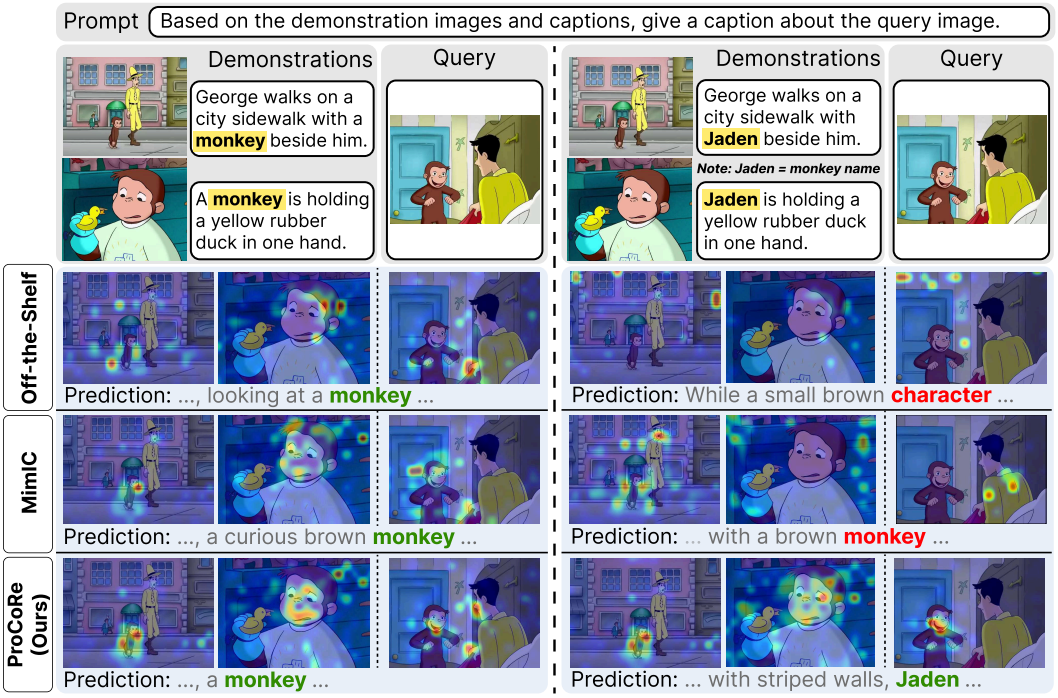
Visual Grounding First, Multimodal In-context Learning Follows

Minhyuk Seo[1,*] **Minjae Lee**[2,*] **Chaeeun Lee**[2] **Wei Lin**[4] **M. Jehanzeb Mirza**[5]
Tinne Tuytelaars[3,†] **Jonghyun Choi**[2,†]

[1] UC Berkeley [2] Seoul National University [3] KU Leuven [4] JKU [5] MIT CSAIL
https://98minjae.github.io/projects/ProCoRe/



(a) ICL w/ Generic Identifier (e.g., **monkey**) (b) ICL w/ Personalized Identifier (e.g., **Jaden**)

Figure 1: **Visual grounding in multimodal ICL with generic vs. personalized identifiers.** Attention maps show visual-token attention when predicting the highlighted token using Qwen-VL-3-8B. (a) When a generic identifier (*e.g.*, "**monkey**") seen during pre-training appears in the ICL demonstrations, both baselines and ProCoRe attend to visual regions relevant to the predicted token (*i.e.*, **monkey**). (b) In contrast, when the generic identifier is replaced with a personalized identifier (*e.g.*, "**Jaden**"), baselines fail to attend to the relevant region and thereby fall back to generic terms (*e.g.*, "**character**", "**monkey**"), whereas ProCoRe correctly predicts the personalized concept (*i.e.*, "**Jaden**") with strong visual grounding. See Sec. A.17 for more examples.

Abstract

LLMs have shown strong in-context learning (ICL) capability, but extending it to Multimodal LLMs (MLLMs) remains challenging. Prior multimodal ICL meth-

*indicates equal contribution. † indicates corresponding authors. Work done while MS is with SNU and KU Leuven.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

ods often rely on ICL-specific datasets for additional training or task vector extraction, which can improve performance on the same ICL benchmark used for adaptation. However, they do not generalize well to other benchmarks and often induce forgetting in previously well-solved tasks. For this reason, we identify 'visual grounding' as the key bottleneck in multimodal ICL; MLLMs often fail to attend to task-relevant visual evidence in demonstrations, instead over-relying on textual cues and producing hallucinated outputs. To address this, we propose ProCoRe to enhance visual grounding through contrastive reinforcement learning on multimodal contrastive data for improving multimodal ICL. ProCoRe generates captions for contrastive image pairs, decomposes them into propositions, and optimizes contrastive rewards that encourage alignment with corresponding images while discouraging mismatched ones. Despite never seeing ICL-formatted training examples, ProCoRe improves average multimodal ICL classification accuracy by 5.6% and captioning ROUGE-L by 13.8% over the off-the-shelf Qwen-VL-3-8B model, outperforming ICL-data-dependent state of the arts across benchmarks. Finally, we introduce PICL, a personalized multimodal ICL captioning benchmark that tests whether MLLMs genuinely understand visual demonstrations rather than relying on pre-trained semantic priors.

1 Introduction

In-context learning (ICL) enables models to adapt to new tasks from a few demonstrations without parameter updates [9, 14]. While this capability is well established in Large Language Models (LLMs), extending it to Multimodal LLMs (MLLMs) remains more challenging [64, 22]. The challenge is not simply that demonstrations contain images. Effective multimodal ICL requires binding visual evidence to textual outputs in demonstrations and transferring this visual–linguistic association to the query. However, recent MLLMs often fail to fully leverage visual information in demonstrations, instead over-relying on textual cues and simply mimicking the text demonstrations [12].

Prior work has largely improved multimodal ICL by selecting useful demonstrations or adapting models to ICL-formatted data. Retrieval- and ordering-based methods [6, 38, 13] retrieve and rank ICL demonstrations based on their similarity to the query. Despite their effectiveness for ICL in LLMs, they are highly sensitive to prompt design and sample order in multimodal ICL, often underperforming even random selection [63]. In contrast, ICL training-based methods tune MLLMs on ICL datasets [15, 12, 26, 22], while task vector-based methods [39, 54] use ICL data to extract task vectors that steer hidden states, replicating the effect of demonstrations. Despite their effectiveness, we observe that their reliance on ICL datasets limits their generalization (Fig. 2): they perform well on the same ICL benchmark used for training or task vector extraction, but drastically degrades on other ICL benchmarks. Moreover, training on ICL datasets can cause catastrophic forgetting on previously well-solved tasks (*e.g.*, simple VQA) (Fig. 5). More fundamentally, since ICL in LLMs is widely considered an emergent capability from pre-training rather than explicit supervision on ICL-formatted datasets [36, 40], this raises a key question: *do ICL data-dependent methods address the root cause of poor multimodal ICL, or merely improve performance on their training benchmarks?*

In this work, rather than relying on ICL benchmarks, we investigate a more fundamental limitation that hinders multimodal ICL by asking *why multimodal ICL fails and what prevents MLLMs from effectively using demonstrations*. Prior work has shown that visual hallucination [56, 30, 17], where outputs include nonexistent objects or inconsistent visual details, is a key factor behind poor multimodal ICL [51]. With further analysis, we find that *poor visual grounding*, a phenomenon attending to irrelevant regions instead of task-relevant visual evidence, significantly contributes to these hallucinations during multimodal ICL. Specifically, we reveal that this issue is particularly pronounced for personalized identifiers unseen during pre-training: while MLLMs can ground correctly on generic identifiers common in training data (Fig. 1-(a)), they struggle with personalized identifiers (Fig. 1-(b)). This limitation is critical, as real-world multimodal ICL often involves personalized references, where demonstration captions include identifiers such as a dog named "John" rather than generic concepts like "dog" [46, 47]. Since such concepts cannot be reliably inferred from pre-trained priors alone, correct predictions rely more directly on accurate visual grounding.

Motivated by this, we propose ProCoRe (**Pro**position-level **Co**ntrastive **Re**ward), which improves multimodal ICL by enhancing visual grounding. Instead of relying on ICL datasets, ProCoRe

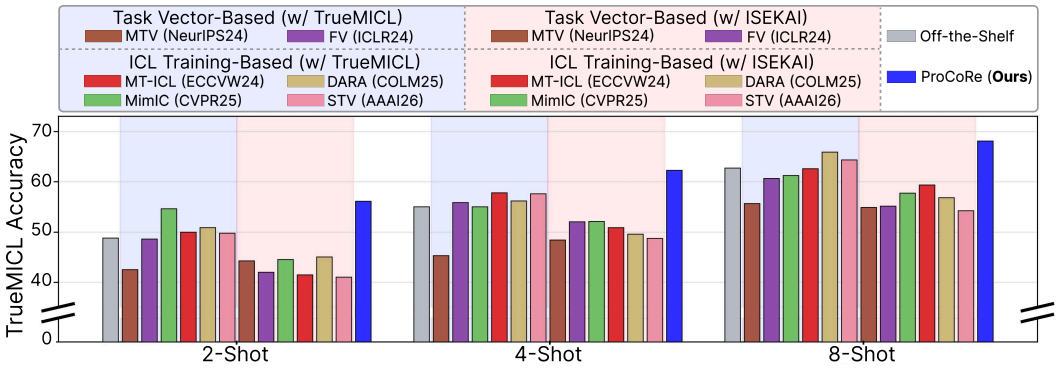


Figure 2: **ICL data-dependent methods show limited generalization.** Blue-shaded bars denote ICL data-dependent baselines trained on TrueMICL ICL dataset, while red-shaded bars denote those based on ISEKAI. They often outperform the off-the-shelf model when evaluated on the same ICL dataset used for training (*i.e.*, train: TrueMICL $\rightarrow$ test: TrueMICL), but degrade significantly when their training and test ICL datasets differ (*i.e.*, train: ISEKAI $\rightarrow$ test: TrueMICL). In contrast, even without using any ICL data, ProCoRe outperforms them through strong visual grounding.

trains MLLMs to distinguish visually contrastive samples (*e.g.*, Fig. 9), drawing inspiration from contrastive learning that fine-grained discrimination improves visual representation [11]. Specifically, it generates captions for contrastive image pairs, decomposes them into propositions (*i.e.*, distinct propositions for each caption and common propositions), and applies contrastive reinforcement learning to align each proposition with its corresponding image while discouraging alignment with mismatched images. By operating at the proposition-level rather than on sentence-level, ProCoRe enables MLLMs to capture both distinctive and shared visual features. As a result, ProCoRe improves grounding for both generic and personalized concepts (Fig. 1), and significantly enhances multimodal ICL, despite using no ICL data for training (Fig. 2). In addition, ProCoRe mitigates forgetting on general VQA tasks (Fig. 5).

Finally, to better assess whether MLLMs truly rely on multimodal demonstrations rather than pre-trained semantic priors, we introduce PICL, a personalized ICL captioning benchmark built around personalized concept names. Existing evaluations often use common concepts such as "dog" or "cat", for which correct predictions may stem from memorized pre-trained semantic priors rather than genuine understanding of the demonstrations. In contrast, by requiring models to identify personalized concepts such as a dog named "John", PICL provides a more reliable test of whether MLLMs truly understand and leverage multimodal ICL demonstrations.

Our main contributions are:

- We show that prior multimodal ICL methods suffer from limited generalization and forgetting.
- We identify weak visual grounding as a key bottleneck in multimodal ICL and propose ProCoRe, a proposition-level contrastive reward that improves grounding without ICL-formatted data.
- We introduce PICL, a personalized multimodal ICL captioning benchmark for evaluating grounded use of demonstrations.

2 Related Work

Multimodal In-context Learning. Building on the emergence of ICL in LLMs, prior work has attempted to extend ICL to MLLMs, primarily by improving the demonstrations or by adapting the model with ICL-formatted data. Retrieval- and ordering-based methods select demonstrations that are similar to the query [6, 38, 13], yet multimodal ICL remains highly sensitive to prompt design, sample order, and cross-modal similarity estimates [37, 43, 63, 18]. As a result, they often underperform random selection [26, 63]. ICL training-based methods instead expose MLLMs to ICL-formatted datasets, either through full-model tuning or additional learnable modules [15, 52, 12, 26, 22]. These methods can improve performance on the same ICL datasets used for adaptation, but often generalize poorly to unseen ICL benchmarks and can forget previously well-performed tasks (Fig. 5).

