COCO | Action hallucination



ICL-R: A woman bending over to **pick up a piece of grass**.
Reference (1/5): A woman **videos a man cutting the cement** around a street pole.

ID: COCO / COCO_val2014_000000522778.jpg

51 COCO | Participant-role confusion



ICL-R: A man is **flying a kite** in a field.
Reference (1/5): A man wearing a helmet **suspended from the air by ropes**.

ID: COCO / COCO_val2014_000000021919.jpg

52 COCO | Object hallucination



ICL-R: **A dog** sitting in the drivers seat of a van.
Reference (1/5): **An empty truck** with white doors and a yellow body.

ID: COCO / COCO_val2014_000000402903.jpg

53 COCO | Frisbee vs. racket



ICL-R: A man holding a **tennis racket** on top of a dirt road.
Reference (1/5): A man is **throwing a frisbee** in a sandy area.

ID: COCO / COCO_val2014_000000403349.jpg

54 COCO | Unsupported movement



ICL-R: A man and a woman are **snowboarding down a hill**.
Reference (1/5): The snowboarders are **posing for a picture** on the slope.

ID: COCO / COCO_val2014_000000267709.jpg

55 COCO | Unsupported activity



ICL-R: A baby sitting on a couch **brushing his teeth**.
Reference (1/5): A small boy **chewing on a blue and white toy**.

ID: COCO / COCO_val2014_000000027805.jpg

56 COCO | Pipe vs. bat



ICL-R: A man sitting on the ground holding a **baseball bat**.
Reference (1/5): A man **holding onto a pipe** with his hand out in a parking lot.

ID: COCO / COCO_val2014_000000164475.jpg

Figure 30: ICL-R failure cases on COCO captioning. The examples include object and equipment confusion, unsupported actions, and hallucinated objects. Red highlights the specific problematic phrase in the generated caption, while green highlights relevant evidence in one of the five reference captions shown for each image.