Task vector-based methods compress demonstrations into activation-space vectors that replicate the effect of steering hidden states during ICL, and apply them at inference in place of demonstrations to steer hidden states [42, 32, 54]. They reduce context and memory overhead, but potentially losing fine-grained visual information needed for complex multimodal tasks [21, 39]. Furthermore, similar to ICL training-based methods, they show limited generalization to unseen ICL benchmarks (Fig. 2). In contrast to these ICL data-dependent routes, ProCoRe targets a more fundamental bottleneck, *i.e.*, poor visual grounding, improving multimodal ICL without relying on ICL-formatted examples.

Visual Grounding and Hallucination in MLLMs. MLLMs often over-rely on language priors and textual context, producing hallucinated objects, attributes, or relations not supported by the image [56, 30, 17, 66]. This is critical for multimodal ICL: if a model does not ground generated text in relevant visual evidence, it cannot reliably leverage the visual information in demonstrations [51]. Prior hallucination-mitigation methods include improving instruction-tuning data [30, 55, 66] and adding auxiliary detectors or revisers [71, 57, 20]. Despite their effectiveness, they often require access to large-scale instruction data and incur substantial parameter and computational overhead [74].

Alternatively, visual grounding approaches [33, 16, 12], often directly amplify attention to visual tokens to counter the strong bias of MLLMs toward textual context over visual inputs. However, directly scaling visual token norms may increase attention to irrelevant visual tokens [23] and potentially distort feature representations by altering embedding magnitudes [67]. To avoid this, rather than modifying token norms, ProCoRe enhances visual grounding through a contrastive reward that promotes capturing discriminative visual cues from image–caption pairs, encouraging generated content to be supported by the corresponding image rather than by language priors alone.

See Sec. A.12 for detailed related work on multimodal ICL and hallucination mitigation methods.

3 Approach

We address multimodal ICL from the perspective of visual grounding. Our key idea is that MLLMs fail at multimodal ICL not primarily because they lack ICL-specific supervision, but because they often fail to ground demonstrations in the relevant visual evidence. The weak grounding leads to over-reliance on textual cues and leads to hallucinations. To address this, we propose ProCoRe, a simple yet effective reinforcement learning framework that improves visual grounding via contrastive rewards on visually contrastive image pairs. The key intuition is that coarse captions missing fine-grained visual details between two similar images (*e.g.*, "A hydrant on the road" in Fig. 3) can also match the wrong image, leading to lower rewards. Consequently, optimizing this objective encourages the model to capture discriminative visual details and ground generation in the image.

However, rewarding only discriminative differences can bias the model toward shortcut descriptions that ignore visual details shared across the two images (*e.g.*, "on the sidewalk next to a sign" in Fig. 3), since describing shared content increases similarity to negative samples and reduces the reward. To overcome this, ProCoRe further decomposes generated captions into propositions and applies the contrastive rewards only to image-specific propositions, while separately encouraging shared propositions to align with content common to both images. This proposition-level design allows the model to preserve complete visual descriptions while still learning to distinguish subtle differences. We provide an overview of ProCoRe in Fig. 3 and a pseudocode in Sec. A.16.

3.1 Multimodal Contrastive Training Pairs

ProCoRe is trained on visually contrastive image pairs that differ in fine-grained visual content such object identity or attributes, along with their captions, as illustrated in Fig. 3 (more in Fig. 9). Formally, each sample consists of two image-text pairs, (I_1, t_1) and (I_2, t_2), where captions t_1 and t_2 describe images I_1 and I_2, respectively. Our method is compatible with any multimodal contrastive dataset containing such image-caption pairs. In this work, we use widely used multimodal contrastive datasets, InpaintCOCO [45] and EQBEN [58], with results on other datasets in Sec. A.10.

3.2 Contrastive Reward Function

Sentence-level Contrastive Reward. We first introduce a sentence-level contrastive reward that encourages MLLMs to capture fine-grained differences between visually similar images. The in-

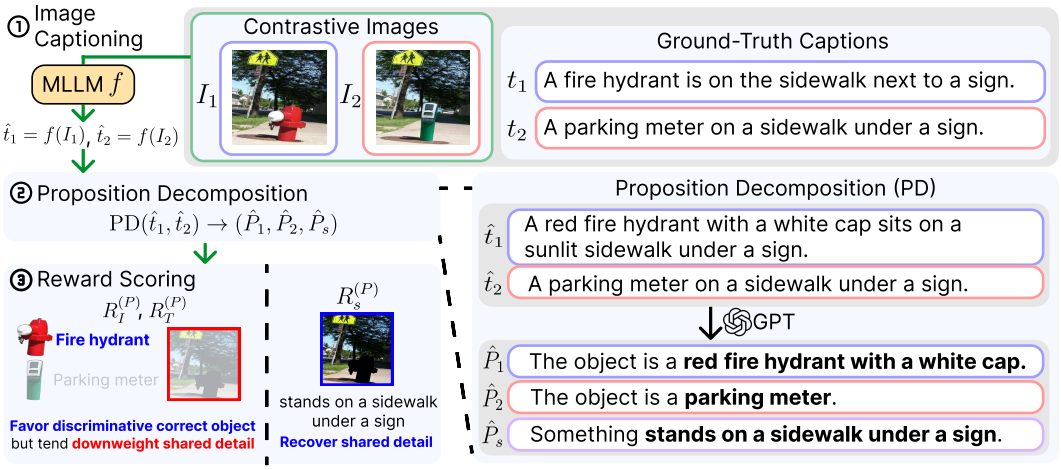


Figure 3: **Overview of ProCoRe.** (1) MLLM f predict captions $\hat{t}_1, \hat{t}_2$ for contrastive images. (2) Decompose them into propositions using ChatGPT: unique to I_1 (*i.e.*, $\hat{P}_1$), I_2 (*i.e.*, $\hat{P}_2$), or shared (*i.e.*, $\hat{P}_s$). (3) Compute proposition-level contrastive reward: R_I and R_T encourage $\hat{t}_1$ to reflect discriminative visual details (*i.e.*, fire hydrant vs. parking meter), while R_s preserves shared details.

tuition is that a caption grounded in the correct image should be more similar to its corresponding image and caption than to a non-corresponding one. By rewarding this contrast, the model is encouraged to attend to discriminative visual evidence rather than generate coarse descriptions. Formally, given a contrastive sample $\{(I_1, t_1), (I_2, t_2)\}$, an MLLM f first generates captions $\hat{t}_1 = f(I_1)$ and $\hat{t}_2 = f(I_2)$. We then define an image-text contrastive reward as:

$$R_I(I_1, I_2, \hat{t}_1, \hat{t}_2) = \sigma\left((\mathrm{sim}_I(I_1, \hat{t}_1) - \mathrm{sim}_I(I_2, \hat{t}_1)) + (\mathrm{sim}_I(I_2, \hat{t}_2) - \mathrm{sim}_I(I_1, \hat{t}_2))\right), \quad (1)$$

where σ is the sigmoid function and $\mathrm{sim}_I(I, t) = \cos(f_v(I), f_t(t))$ denotes cosine similarity between image and text embeddings produced by a vision encoder f_v and a text encoder f_t. We use CLIP-ViT-B/32 [44] as f_v and the CLIP text encoder as f_t. This reward encourages each generated caption to align with its corresponding image while discouraging alignment with the non-corresponding image.

In addition, we define a text-text contrastive reward that encourages generated captions to align with their corresponding ground-truth captions and differ from non-corresponding ones:

$$R_T(t_1, t_2, \hat{t}_1, \hat{t}_2) = \sigma\left((\mathrm{sim}_T(t_1, \hat{t}_1) - \mathrm{sim}_T(t_2, \hat{t}_1)) + (\mathrm{sim}_T(t_2, \hat{t}_2) - \mathrm{sim}_T(t_1, \hat{t}_2))\right), \quad (2)$$

where $\mathrm{sim}_T(t_i, t_j) = \cos(f_t(t_i), f_t(t_j))$ is cosine similarity between text embeddings.

The final sentence-level reward is $R = \lambda_I R_I + \lambda_T R_T$ where λ_I and λ_T control the relative importance of the two terms. We optimize this reward using Group Relative Policy Optimization (GRPO) [50].

Proposition-level Contrastive Reward. Sentence-level reward encourages discriminative grounding, but it can also bias the model toward incomplete descriptions. Since shared content appears in both images, describing it may increase similarity to the non-corresponding sample and thus reduce the reward. For example, in Fig. 3, the images differ only in the object (*i.e.*, hydrant versus parking meter), while other details such as the sidewalk and nearby sign are shared. Describing this shared scene content (*e.g.*, "on the sidewalk next to a sign") increases similarity to both images and may therefore lower the reward. As a result, the model may learn shortcut captions that mention only the distinguishing difference (*e.g.*, "The object is a hydrant") while omitting shared visual details.

To address this limitation, ProCoRe computes contrastive rewards at the proposition level rather than on full-sentence captions. Specifically, we decompose the generated captions $\hat{t}_1$ and $\hat{t}_2$ into three types of propositions: (i) $\hat{P}_1$, content unique to $\hat{t}_1$; (ii) $\hat{P}_2$, content unique to $\hat{t}_2$; and (iii) $\hat{P}_s$, content shared by both captions (see Fig. 3 for an example), using GPT-4o-mini [1]. See Sec. A.9 for the

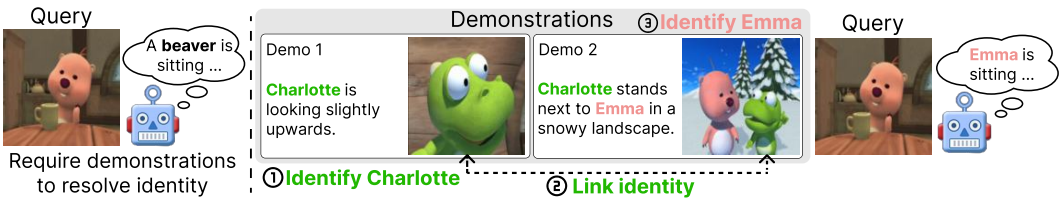


Figure 4: **Example of 2-shot PICL.** (left) Query image alone cannot infer the personalized identity. (right) (1) Identifying the dinosaur as "Charlotte" in the first demo and (2) linking it to the dinosaur in the second enables (3) inferring that the pink beaver is "Emma" in the second, allowing correct prediction of the personalized identifier for the query. See Fig. 10 and Fig. 11 for more examples.

details of decomposition process. We then compute the proposition-level contrastive reward using only the image-specific propositions:

$$R_I^{(P)}(I_1, I_2, \hat{P}_1, \hat{P}_2) = \sigma\left((\text{sim}_I(I_1, \hat{P}_1) - \text{sim}_I(I_1, \hat{P}_2)) + (\text{sim}_I(I_2, \hat{P}_2) - \text{sim}_I(I_2, \hat{P}_1))\right). \quad (3)$$

By excluding shared propositions from the contrastive terms, this formulation encourages the model to distinguish the two images without penalizing content that is true for both.

Similarly, we replace sentence-level text similarity with proposition-level entailment. Since propositions are sub-sentential units, cosine similarity between sentence embeddings is not ideal. Instead, we use an NLI-based entailment score $\text{entail}(t, \hat{P})$ [8], which measures how strongly the ground-truth caption t entails the predicted proposition $\hat{P}$. The proposition-level text reward is then defined as:

$$R_T^{(P)}(t_1, t_2, \hat{P}_1, \hat{P}_2) = \sigma\left((\text{entail}(t_1, \hat{P}_1) - \text{entail}(t_2, \hat{P}_1)) + (\text{entail}(t_2, \hat{P}_2) - \text{entail}(t_1, \hat{P}_2))\right). \quad (4)$$

To preserve shared visual details, we further introduce a shared-content reward for $\hat{P}_s$:

$$R_s^{(P)}(I_1, I_2, t_1, t_2, \hat{P}_s) = \sigma\left(\text{sim}_I(I_1, \hat{P}_s) + \text{sim}_I(I_2, \hat{P}_s) + \text{entail}(t_1, \hat{P}_s) + \text{entail}(t_2, \hat{P}_s)\right). \quad (5)$$

This term encourages $\hat{P}_s$ to capture content supported by both images and both ground-truth captions.

The final proposition-level reward used in ProCoRe is below:

$$R^{(P)} = \lambda_I R_I^{(P)} + \lambda_T R_T^{(P)} + \lambda_s R_s^{(P)}, \quad (6)$$

where λ_I, λ_T, and λ_s are weighting coefficients.

4 PICL Benchmark

Standard multimodal ICL benchmarks [12, 29, 69, 75] often involve generic identifiers (*e.g.*, 'dog', 'he') that MLLMs have likely seen during pre-training. As a result, models can answer correctly using pre-trained priors without truly leveraging demonstrations. PICL mitigates this shortcut by evaluating personalized multimodal ICL captioning, where demonstrations introduce identifiers not recoverable from the query image alone (*e.g.*, particular character names). Correct prediction requires grounding these personalized identifiers in the demonstration images and transfer that association to the query.

We construct PICL from visual storytelling datasets—PororoSV [27], FlintstonesSV [19], and StoryStream [61]—where the same characters appear across frames, each with caption. We first manually select frames to form demonstrations and queries. We then replace original identifiers in the captions with personalized identifiers, randomly sampled via GPT-4o-mini [1], requiring models to use the correct identifier when the same entity appears in the query. This benchmark thus complements standard ICL evaluations by testing whether models truly leverage visual context, rather than merely recognizing familiar categories. From each dataset (PororoSV, FlintstonesSV, and StoryStream), we form 1,000 training samples (for ICL data-dependent methods) and 100 test samples,

Category	ICL Training Dataset	Method	MS COCO					PICL				
			0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	1-shot	2-shot	3-shot	4-shot
Off-the-shelf	-	-	15.53	19.02	21.21	23.52	25.65	24.50	25.40	24.78	28.30	29.29
Retrieval and ordering	-	RICES (NeurIPS22)	15.53	19.02	21.08	23.36	25.70	24.50	25.40	26.44	27.14	28.20
		VICL (ACL24)	17.41	21.52	23.44	24.16	27.63	23.12	26.77	27.38	25.96	29.41
		MMICES (WACV25)	15.53	19.02	21.65	22.52	25.16	24.50	25.40	21.70	23.82	24.79
Task vector	MS COCO	MTV (NeurIPS24)	15.84	18.36	18.68	18.53	19.02	12.11	15.02	16.41	15.88	17.63
		FV (ICLR24)	14.56	17.97	20.65	22.93	25.33	22.31	23.88	24.12	26.72	28.41
	PICL	MTV (NeurIPS24)	9.11	14.03	13.74	15.88	12.84	9.18	14.87	13.96	15.27	13.02
		FV (ICLR24)	10.73	13.88	16.21	18.44	20.91	24.04	25.11	25.57	28.50	29.84
ICL training	MS COCO	MT-ICL (ECCVW24)	37.70	42.95	42.76	42.37	42.48	30.41	29.72	30.88	31.26	32.02
		DARA (COLM25)	25.27	29.10	30.83	32.25	38.40	22.03	23.84	23.16	25.72	26.19
		MimIC (CVPR25)	29.38	36.06	35.64	35.51	37.21	21.88	26.41	24.92	26.76	29.03
		STV (AAAI26)	38.72	44.42	42.33	43.04	42.95	29.43	30.01	30.88	31.92	32.27
	PICL	MT-ICL (ECCVW24)	34.18	39.77	40.12	39.02	39.88	32.63	32.63	32.59	33.73	34.55
		DARA (COLM25)	21.92	25.84	27.66	29.11	35.07	24.50	25.22	25.47	28.20	29.14
		MimIC (CVPR25)	25.74	32.88	31.92	31.73	33.44	24.56	28.53	27.71	28.85	31.41
		STV (AAAI26)	35.41	41.26	39.74	40.12	40.33	32.11	32.99	32.61	34.48	34.74
Mitigating hallucination	-	PAI (ECCV24)	13.92	20.11	18.84	24.33	23.07	21.74	24.62	23.18	25.11	27.89
		VOLCANO (NAACL24)	31.72	38.41	33.88	37.12	41.04	23.47	26.82	24.91	27.66	29.84
		LURE (ICLR24)	16.41	17.36	19.58	22.07	24.12	22.18	27.04	25.32	29.88	30.17
		LRV (ICLR24)	17.41	29.22	26.77	23.89	30.26	26.73	31.21	28.96	30.41	32.55
		IBD (CVPRW25)	30.94	34.28	37.51	38.67	35.88	20.33	23.91	22.48	23.76	26.12
		SPARC (ICML25)	33.18	37.73	34.26	39.84	42.11	24.91	30.12	26.44	27.03	33.08
		ProCoRe (Ours)	**39.12**	**46.88**	**46.41**	**46.26**	**46.03**	**33.54**	**36.11**	**35.89**	**36.71**	**36.76**

Table 1: **Comparison in multimodal ICL captioning benchmarks.** Best and second-best scores are bolded and underlined per dataset and shot. We use Qwen-VL-3-8B and apply random ordering of demonstrations for all baselines except retrieval- and ordering-based methods.

yielding 3,300 samples per shot setting (1-4 shot). Notably, higher-shot settings are designed to require reasoning over increasingly complex relationships among demonstrations (*e.g.*, Fig. 11), rather than simply providing more context; thus, performance does not necessarily improve with more shots. See Fig. 4 for an example (more in Sec. A.13) and Sec. A.14 for more details.

5 Experiments

5.1 Experimental Setups

Models. We use Qwen-VL-3-8B [5] as the main MLLM for all primary comparisons. We further evaluate ProCoRe across multiple MLLMs, including LLaVA-OneVision-1.5-8B [3] and InternVL-3.5-8B [59] in Sec. A.7, as well as across diverse model sizes, including 4B and 14B (Sec. A.11).

Baselines. We compare ProCoRe with SOTA multimodal ICL methods: (i) retrieval- and ordering-based methods (RICES [2], VICL [72], MMICES [13]), (ii) ICL training-based methods (STV [39], MimIC [22], MT-ICL [15], DARA [12]), and (iii) task vector-based methods (MTV [21], FV [54]). We also compare with hallucination-mitigation baselines (PAI [33], VOLCANO [25], LURE [70], LRV [30], IBD [74], SPARC [23]) to separate gains from improved visual grounding from those of generic hallucination reduction. For fair comparison, all training-based baselines (including ProCoRe) use the same number of training samples, though ICL data-dependent methods use ICL datasets while others use non-ICL data, as more detailed in Sec. A.2. The off-the-shelf model with random demonstrations is included as a reference. Further baseline details are in Sec. A.8.

Metrics. We report accuracy for multimodal ICL classification and ROUGE-L [28] for multimodal ICL captioning, as a commonly used metric for evaluating generation quality against ground truth [65, 49]. Moreover, to quantify visual grounding, we use target attention score (TAS) [41], which averages attention from each generated token to the corresponding target-region image tokens over attention heads; higher TAS indicates stronger region-token grounding. Target regions are obtained from SAM3 segmentation maps [10]. For hallucination analysis, we report OpenCHAIR scores, where lower values indicate fewer object hallucinations. See Sec. A.2 for more details of metrics.

Benchmarks. For multimodal ICL classification benchmarks, we evaluate on TrueMICL [13] and ISEKAI [52]. For multimodal ICL captioning, we employ MS COCO [29] and our proposed personalized ICL benchmark, PICL, where demonstrations introduce personalized character names that must be visually grounded. We also evaluate on MIC [69] and VL-ICL [75] in Sec. A.6.

Category	ICL Training Dataset	Method	TrueMICL					ISEKAI				
			0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	2-shot	4-shot	8-shot	12-shot
Off-the-shelf	-	-	27.11	44.79	48.83	55.03	62.70	0.16	37.98	83.68	93.79	96.42
Retrieval and ordering	-	RICES (NeurIPS22)	27.11	44.79	47.16	51.83	61.86	0.16	36.31	81.21	90.53	92.36
		VICL (ACL24)	28.22	45.44	50.01	55.12	61.80	1.12	37.81	74.64	88.93	88.41
		MMICES (WACV25)	27.11	44.79	48.40	58.07	63.04	0.16	34.55	73.09	85.27	84.87
Task vector	TrueMICL	MTV (NeurIPS24)	23.80	36.45	42.52	45.30	55.70	3.04	32.88	81.98	94.99	90.12
		FV (ICLR24)	28.82	43.88	48.61	55.87	60.65	0.96	31.58	78.78	93.19	97.92
	ISEKAI	MTV (NeurIPS24)	21.82	40.11	44.27	48.39	54.92	2.15	28.65	84.55	94.11	96.58
		FV (ICLR24)	26.44	38.76	41.98	52.02	55.18	1.21	35.44	84.72	92.18	97.03
ICL training	TrueMICL	MT-ICL (ECCVW24)	28.12	45.26	49.96	57.80	62.59	3.47	31.72	70.94	93.18	84.02
		DARA (COLM25)	29.11	45.99	50.86	56.20	65.93	0.82	27.64	79.11	84.27	91.08
		MimIC (CVPR25)	27.08	48.62	54.65	55.00	61.21	3.02	34.86	81.03	85.51	92.19
		STV (AAAI26)	29.17	45.35	49.78	57.62	64.33	**5.73**	24.39	72.58	90.92	86.44
	ISEKAI	MT-ICL (ECCVW24)	21.94	38.21	40.47	49.89	58.36	0.40	48.06	78.14	96.46	95.58
		DARA (COLM25)	24.02	39.15	44.06	48.58	55.84	1.51	40.41	75.92	87.63	88.77
		MimIC (CVPR25)	27.08	38.74	43.55	51.12	56.76	0.16	44.87	83.02	94.74	94.11
		STV (AAAI26)	19.82	37.41	40.02	47.73	53.26	2.92	49.83	86.02	96.11	95.77
Mitigating hallucination	-	PAI (ECCV24)	24.91	44.12	47.83	56.02	59.67	1.87	39.22	73.11	91.54	90.02
		VOLCANO (NAACL24)	31.41	47.76	50.68	57.43	64.12	0.21	42.08	74.96	94.73	92.11
		LURE (ICLR24)	24.08	42.31	45.92	52.77	60.18	0.16	33.74	80.21	89.46	92.87
		LRV (ICLR24)	29.02	45.11	51.94	57.27	63.46	0.40	39.77	78.11	**97.42**	95.06
		IBD (CVPRW25)	27.21	45.27	50.02	55.44	62.18	1.42	45.63	72.88	91.56	89.47
		SPARC (ICML25)	25.38	46.21	47.09	56.82	60.94	0.16	45.63	83.02	92.11	93.63
		ProCoRe (**Ours**)	**33.02**	**51.47**	**56.11**	**62.26**	**68.08**	3.94	**50.72**	**86.23**	96.18	**98.41**

Table 2: **Comparison in multimodal ICL classification benchmarks.** Best and second-best scores are bolded and underlined per dataset and shot. We use Qwen-VL-3-8B and apply random ordering of demonstrations for all baselines except retrieval- and ordering-based methods.

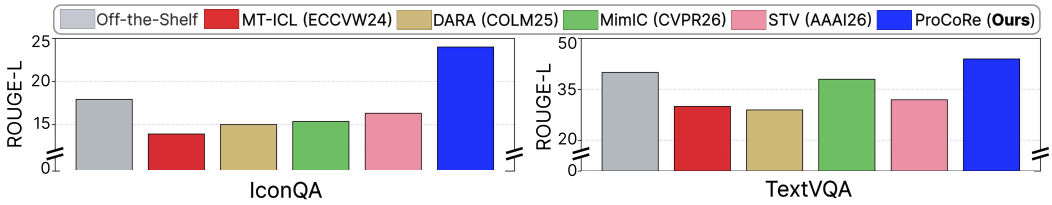


Figure 5: **Effect of ICL training on simple VQA tasks.** Additional instruction tuning on ICL datasets degrades performance on previously well-solved simple VQA tasks (e.g., IconQA and TextVQA) due to forgetting. In contrast, ProCoRe improves these tasks by enhancing visual grounding, while also achieving strong multimodal ICL performance (Tab. 2).

5.2 Quantitative Analysis

Multimodal ICL Captioning Tasks. Table 1 evaluates whether stronger visual grounding improves ICL captioning. As shown, ProCoRe significantly outperforms baselines across MS COCO and PICL at all settings, despite using no ICL-formatted training data. These results support our hypothesis: strong visual grounding elicits stronger ICL capability without directly fitting to specific ICL benchmarks. In contrast, ICL data-dependent methods perform well on the benchmark used for training or task-vector extraction, but degrades significantly on other ICL datasets, often falling below the off-the-shelf model, indicating limited generalization. Retrieval- and ordering-based methods often underperform the random selection (*i.e.*, off-the-shelf model), as also observed in prior work [63, 26].

Moreover, hallucination-mitigation baselines show only marginal gains and often underperform the off-the-shelf model. We attribute this to their approach of enhancing visual grounding through direct amplification of visual attention norms, which fails to focus on task-relevant regions (Sec. A.17). In contrast, ProCoRe naturally enhances visual grounding through a contrastive reward, resulting in more accurate visual grounding.

Multimodal ICL Classification Tasks. We also evaluate ProCoRe on ICL classification tasks and report results in Tab. 2. As shown in the table, consistent with the ICL captioning task, ProCoRe significantly outperforms baselines across all shot settings, demonstrating that ProCoRe generalizes strongly across diverse ICL tasks, while ICL data-dependent baselines continue to show limited generalization. We provide additional ICL classification results in Sec. A.6.

Category	Method	PICL-PororoSV			
		1-shot	2-shot	3-shot	4-shot
Off-the-shelf	-	38	21	29	31
Retrieval and ordering	RICES (NeurIPS22)	33	27	29	35
	VICL (ACL24)	36	28	30	35
	MMICES (WACV25)	30	24	27	32
Task vector	MTV (NeurIPS24)	34	13	26	25
	FV (ICLR24)	36	22	28	32
ICL training	MT-ICL (ECCVW24)	54	54	56	**59**
	DARA (COLM25)	43	37	41	47
	MimIC (CVPR25)	51	43	44	49
	STV (AAAI26)	59	52	55	56
Mitigating hallucination	PAI (ECCV24)	26	17	25	27
	VOLCANO (NAACL24)	54	41	47	55
	LURE (ICLR24)	24	19	24	28
	LRV (ICLR24)	42	29	25	34
	IBD (CVPRW25)	50	45	46	42
	SPARC (ICML25)	54	42	51	55
	ProCoRe (Ours)	**62**	**57**	**57**	**59**

Table 3: **Comparison of personalized identifier accuracy.** We measure whether predicted captions correctly match personalized identifiers in the ground-truth captions for each query. ProCoRe achieves the highest accuracy across all shots.

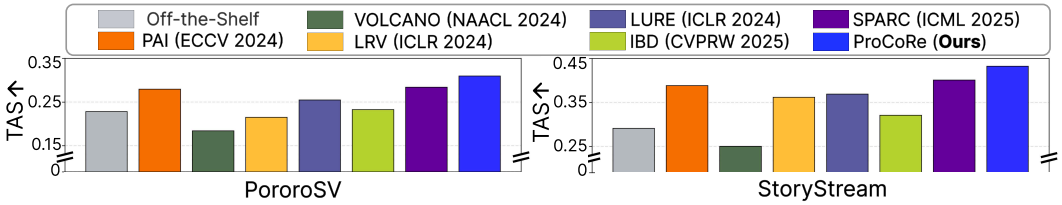


Figure 6: **Comparison of target attention score (TAS) on multimodal ICL captioning benchmarks.** While hallucination-mitigation baselines sometimes enhance TAS compared to off-the-shelf model, our proposed ProCoRe significantly outperforms both off-the-shelf model and baselines.

Personalized Identifier Accuracy. We further provide further analysis of Tab. 1. Importantly, PICL shows high zero-shot performance, even though correct answers require visual grounding of demonstrations in our proposed PICL. This is because we use ROUGE-L for captioning evaluation: even if the personalized identifier is incorrect, models can still achieve high scores by correctly describing generic visual content (*e.g.*, actions or scenes). Despite this limitation, ROUGE-L remains appropriate as captioning evaluation not only require evaluating correct identifier usage but also the overall quality of visual descriptions.

However, to isolate personalized identifier accuracy from overall description quality, we use GPT to evaluate whether generated captions correctly match the identifiers in the ground-truth captions. See Sec. A.1 for the prompts used for the evaluation.

As shown in Tab. 3, ProCoRe achieves the highest personalized identifier accuracy across all shots. Together with its ROUGE-L gains (Tab. 1), this indicates that ProCoRe better grounds personalized identifiers, leading to more accurate descriptions of both the target entity and other visual details.

Visual Grounding and Hallucination Evaluation. Next we test whether the ICL gains in ProCoRe come with strong visual grounding. Visual attention map (Fig. 1 and Fig. 12) shows that ProCoRe attends to relevant visual regions when generating response tokens. Moreover, Fig. 6 shows that ProCoRe achieves the highest TAS, indicating stronger focus on task-relevant regions.

This improved visual grounding mitigates hallucination in multimodal ICL. As shown in Fig. 7, ProCoRe consistently shows lower OpenCHAIR scores than baselines. We believe that this reduction in

Method	MS COCO					PICL				
	0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	1-shot	2-shot	3-shot	4-shot
Off-the-Shelf (Qwen-VL-3-8B)	15.53	19.02	21.21	23.52	25.65	24.50	25.40	24.78	28.30	29.29
SFT	19.48	27.92	30.11	26.84	37.62	22.13	27.88	23.41	30.74	26.81
Sentence-Level Contrastive Reward	36.02	45.73	43.21	44.11	42.58	32.88	34.74	31.92	34.02	33.41
ProCoRe (**Ours**)	**39.12**	**46.88**	**46.41**	**46.26**	**46.03**	**33.54**	**36.11**	**34.89**	**36.71**	**36.76**

Table 4: **Ablations of ProCoRe.** Gains in ProCoRe stem from proposition-level contrastive rewards, not contrastive-pair training it self, with clear improvements over sentence-level rewards.

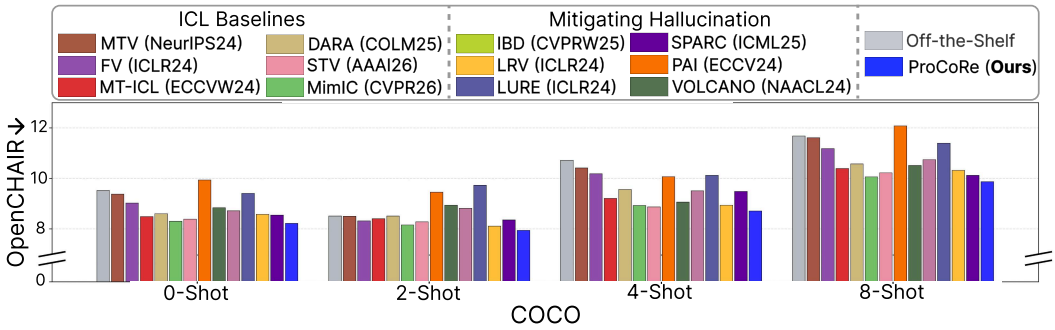


Figure 7: **Comparison of hallucinations on MS COCO ICL captioning benchmark.** ProCoRe consistently reduces hallucinations relative to baselines across different numbers of demonstrations.

hallucination, driven by strong visual grounding, contributes to strong multimodal ICL performance. We provide further visual grounding analysis in Sec. A.17.

General VQA Evaluation. We test whether general vision–language ability is preserved. Figure 5 shows that ProCoRe improves standard VQA performances, whereas ICL-training baselines often degrade. This suggests that ProCoRe does not trade general vision-language ability for benchmark-specific ICL behavior, but instead improves the fundamental visual grounding ability of MLLMs.

Analysis across Diverse MLLMs and Model Sizes. In addition to Qwen-VL-3 [5], we compare ProCoRe with baselines across multiple MLLMs, including LLaVA-OneVision-1.5 [3] and InternVL-3.5 [59] in Sec. A.7, as well as 4B and 13B scales in Sec. A.11, showing that ProCoRe consistently achieves strong multimodal ICL performance across models and scales.

Additional Experiments and Analysis. Additional results include: ProCoRe across diverse contrastive benchmarks (Sec. A.10) and hyperparameter analysis (Sec. A.3).

Ablation Study. We analyze the contribution of ProCoRe by comparing to (i) supervised fine-tuning (SFT) on contrastive pairs to disentangle gains beyond those achievable from contrastive-pair training alone, (ii) sentence-level contrastive rewards, to assess the effect of proposition-level decomposition in ProCoRe. As shown in Tab. 4, SFT with contrastive pairs shows only marginal gains, indicating that improvements in ProCoRe stem from the proposed contrastive reward rather than contrastive-pair training itself. Moreover, the significant gap between ProCoRe and sentence-level rewards highlights the effectiveness of proposition-level contrastive rewards.

6 Conclusions

We show that existing multimodal ICL data-dependent methods suffer from limited generalization and catastrophic forgetting. Rather than relying on ICL benchmarks, we address a more fundamental issue: weak visual grounding in MLLMs. To this end, we propose ProCoRe, a grounding-first approach that strengthens visual–textual associations without using ICL-formatted data. By learning from visually contrastive pairs with proposition-level rewards, ProCoRe promotes fine-grained grounding without shortcut descriptions. Across ICL benchmarks, ProCoRe improves performance and generalizes beyond the training data of ICL data-dependent methods. These gains are accompanied by stronger target-region attention, reduced hallucinations, and less forgetting on standard VQA, supporting our central claim: stronger visual grounding can elicit multimodal ICL without explicit ICL training.

Acknowledgement

This work was partly supported by AICA GPU support program, the IITP grants (RS-2022-II220077, RS-2022-II220113, RS-2022-II220959, RS-2026-25507282, RS-2026-25518317, RS-2021-II211343 (SNU AI), RS-2025-25442338 (AI Star Fellowship-SNU)) funded by the Korea government (MSIT), and the BK21 FOUR program, SNU in 2025.

This work was also partly supported by the AI Seoul Tech Research Support Program of the Seoul Future Foundation and European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 101021347). We also acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputers MareNostrum5 at BSC, Spain; LEONARDO at CINECA, Italy; VEGA at IZUM, Slovenia; Karolina at IT4Innovations, Czech Republic; MeluXina at LuxProvide, Luxembourg; Discoverer at Sofia Tech Park, Bulgaria; and Deucalion at Minho Advanced Computing Centre, Portugal, under project IDs EHPC-DEV-2026D06-170 and EHPC-DEV-2026D04-219, through EuroHPC Development Access calls.

This work was done while M. Jehanzeb was affiliated with MIT CSAIL; at the time of publication, he is affiliated with Xero. Jonghyun Choi is with ECE, IPAI, ASI in SNU.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] 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.

[3] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

[4] Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. Vismin: Visual minimal-change understanding. *Advances in Neural Information Processing Systems*, 37:107795–107829, 2024.

[5] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

[7] Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations. In *EMNLP*, 2024.

[8] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing*, 2015. URL `https://api.semanticscholar.org/CorpusID:14604520`.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[10] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[11] Delong Chen, Zhao Wu, Fan Liu, Zaiquan Yang, Shaoqiu Zheng, Ying Tan, and Erjin Zhou. Protoclip: Prototypical contrastive language image pretraining. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):610–624, 2023.

[12] Shuo Chen, Jianzhe Liu, Zhen Han, Yan Xia, Daniel Cremers, Philip Torr, Volker Tresp, and Jindong Gu. True multimodal in-context learning needs attention to the visual context. In *COLM*, 2024.

[13] Shuo Chen, Zhen Han, Bailan He, Jianzhe Liu, Mark Buckley, Yao Qin, Philip Torr, Volker Tresp, and Jindong Gu. Can multimodal large language models truly perform multimodal in-context learning? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6000–6010. IEEE, 2025.

[14] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 1107–1128, 2024.

[15] Sivan Doveh, Shaked Perek, M Jehanzeb Mirza, Wei Lin, Amit Alfassy, Assaf Arbelle, Shimon Ullman, and Leonid Karlinsky. Towards multimodal in-context learning for vision and language models. In *European Conference on Computer Vision*, pages 250–267. Springer, 2024.

[16] Jun Gao, Qian Qiao, Tianxiang Wu, Zili Wang, Ziqiang Cao, and Wenjie Li. Aim: Let any multimodal large language models embrace efficient in-context learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[17] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI*, 2024.

[18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[19] Tanmay Gupta, Dustin Schwenk, Ali Farhadi, Derek Hoiem, and Aniruddha Kembhavi. Imagine this! scripts to compositions to videos. In *Proceedings of the European conference on computer vision (ECCV)*, 2018.

[20] Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30787–30796, 2025.

[21] Brandon Huang, Chancharik Mitra, Leonid Karlinsky, Assaf Arbelle, Trevor Darrell, and Roei Herzig. Multimodal task vectors enable many-shot multimodal in-context learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL `https://openreview.net/forum?id=W0okTgsPvM`.

[22] Yuchu Jiang, Jiale Fu, Chenduo Hao, Xinting Hu, Yingzhe Peng, Xin Geng, and Xu Yang. Mimic in-context learning for multimodal tasks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29825–29835, 2025.

[23] Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. In *International Conference on Machine Learning*, pages 28527–28551. PMLR, 2025.

[24] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[25] Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *NAACL*, pages 391–404, 2024.

[26] 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. *arXiv preprint arXiv:2504.04633*, 2025.

[27] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6329–6338, 2019.

[28] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013/`.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[30] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*, 2024.

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

[32] Sheng Liu, Haotian Ye, Lei Xing, and James Y Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *ICML*, 2024.

[33] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024.

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[35] Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. Evaluation and enhancement of semantic grounding in large vision-language models. *arXiv preprint arXiv:2309.04041*, 2023.

[36] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5098–5139, 2024.

[37] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, 2022.

[38] Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. How does the textual information affect the retrieval of multimodal in-context learning? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5321–5335, 2024.

[39] 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, pages 7892–7900, 2026.

[40] Sherin Muckatira, Vijeta Deshpande, Vladislav Lialin, and Anna Rumshisky. Emergent abilities in reduced-scale generative language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1242–1257, 2024.

[41] Siqu Ou, Tianrui Wan, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Do mllms really see it: Reinforcing visual attention in multimodal llms. *arXiv preprint arXiv:2602.08241*, 2026.

[42] 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.

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

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[45] Philipp J Rösch, Norbert Oswald, Michaela Geierhos, and Jindřich Libovickỳ. Enhancing conceptual understanding in multimodal contrastive learning through hard negative samples. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 102–115, 2024.

[46] Soroush Seifi, Vaggelis Dorovatas, Daniel Olmeda Reino, and Rahaf Aljundi. Personalization toolkit: Training free personalization of large vision language models. *arXiv preprint arXiv:2502.02452*, 2025.

[47] Soroush Seifi, Simon Gardier, Vaggelis Dorovatas, Daniel Olmeda Reino, and Rahaf Aljundi. Ego: Embedding-guided personalization of vision-language models. *arXiv preprint arXiv:2603.09771*, 2026.

[48] Minhyuk Seo, Seongwon Cho, Minjae Lee, Diganta Misra, Hyeonbeom Choi, Seon Joo Kim, and Jonghyun Choi. GenOL: Generating diverse examples for name-only online learning. *Transactions on Machine Learning Research*, 2025.

[49] Minhyuk Seo, Taeheon Kim, Hankook Lee, Jonghyun Choi, and Tinne Tuytelaars. Co-loRA: Collaborative model personalization on heterogeneous multi-modal clients. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=0g5Dk4Qfh0`.

[50] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[51] Mustafa Shukor, Alexandre Rame, Corentin Dancette, and Matthieu Cord. Beyond task performance: evaluating and reducing the flaws of large multimodal models with in-context-learning. In *ICLR*, 2024.

[52] Yan Tai, Weichen Fan, Zhao Zhang, and Ziwei Liu. Link-context learning for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27176–27185, 2024.

[53] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[54] Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau. Function vectors in large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[55] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. In *AAAI*, 2024.

[56] Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

[57] Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer, 2024.

[58] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11998–12008, 2023.

[59] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[60] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10637–10647, 2023.

[61] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Ying-Cong Chen. Seed-story: Multimodal long story generation with large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1850–1860, 2025.

[62] Xu Yang, Yongliang Wu, Mingzhuo Yang, Haokun Chen, and Xin Geng. Exploring diverse in-context configurations for image captioning. *NeurIPS*, 2023.

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

[64] Chao Yi, Yuhang He, De-Chuan Zhan, and Han-Jia Ye. Bridge the modality and capability gaps in vision-language model selection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=01qa1ZJs65.

[65] yiyuan yang, Guodong Long, Tao Shen, Jing Jiang, and Michael Blumenstein. Dual-personalizing adapter for federated foundation models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[66] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024.

[67] Yanhong Zeng, Huan Yang, Hongyang Chao, Jianbo Wang, and Jianlong Fu. Improving visual quality of image synthesis by a token-based generator with transformers. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=lGoKo9WS2A_.

[68] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[69] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. MMICL: Empowering vision-language model with multi-modal in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5KojubHBr8.

[70] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[71] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*, 2024.

[72] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15890–15902, 2024.

[73] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*, 2024.

[74] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633, 2025.

[75] Yongshuo Zong, Ondrej Bohdal, and Timothy Hospedales. VL-ICL bench: The devil in the details of multimodal in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=cpGPPLLYYx.

A Technical Appendices and Supplementary Material

A.1 Additional Analyses on Multimodal ICL Captioning Tasks

To isolate personalized identifier accuracy from overall description quality, we use GPT to evaluate whether generated captions correctly match the identifiers in the ground-truth captions. We summarize the results in (Tab. 3). The evaluation prompt used is as follows:

```
You will be given:
1. A predicted caption
2. A ground-truth (GT) caption


Your task is to identify the characters described in the GT caption and
determine whether the predicted caption correctly assigns and uses the
character names.


Evaluation instructions:


1. Carefully understand the GT caption:
- Identify all characters mentioned or implied.
- Identify which actions, attributes, clothing, positions, or interactions
correspond to each character.


2. Analyze the predicted caption:
- Determine which predicted character corresponds to which GT character.
- Check whether the predicted character names are correctly assigned.


3. A prediction is considered incorrect if:
- A character is given the wrong name.
- Two characters' names are swapped.
- A character from the GT caption is missing.
- An extra unsupported character name is introduced.

4. A prediction is considered correct only if:
- Every character identity matches the GT caption.
- The names are assigned to the correct corresponding characters.
```

Human Validation of the GPT-4o-mini Identifier-Accuracy Metric. We perform a human-agreement study to validate that the GPT-4o-mini-based identifier-accuracy metric accurately evaluates personalized identifier binding. Specifically, we randomly sample 30 query examples and present each ground-truth/predicted caption pair to 50 independent human annotators through a questionnaire. Consistent with the GPT-4o-mini evaluation protocol, human evaluators are provided only with the ground-truth and predicted captions, without access to the corresponding query images. For each example, human evaluators answer the question: "Are all personalized identifiers correctly assigned?" with either Correct or Incorrect. We then evaluate the same caption pairs using our GPT-4o-mini-based identifier-accuracy metric and compare its predictions with the human judgments.

Metric	Mean±SD	Median
Agreement (%)	95.91±2.98	96.67
Cohen's κ	0.92±0.08	0.93
Precision (%)	96.97±4.72	100
Recall (%)	95.43±5.00	100
F1-score (%)	95.97±3.88	96.77
Raw score (/30)	28.77±1.19	29

Table 5: Human agreement study on personalized identifier accuracy.

We summarize the results in Tab. 5, reporting (i) Agreement: the percentage of examples on which the GPT-4o-mini-based identifier-accuracy metric agrees with the majority human judgment, (ii)

Cohen's κ: a standard measure of inter-rater agreement that quantifies the consistency between GPT-4o-mini and human judgments beyond chance agreement, and (iii) precision, recall, and F1-score. As shown, the GPT-4o-mini-based metric achieves high agreement with human judgments across all metrics, demonstrating that it provides a reliable evaluation of personalized identifier binding.

A.2 Details of Experimental Setups

Metrics. To quantify visual grounding, we examine whether MLLMs attend to relevant regions during token generation. To this end, we use target attention score (TAS) [41] measures the average attention weight assigned across all attention heads from each generated token to image tokens corresponding to a target region, where the region is defined by a segmentation map obtained from SAM3 [10]. Formally, let $\alpha_{t_i,t_g}^{(h)}$ denote the attention weight from the generated token t_g to image token t_i at attention head h, and let $\mathcal{T}_{\text{target}}$ denote the set of image tokens corresponding to the target region associated with t_g. Then, TAS a is defined as follows:

$$a = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{|\mathcal{T}_{\text{target}}|} \sum_{t_i \in \mathcal{T}_{\text{target}}} \alpha_{t_i,t_g}^{(h)}. \tag{7}$$

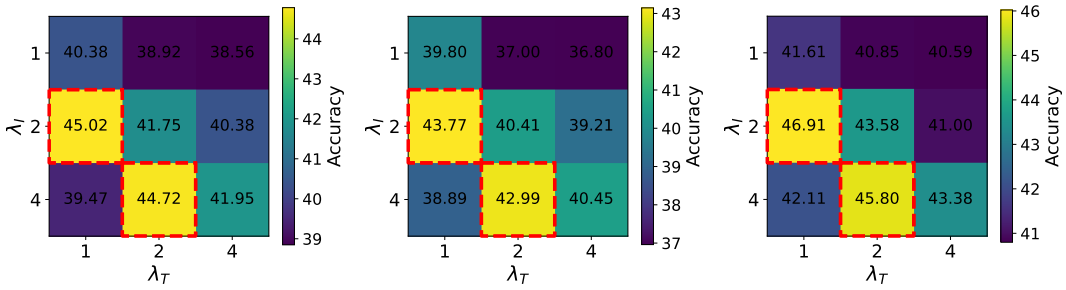
Moreover, to quantitatively evaluate visual hallucination, we use the OpenCHAIR metric [7], which decomposes predicted captions into words and measures the proportion of hallucinated words by checking whether each predicted word is supported by the ground-truth caption, considering synonyms through LLMs.

Implementation Details and Hyperparameters. We set the batch size to 16, the learning rate for Qwen-VL-3 and LLaVA-OneVision-1.5 to 2×10^{-5}, and InternVL to 2×10^{-6}, the Cosine LR scheduler and the Adam optimizer [34], and LoRA rank r of 128 attached to all linear layers in the LLM backbone for all baselines require training, such as ICL data-dependent baselines and hallucination mitigation baselines (including ProCoRe). For ProCoRe, we set the reward weighting coefficients as $\lambda_I = 2$, $\lambda_T = 1$, and $\lambda_s = 4$ in all experiments. See Sec. A.3 for detailed hyperparameter analysis. All experiments are executed in Python 3.10, on four Ubuntu 20.04 machines, with 8 NVIDIA RTX A6000 GPUs each. Each experiment runs on a single RTX A6000 GPU. We report single-run results, since repeated ICL with the same inputs yields low variance in MLLM outputs [2, 38]. For trainable baselines, we use 1,000 training samples for fair comparison, following the training datasets used in each method's original paper: ICL-dependent methods sample from the evaluated ICL benchmark; LURE uses MS COCO [29], VOLCANO uses LLaVA-SFT-127k [31], LRV uses Visual Genome [24], and ProCoRe uses 1,000 samples from each of InpaintCOCO [45] and EQBEN [58].

A.3 Hyperparameter Analysis

We analyze the effect of reward weights λ_I, λ_T, and λ_s, which weight the image contrastive reward $R_I^{(p)}$, text contrastive reward $R_T^{(p)}$, and shared-content reward $R_s^{(p)}$, respectively. Since these coefficients control relative contributions of each reward term, their ratios, rather than absolute values, are key. Therefore, we fix one weight λ_s and vary the others (*i.e.*, λ_I, λ_T) to study the effect of different relative ratios.

As shown in Fig. 8, placing more weight on the text contrastive reward than the image contrastive reward (*i.e.*, $\lambda_T > \lambda_I$, strictly upper triangle in the heatmap) leads to lower performance. This is because it may bias the reward toward over-reliance on ground-truth captions in contrastive pairs, encouraging simple imitation of the corresponding caption rather than enhancing visual grounding. Conversely, overly emphasizing the image contrastive reward (*i.e.*, $\lambda_I = 4, \lambda_T = 1$) can underutilize useful details from ground-truth descriptions, also leading to performance drops. Considering this trade-off and our goal of improving visual grounding, we set $\lambda_I > \lambda_T$ while keeping their ratio moderate to avoid overemphasis. Specifically, we use $\lambda_I = 2$, $\lambda_T = 1$, and $\lambda_s = 4$ across all experiments.



(a) 1-shot ICL Accuracy (b) 4-shot ICL Accuracy (c) 8-shot ICL Accuracy

Figure 8: **Accuracy heatmap over reward weights (λ_I, λ_T) with $\lambda_s = 4$.** If the text reward weight (λ_T) outweighs the visual contrastive reward weight (λ_I), the model over-focuses on textual demonstrations, weakening visual grounding and degrading performance. Conversely, if λ_I is too large compared to λ_T, the model underutilizes useful information from ground-truth captions, also leading to performance drops. Moderate ratios that emphasize visual rewards while retaining text signals (*e.g.*, $\lambda_I/\lambda_T = 2$, red dotted box) consistently show strong performance across shots. We evaluate ICL performance on the MS COCO captioning benchmark, with InpaintCOCO used for contrastive RL training in ProCoRe.

A.4 Additional Details and Analysis of Proposition Decomposition

Detailed Procedure. Proposition decomposition with GPT-4o-mini is performed online during RL training. More concretely, at each GRPO step, the MLLM first generates captions for the contrastive image pair. We then send the generated captions together with the corresponding ground-truth captions to the GPT-4o-mini API to decompose them into propositions.

Computational Cost. Importantly, the decomposition operates only on textual captions (*i.e.*, the generated and ground-truth captions), without requiring any image inputs, resulting in minimal computational and API overhead. The additional API cost is approximately 4×10^{-4} USD per training sample, while the wall-clock training time increases by only 5.32%. This modest overhead is because the API calls are lightweight text-only requests and can be executed in parallel with GRPO rollout generation using multiple threads.

Alternative Proposition Decomposers. We emphasize that proposition decomposition is a lightweight text-only preprocessing step that simply partitions the generated and ground-truth captions into unique and shared propositions. Therefore, GPT-4o-mini is not essential to our method and can be readily replaced by much smaller models with comparable performance. As shown in Tab. 6 and Tab. 7, small open-source LLMs (*e.g.*, Qwen3-1.7B, LLaMA3.2-3B) achieve comparable performance. This demonstrates that the performance gains stem from the proposition decomposition itself, rather than from the high capability of GPT-4o-mini.

Proposition Decomposer	0-shot	1-shot	2-shot	4-shot
GPT	33.04±1.52	35.82±1.44	35.31±1.88	35.30±1.19
Qwen3-1.7B	32.55±1.30	32.74±2.30	36.74±1.29	35.20±1.66
LLaMA3.2-3B	32.32±1.41	34.44±1.72	36.87±1.66	37.60±2.15
Qwen3-4B	31.88±1.14	34.97±1.87	37.62±2.05	34.41±0.96

Table 6: Comparison of different LLM-based proposition decomposers in ProCoRe on PICL.

A.5 Detailed ablation study of ProCoRe's contrastive reward

We conduct detailed ablation studies with the following variants: (i) Image contrastive only (*i.e.*, $\lambda_T = 0, \lambda_s = 0$ in Eq. 6), (ii) Text contrastive only (*i.e.*, $\lambda_I = 0, \lambda_s = 0$ in Eq. 6), (iii) No shared-content reward (*i.e.*, $\lambda_s = 0$) and (iv) No proposition decomposition (*i.e.*, sentence-level contrastive reward). We summarize the results in Tab. 8 and Tab. 9. As shown, using either the contrastive

Proposition Decomposer	0-shot	1-shot	2-shot	4-shot
GPT	32.44±1.89	50.91±2.74	55.44±1.59	61.37±1.68
Qwen3-1.7B	31.98±1.79	53.90±1.34	54.04±1.34	63.73±1.45
LLaMA3.2-3B	31.32±1.72	50.02±1.56	53.33±1.60	60.90±1.69
Qwen3-4B	33.47±1.42	50.26±2.18	54.91±0.89	62.38±1.67

Table 7: Comparison of different LLM-based proposition decomposers in ProCoRe on TrueMICL.

image reward or the contrastive text reward alone yields only marginal improvements over the off-the-shelf model, while combining both rewards leads to substantially better performance. These results indicate that jointly encouraging the model to attend to discriminative visual evidence and to generate semantically faithful descriptions is important for improving multimodal ICL.

Furthermore, incorporating the shared-content reward consistently improves performance, suggesting that explicitly rewarding propositions shared across the contrastive pair prevents the model from focusing solely on discriminative differences and encourages it to capture visual attributes common to both images. Finally, replacing proposition-level decomposition with a sentence-level contrastive reward leads to a noticeable performance drop, highlighting the importance of fine-grained proposition-level supervision. Together, these ablations demonstrate the complementary contributions of the individual reward components and provide further support for the design of our proposition-level contrastive reward.

Method	1-shot	2-shot	3-shot	4-shot
Off-the-shelf	23.33	27.00	31.67	29.00
SFT	9.67	9.33	15.67	10.33
No-proposition-decomposition	42.67	38.67	40.00	32.00
Image contrastive only ($\lambda_T = 0, \lambda_s = 0$)	40.00	42.67	37.67	30.33
Text contrastive only ($\lambda_I = 0, \lambda_s = 0$)	33.00	30.33	31.67	32.00
No shared-content reward ($\lambda_s = 0$)	53.33	38.00	44.00	39.67
ProCoRe (Ours)	**62.67**	**58.67**	**53.00**	**50.33**

Table 8: **Ablation study of ProCoRe's individual reward components on PICL using personalized identifier accuracy.**

Method	0-shot	1-shot	2-shot	4-shot	8-shot
Off-the-shelf	15.53	19.02	21.21	23.52	25.65
SFT	19.48	27.92	30.11	26.84	37.62
No-proposition-decomposition	36.02	44.73	43.21	44.11	42.58
Image contrastive only ($\lambda_T = 0, \lambda_s = 0$)	22.91	35.44	29.37	33.88	31.26
Text contrastive only ($\lambda_I = 0, \lambda_s = 0$)	24.12	36.91	27.54	34.26	28.93
No shared-content reward ($\lambda_s = 0$)	30.06	30.58	30.11	35.49	37.82
ProCoRe (Ours)	**39.12**	**46.88**	**46.41**	**46.26**	**46.03**

Table 9: **Ablation study of ProCoRe's individual reward components on TrueMICL using ROUGE-L.**

A.6 Additional Experiments on Multimodal ICL Classification Tasks

In addition to TrueMICL and ISEKAI, we also evaluate on MIC [69] and VL-ICL [75], and summarize the results in Tab. 10. ICL data-dependent baselines (*e.g.*, task-vector and ICL-training methods) sometimes outperform ProCoRe on the benchmarks used for their ICL training or extraction, but degrade drastically on others, indicating poor generalization.

Category	ICL Training Dataset	Method	VL-ICL					MIC				
			0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	1-shot	2-shot	4-shot	8-shot
Off-the-shelf	-	-	19.29	45.13	59.25	57.42	63.25	16.49	24.90	30.23	35.48	42.46
Retrieval and ordering	-	RICES (NeurIPS22)	19.29	45.13	44.67	53.13	57.88	16.49	24.90	30.70	37.32	45.48
		VICL (ACL24)	15.84	41.72	41.03	53.68	58.21	14.11	27.42	32.88	33.21	45.03
		MMICES (WACV25)	19.29	45.13	44.08	53.33	59.84	16.49	24.90	30.68	36.62	41.94
Task vector	VL-ICL	MTV (NeurIPS24)	20.96	37.79	38.92	44.04	48.88	33.74	35.21	36.02	37.88	38.41
		FV (ICLR24)	20.11	46.02	60.47	56.88	64.73	14.11	22.03	26.18	30.12	34.27
	MIC	MTV (NeurIPS24)	16.42	36.11	34.28	42.63	47.21	38.14	39.97	40.46	41.99	42.95
		FV (ICLR24)	15.92	41.18	55.84	52.11	59.27	18.78	26.77	30.97	34.69	39.38
ICL training	VL-ICL	MT-ICL (ECCVW24)	23.18	47.92	48.63	49.11	50.02	35.42	32.03	31.88	34.01	43.92
		DARA (COLM25)	19.29	47.27	61.41	56.06	64.72	16.49	21.88	24.17	29.04	38.11
		MimIC (CVPR25)	23.61	46.55	60.12	**62.48**	64.09	22.88	19.21	20.11	21.04	35.18
		STV (AAAI26)	21.07	48.63	59.77	59.34	**68.11**	36.41	33.18	32.74	34.21	43.03
	MIC	MT-ICL (ECCVW24)	20.41	32.88	49.21	47.63	54.02	49.01	54.41	55.18	55.28	55.84
		DARA (COLM25)	17.84	34.11	45.73	41.92	50.88	16.49	44.18	46.72	41.63	49.91
		MimIC (CVPR25)	22.03	36.74	50.11	46.28	53.42	36.02	41.92	42.03	42.85	46.38
		STV (AAAI26)	19.18	37.21	47.88	48.03	56.62	49.52	55.14	**55.30**	55.53	55.38
Mitigating hallucination	-	PAI (ECCV24)	15.94	40.88	55.73	53.62	59.11	30.18	35.41	36.74	37.92	40.11
		VOLCANO (NAACL24)	20.88	46.72	60.41	58.03	65.74	38.74	44.11	39.88	46.21	49.03
		LURE (ICLR24)	17.02	42.51	54.88	52.77	60.42	34.18	39.74	43.92	41.27	45.11
		LRV (ICLR24)	18.41	47.09	61.02	59.11	64.38	33.41	39.12	45.28	40.74	43.91
		IBD (CVPRW25)	18.37	46.44	60.03	55.88	64.71	39.27	43.83	40.11	45.63	47.88
		SPARC (ICML25)	21.67	44.58	58.12	56.04	66.03	32.18	38.74	44.62	41.92	50.21
		ProCoRe	**25.91**	**49.88**	**62.54**	61.47	67.03	**52.11**	**58.29**	53.88	**58.01**	**56.63**

Table 10: **Quantitative comparison on the VL-ICL benchmark and MIC.** Best results for each shot are highlighted in bold, while second-best results are underlined. All methods use Qwen-VL-3-8B as the base MLLM. Demonstrations are randomly ordered for all methods except retrieval- and ordering-based approaches.

Category	ICL Training Dataset	Method	MS COCO					PICL				
			0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	1-shot	2-shot	3-shot	4-shot
Off-the-shelf	-	Random	26.92	31.22	33.25	35.83	36.26	18.44	21.86	24.56	25.84	28.01
Retrieval and ordering	-	RICES (NeurIPS22)	26.92	31.22	31.84	33.70	35.41	18.44	21.86	25.01	26.31	27.08
		VICL (ACL24)	27.14	32.81	35.02	37.06	36.41	18.72	23.11	23.56	24.84	27.01
		MMICES (WACV25)	26.92	31.22	29.62	33.71	36.65	18.44	21.86	20.01	20.99	23.31
Task vector	MS COCO	MTV (NeurIPS24)	22.56	22.75	24.40	25.00	26.63	9.84	13.72	12.06	12.88	11.91
		FV (ICLR24)	26.00	31.29	32.18	32.94	36.58	17.42	21.08	21.76	20.94	24.71
	PICL	MTV (NeurIPS24)	20.87	21.12	23.18	23.74	25.41	11.56	15.19	14.25	14.61	13.53
		FV (ICLR24)	24.31	29.48	30.21	31.02	34.26	19.53	22.81	23.49	23.03	26.33
ICL training	MS COCO	MT-ICL (ECCVW24)	41.39	45.18	44.56	45.64	43.19	16.12	19.73	21.04	21.58	23.41
		DARA (COLM25)	26.92	32.50	34.52	35.29	37.66	16.35	19.42	21.18	22.74	24.63
		MimIC (CVPR25)	26.73	29.64	33.95	34.86	37.88	17.23	20.16	22.38	22.06	24.72
		STV (AAAI26)	42.52	42.21	43.34	43.51	42.53	14.81	21.56	22.17	22.84	23.09
	PICL	MT-ICL (ECCVW24)	28.04	32.17	31.28	32.36	39.02	18.44	21.86	23.38	23.69	25.66
		DARA (COLM25)	24.48	29.12	30.67	32.06	34.19	18.44	21.37	23.37	24.67	26.06
		MimIC (CVPR25)	23.11	26.32	30.24	31.15	34.02	19.31	21.97	24.66	24.21	26.27
		STV (AAAI26)	28.64	28.98	29.87	30.05	38.81	24.27	23.43	24.24	24.36	25.01
Mitigating hallucination	-	PAI (ECCV24)	22.87	28.92	29.08	31.74	34.21	15.83	18.92	22.01	23.14	24.76
		VOLCANO (NAACL24)	28.21	34.44	36.02	37.88	39.37	21.02	24.47	27.89	28.21	30.04
		LURE (ICLR24)	30.41	35.98	38.61	39.44	41.72	16.72	20.11	23.58	24.79	26.04
		LRV (ICLR24)	27.05	32.62	35.11	36.47	38.02	17.21	20.34	24.12	25.08	26.02
		IBD (CVPRW25)	26.88	33.16	34.73	37.02	39.05	17.87	23.11	26.42	27.05	28.66
		SPARC (ICML25)	29.67	35.21	37.54	38.73	40.81	16.95	19.73	23.56	24.67	25.41
		ProCoRe (Ours)	**43.22**	**47.91**	**46.04**	**46.21**	**45.23**	**25.36**	**27.81**	**29.22**	**31.54**	**33.67**

Table 11: **Quantitative comparison in multimodal ICL captioning benchmarks.** For each dataset and shot, the top scores are highlighted in bold, and the second-best scores are underlined. We use LLaVA-OneVision-1.5-8B as the MLLM and apply random ordering of demonstrations for all baselines except retrieval- and ordering-based methods.

In contrast, despite using no ICL data, ProCoRe matches or even exceeds these baselines on their training benchmarks and significantly outperforms them on unseen ones. This supports our hypothesis that improving visual grounding yields broadly generalizable multimodal ICL gains, rather than benchmark-specific improvements.

A.7 Quantitative Analysis across Diverse MLLMs

Beyond Qwen-VL-3, we also evaluate ProCoRe with LLaVA-OneVision-1.5-8B [3] and InternVL-3.5-8B [59] to demonstrate robustness across MLLMs. As shown in Tab. 11 and Tab. 12, ProCoRe consistently outperforms baselines, demonstrating strong cross-model applicability.

Category	ICL Training Dataset	Method	MS COCO					PICL				
			0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	1-shot	2-shot	3-shot	4-shot
Off-the-shelf	-	Random	27.57	31.78	33.47	34.11	33.99	15.68	14.58	17.37	16.82	19.95
Retrieval and ordering	-	RICES (NeurIPS22)	27.57	31.78	33.29	33.05	34.06	15.68	14.58	18.27	22.19	25.02
		VICL (ACL24)	26.01	33.02	34.56	34.73	35.21	16.41	15.92	15.36	21.88	21.47
		MMICES (WACV25)	27.57	31.78	33.48	33.83	33.74	15.68	14.58	14.85	20.54	20.56
Task vector	MS COCO	MTV (NeurIPS24)	28.41	32.67	34.21	34.11	35.72	13.02	13.74	15.26	13.18	16.72
		FV (ICLR24)	27.98	33.02	34.05	33.64	35.18	14.11	12.92	15.21	14.03	17.48
	PICL	MTV (NeurIPS24)	26.03	30.11	31.84	32.76	34.02	15.27	15.48	17.02	14.63	18.41
		FV (ICLR24)	25.74	31.18	32.41	31.92	33.47	16.08	14.36	17.68	15.85	19.05
ICL training	MS COCO	MT-ICL (ECCVW24)	36.63	37.91	38.11	38.30	37.74	22.18	22.01	21.04	21.37	22.09
		DARA (COLM25)	27.57	33.06	34.48	37.05	36.78	13.84	15.02	17.41	20.12	22.03
		MimIC (CVPR25)	36.19	36.14	38.05	37.18	37.99	24.11	24.02	26.07	21.91	18.02
		STV (AAAI26)	34.38	37.88	37.11	37.07	37.67	14.73	20.11	18.42	21.18	21.44
	PICL	MT-ICL (ECCVW24)	32.64	33.52	33.19	32.87	32.02	24.52	24.38	23.28	23.60	24.56
		DARA (COLM25)	23.83	28.37	29.88	31.12	31.47	15.68	17.26	19.90	22.65	24.17
		MimIC (CVPR25)	31.41	32.08	33.42	32.66	32.37	**26.89**	26.37	28.29	26.83	29.95
		STV (AAAI26)	29.61	32.54	32.01	31.88	32.12	**26.89**	30.87	30.59	33.46	33.90
Mitigating hallucination	-	PAI (ECCV24)	25.18	29.62	31.08	31.43	32.07	15.48	12.67	15.26	15.84	20.72
		VOLCANO (NAACL24)	32.98	33.92	34.61	35.43	34.57	22.55	30.89	23.92	23.76	24.96
		LURE (ICLR24)	29.06	33.41	34.55	36.02	35.11	17.63	25.94	19.81	19.42	22.11
		LRV (ICLR24)	33.05	34.47	34.92	35.11	34.02	20.74	28.96	22.41	22.18	23.07
		IBD (CVPRW25)	35.44	36.83	35.27	38.02	36.88	21.83	30.12	21.95	24.11	22.64
		SPARC (ICML25)	34.12	36.02	37.08	36.14	36.21	19.92	27.88	20.77	21.36	23.58
		ProCoRe (Ours)	**38.02**	**40.71**	**39.36**	**41.22**	**39.18**	24.81	**32.46**	**33.92**	**34.37**	**35.74**

Table 12: **Quantitative comparison in multimodal ICL captioning benchmarks.** For each dataset and shot, the top scores are highlighted in bold, and the second-best scores are underlined. We use InternVL-3.5-8B as the MLLM and apply random ordering of demonstrations for all baselines except retrieval- and ordering-based methods.

A.8 Details of Baselines

Multimodal ICL baselines can be grouped into (i) retrieval- and ordering-based methods (MMICES [13], VICL [72], RICES [2]), (ii) ICL training-based methods (STV [39], MimIC [22], MT-ICL [15], DARA [12]), and (iii) task vector-based methods (MTV [21], FV [54]).

A.8.1 Retrieval- and Ordering-based Baselines

MMICES [13] shows that selecting demonstrations based solely on text or visual similarity to the query is suboptimal for in-context learning. While text similarity is generally more informative than image similarity for selecting ICL demonstrations [62], high text similarity often corresponds to generic questions (*e.g.*, "What does the picture depict?"), which can lead to irrelevant context selection. Therefore, MMICES considers both modalities: to select K ICL demonstrations, it first retrieves N $(N > K)$ candidates with high visual similarity to the query, and then selects the final K based on text similarity.

VICL [72] also selects and orders ICL demonstrations considering both visual and text similarity. Specifically, to select K demonstrations, it first retrieves N $(N > K)$ candidates with high visual similarity to the query image. It then performs cross-modal reranking to ensure that the selected demonstrations are not only visually similar but also semantically aligned with the query. Concretely, it generates a caption T_q describing the query image, and reranks (*i.e.*, ascending order) the N candidates based on cross-modal similarity using CLIP, selecting the top K.

RICES [6] selects and orders ICL demonstrations based on multimodal similarity to the query. Specifically, it computes cosine similarity between the query and candidate samples using embeddings from a multimodal encoder that jointly considers image and text, and selects the most similar samples as demonstrations, ordering them in ascending similarity. Following Qin et al. [43], we use BridgeTower [60] as the multimodal encoder.

A.8.2 ICL training-based Baselines

MT-ICL [15] proposes a multi-shot ICL instruction tuning framework for MLLMs. Instead of using a single-instruction format (*i.e.*, image + instruction + response), it employs multiple instructions in an ICL format during training, enabling the model to effectively leverage in-context examples. In other words, it constructs instruction tuning data in an ICL format and trains MLLMs to utilize demonstrations, thereby explicitly enabling multimodal ICL.

DARA [12] shows that MLLMs mainly rely on textual inputs and neglect visual information in the ICL context during multimodal ICL. Based on this observation, DARA aims to increase the norm of visual tokens in ICL demonstrations. To this end, it trains scaling parameters that are multiplied to the attention norms of visual tokens in the ICL context, thereby enhancing their contribution. However, since these parameters are trained on ICL datasets, DARA performs well on test sets drawn from the same datasets (*i.e.*, ID datasets) but generalizes poorly to unseen ICL datasets (*i.e.*, OOD datasets), as shown in Fig. 2. Moreover, DARA learns shot-specific scaling parameters for n-shot ICL, requiring training with n-shot data and preventing its use in other shot settings.

MimIC [22] argues that prior methods apply shifting vectors to FFN hidden states after attention, whereas such shifts should be applied at the level of individual attention heads. Motivated by this, it introduces a modified attention module with learnable parameters that capture the shifting effect. Specifically, it replaces standard self-attention heads and trains these parameters to reproduce the outputs obtained when processing ICL demonstrations with the original attention. At inference, it applies the learned shift to the query without requiring demonstrations, enabling ICL-like behavior similar to task vector-based methods.

STV [39] compute activation deltas between queries with and without demonstrations, and select the top-k heads with the largest difference. It then collect activations from these selected heads under ICL and cluster them to form a candidate steering vector set. A learnable logit selects which vectors to apply, optimized to minimize the loss between outputs with ICL demonstrations and those with activation steering using selected vectors. It requires ICL data to construct candidate steering vectors and train the coefficients.

A.8.3 Task vector-based Baselines

MTV [21] compresses ICL data into task vectors stored in attention-head activations. It runs multiple few-shot inferences and averages the last-token attention-head activations to obtain mean task vectors. At inference, these vectors steer activations at task-relevant attention-heads that are most useful for the downstream task, allowing the model to exploit many-shot multimodal demonstrations without explicitly placing all examples in the context window. Similar to STV, it requires ICL data for task vector extraction and task-specific head selections.

FV [54] estimates *function vectors*, which trigger specific functions (*e.g.*, translation or synonym generation) when applied to the transformer hidden states. Specifically, it identifies attention heads whose output probabilities are most sensitive to label shuffling in ICL demonstrations, and defines a task vector by summing their activations over an ICL dataset. During inference, this task vector is added to the hidden states for a query, enabling in-context learning behavior even without explicit demonstrations and alleviating context length constraints.

A.9 Details of Prompt Decomposition in ProCoRe

In ProCoRe, we decompose captions generated by the MLLM into proposition-level components using GPT-4o-mini. During decomposition, we use the following prompt:

```
You are given two sentences: Sentence 1 and Sentence 2. Decompose them into
three full-sentence propositions.


Rules:

1. Output exactly three fields:
- shared
- variant 1
- variant 2

2. shared:
- must include only the meaning that is common to both Sentence 1 and Sentence
2
- must not include content that differs between the two sentences

3. variant 1:
- must include only the meaning that is unique to Sentence 1
- must be a full sentence
- must not repeat the shared content unless needed for grammaticality

4. variant 2:
- must include only the meaning that is unique to Sentence 2
- must be a full sentence
- must not repeat the shared content unless needed for grammaticality

5. Preserve the original meaning of both sentences.

6. Do not add extra facts.
```

A.10 Details of Multimodal Contrastive Training Pairs

We first illustrate examples of multimodal contrastive training pairs in Fig. 9 We also argue that our method generalizes to any multimodal contrastive dataset with such image–caption pairs. In Tab. 15, we report the performance of ProCoRe across diverse multimodal contrastive benchmarks (InpaintCOCO [45], VisMin [4], Winoground [53], and EQBEN [58]) when trained on each of them. As shown, ProCoRe achieves consistent performance regardless of the contrastive training benchmark, indicating that gains stem from the proposed contrastive reward rather than any specific dataset. To better capture diverse visual contrastive characteristics, we construct a mixed training set by randomly sampling 500 pairs each from EQBEN and InpaintCOCO, which exhibit the most distinct and complementary properties. Specifically, EQBEN primarily varies object behavior and count, whereas InpaintCOCO focuses more on differences in object identity, attributes, and relations. We expect that combining ProCoRe with more diverse, high-quality multimodal contrastive datasets will further enhance visual grounding.

A.11 Quantitative Analysis across Model Sizes

We employ Qwen-VL-3-8B model main experiments, as well as InternVL-3.5-8B and LLaVA-OneVision-1.5-8B in Sec. A.7. To validate that ProCoRe improves multimodal ICL performance regardless of the base MLLM size, we further conduct experiments with 4B and 14B models. Specifically, we include Qwen-VL-3-4B and InternVL-3.5-14B. Since Qwen-VL-3 is only available up to the 8B scale, we use the 14B variant of InternVL-3.5 to evaluate the effectiveness of ProCoRe on larger models. As shown in Tab. 13 and Tab. 14, ProCoRe consistently outperforms baselines, which aligns with the results observed on 8B models, demonstrating that its effectiveness is not limited to a specific model scale.



Figure 9: **Examples of multimodal contrastive datasets.** Each sample consists of two image-text pairs, (I_1, t_1) and (I_2, t_2), with captions t_1 and t_2 describe images I_1 and I_2, respectively. I_1 and I_2 are visually contrastive (*e.g.*, differing in objects (left) or object attributes (right)).

A.12 Extended Related Work

Multimodal In-Context Learning. Building on the emergence of ICL in LLMs, prior work has attempted to extend ICL to MLLMs, which can be broadly categorized into three groups: (i) retrieval- and ordering-based methods, (ii) ICL training-based methods, and (iii) task vector-based methods.

Category	ICL Training Dataset	Method	MS COCO					PICL				
			0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	1-shot	2-shot	3-shot	4-shot
Off-the-shelf	-	-	11.02	14.87	16.55	21.11	22.34	22.31	26.94	22.17	30.88	31.42
Retrieval and ordering	-	RICES (NeurIPS22)	13.21	17.44	20.02	21.08	24.91	26.77	24.01	27.83	25.16	29.94
		VICL (ACL24)	12.88	19.13	20.76	22.05	24.77	20.42	29.63	28.91	28.04	32.11
		MMICES (WACV25)	10.74	15.62	18.03	20.14	23.09	21.55	27.48	18.92	25.77	27.61
Task vector	MS COCO	MTV (NeurIPS24)	6.11	10.84	14.73	17.02	20.88	24.41	22.88	18.72	28.11	30.92
		FV (ICLR24)	9.87	13.55	17.12	20.44	23.02	23.88	22.11	26.73	29.04	30.55
	PICL	MTV (NeurIPS24)	2.84	7.11	11.02	13.74	16.42	22.18	24.91	20.11	26.03	28.72
		FV (ICLR24)	5.98	10.44	13.37	16.09	19.11	25.92	27.83	23.74	29.62	32.41
ICL training	MS COCO	MT-ICL (ECCVW24)	31.85	35.92	36.14	37.21	**39.66**	**31.66**	27.04	33.42	28.71	34.87
		DARA (COLM25)	18.54	23.61	25.74	28.33	31.88	20.11	21.83	26.34	24.01	29.41
		MimIC (CVPR25)	22.07	28.19	29.44	30.01	33.02	23.64	28.94	26.77	29.91	30.72
		STV (AAAI26)	31.44	36.28	35.11	36.52	37.02	31.88	33.27	29.44	35.02	34.11
	PICL	MT-ICL (ECCVW24)	28.03	33.12	33.77	34.88	35.14	29.84	35.27	30.91	36.88	33.02
		DARA (COLM25)	14.77	20.05	21.33	24.61	28.22	22.44	28.13	24.28	30.55	31.92
		MimIC (CVPR25)	19.31	25.77	26.02	27.88	29.65	25.31	31.44	24.88	29.77	33.65
		STV (AAAI26)	27.66	32.44	33.18	34.05	35.27	30.22	35.91	31.08	33.44	37.11
Mitigating hallucination	-	PAI (ECCV24)	8.12	13.44	14.02	18.76	20.01	18.88	27.02	24.76	22.48	29.91
		VOLCANO (NAACL24)	24.88	30.22	28.44	32.03	36.55	25.55	29.84	22.13	29.02	31.66
		LURE (ICLR24)	10.23	12.11	15.04	18.33	20.52	23.77	24.63	27.91	31.18	27.94
		LRV (ICLR24)	12.77	22.94	21.33	19.01	24.12	28.91	28.55	31.72	32.87	35.44
		IBD (CVPRW25)	25.44	28.01	30.88	31.55	29.03	17.91	24.84	24.72	21.03	23.98
		SPARC (ICML25)	27.11	30.42	29.18	34.02	36.77	26.42	33.14	28.11	30.52	35.62
		ProCoRe (Ours)	**32.88**	**39.12**	**38.77**	**39.55**	38.21	27.55	**38.72**	**37.41**	**39.84**	**35.93**

Table 13: **Quantitative comparison in multimodal ICL captioning benchmarks.** For each dataset and shot, the top scores are highlighted in bold, and the second-best scores are underlined. We use Qwen-VL-3-4B as the MLLM and apply random ordering of demonstrations for all baselines except retrieval- and ordering-based methods.

Category	ICL Training Dataset	Method	MS COCO					PICL				
			0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	1-shot	2-shot	3-shot	4-shot
Off-the-shelf	-	Random	26.41	35.02	35.11	32.84	35.91	24.82	17.77	28.66	22.48	29.14
Retrieval and ordering	-	RICES (NeurIPS22)	24.91	32.66	36.58	34.11	31.94	26.04	28.85	25.71	34.72	30.63
		VICL (ACL24)	23.84	36.02	36.44	35.97	34.82	22.51	19.47	27.18	31.11	33.62
		MMICES (WACV25)	28.73	30.42	34.95	36.48	37.28	19.18	15.77	21.42	29.83	34.66
Task vector	MS COCO	MTV (NeurIPS24)	22.74	34.18	29.41	33.02	35.88	21.92	18.74	30.11	23.84	31.27
		FV (ICLR24)	30.01	36.47	37.88	31.77	36.41	27.72	24.83	28.96	26.15	32.77
	PICL	MTV (NeurIPS24)	20.11	30.74	27.02	29.88	31.41	23.18	16.21	28.74	25.92	29.03
		FV (ICLR24)	24.63	34.19	30.55	33.26	35.94	27.64	20.55	32.41	29.84	33.76
ICL training	MS COCO	MT-ICL (ECCVW24)	37.91	41.22	34.88	41.03	35.66	28.11	28.84	33.96	28.74	33.88
		DARA (COLM25)	31.94	32.61	35.27	40.72	37.83	17.11	25.76	19.22	32.47	25.88
		MimIC (CVPR25)	39.55	36.83	41.28	39.47	41.51	36.88	30.91	33.77	32.13	29.06
		STV (AAAI26)	31.74	35.08	37.44	40.61	36.02	26.73	28.12	20.91	33.76	28.55
	PICL	MT-ICL (ECCVW24)	30.02	34.62	36.11	30.14	35.48	29.22	25.18	30.74	27.84	37.61
		DARA (COLM25)	24.44	30.98	26.84	32.58	29.41	25.61	21.72	32.94	35.55	26.94
		MimIC (CVPR25)	28.11	34.55	36.74	30.62	34.83	**38.66**	35.03	30.71	40.66	37.94
		STV (AAAI26)	32.44	35.22	31.55	34.91	35.76	30.55	39.83	38.61	44.42	38.77
Mitigating hallucination	-	PAI (ECCV24)	23.74	26.55	29.02	33.84	31.12	18.14	25.94	27.11	29.88	24.02
		VOLCANO (NAACL24)	35.12	34.84	36.52	31.98	37.81	30.55	41.88	32.31	38.62	33.76
		LURE (ICLR24)	28.81	31.27	36.84	38.12	32.66	30.83	28.91	29.74	24.02	25.18
		LRV (ICLR24)	36.44	32.11	38.52	36.73	31.62	28.77	37.03	33.91	36.72	26.18
		IBD (CVPRW25)	39.27	37.66	31.88	39.94	40.21	29.11	42.74	32.63	36.83	28.41
		SPARC (ICML25)	35.94	39.88	40.44	35.12	39.62	32.77	39.66	24.61	33.88	34.92
		ProCoRe (Ours)	**40.88**	**43.61**	**42.84**	**45.33**	**41.77**	36.62	**43.91**	**40.88**	**45.27**	**40.42**

Table 14: **Quantitative comparison in multimodal ICL captioning benchmarks.** For each dataset and shot, the top scores are highlighted in bold, and the second-best scores are underlined. We use InternVL-3.5-14B as the MLLM and apply random ordering of demonstrations for all baselines except retrieval- and ordering-based methods.

Retrieval- and ordering-based methods [6, 38, 13] retrieve and rank ICL demonstrations by their similarity to the query, improving performance with appropriate sample selection and ordering. However, they are highly sensitive to prompt design and exhibit substantial performance fluctuations [18], especially in multimodal settings, where aligning textual and visual features introduces additional complexity [37, 43, 26]. As a result, they often perform worse than random selection [63].

ICL training-based methods [15, 52, 12, 26, 22] enable multimodal ICL by further training MLLMs on ICL datasets, either over full model parameters [15, 52] or additional learnable modules [12]. While these approaches achieve strong performance on the same ICL datasets used for training, they show limited generalization, with significant performance drops on ICL benchmarks not seen

during training. Moreover, additional training with ICL data can lead to forgetting of previously well-performing tasks (*e.g.*, simple VQA tasks), as shown in Fig. 5.

Task vector-based methods [42, 32, 21, 39] compress ICL demonstrations into compact task vectors that steer model activations during inference, achieving effects comparable to conditioning on explicit demonstrations. While they reduce memory overhead and context length constraints and perform well on simple tasks, they struggle to generalize to complex multimodal scenarios [42, 26] and may lose task-specific details due to limited representational richness [21, 39]. Furthermore, similar to ICL training-based methods, they show limited generalization to unseen ICL benchmarks (Fig. 2). Different from prior work that improves multimodal ICL through better demonstration selection or ICL data-dependent adaptation, our method targets a more fundamental bottleneck, *i.e.*, poor visual grounding, improving multimodal ICL without relying on any ICL datasets.

Mitigating Hallucinations in MLLMs. Instruction tuning data enhancement methods [30, 55, 66] mitigate bias and hallucinations in visual instruction tuning data, improving factual grounding. Zhu et al. [73] show that these data are biased toward positive responses, while Liu et al. [30], Yu et al. [66] report that over 30% of LLaVA-Instruct-150K contains hallucinated instructions. To address this, prior work performs additional instruction tuning on remedial datasets [35, 73] (*e.g.*, with balanced positive and negative responses) or preprocesses instruction tuning data to detect and correct hallucination [66, 55]. However, these approaches require access to instruction tuning data and incur substantial computational cost due to their large scale.

Supplementary network-based approaches [71, 57, 20] use auxilary networks to identify hallucination of the output or to revise the output text from LVLMs. LURE [71] trains a reviser network on hallucinated–clean caption pairs to refine MLLM outputs, while FactChecker [20] generates report-specific code to verify facts in each medical report and update them accordingly. However, these approaches typically require a substantial number of additional parameters or training data [74].

Visual grounding enhancement approaches often directly amplify attention to visual tokens during self-attention to counter the strong bias of MLLMs toward textual context over visual inputs [33, 16, 12]. PAI [33] amplifies all visual token attention by multiplying with a model-specific scaling factor, but this indiscriminate scaling can increase attention to irrelevant tokens and exacerbate hallucinations [23]. SPARC [23] instead selectively amplifies only the attention of relevant visual tokens through important token selection. However, directly scaling visual token norms can distort feature representations, as it alters the magnitude of embeddings [67].

A.13 Additional Examples of PICL

In addition to Fig. 10, we further provide examples of PICL by comparing it with MS COCO [29], existing multimodal ICL captioning benchmark. As shown, MS COCO ICL mainly involves generic captioning tasks, where query captions can often be predicted using pre-trained semantic priors without truly understanding the demonstrations, although the demonstrations may still guide the output format.

In contrast, PICL requires visual understanding across all demonstrations to answer the query. For example, in the first row of Fig. 10, the first demonstration alone is insufficient to identify which character is 'Jack' or 'Mia'. However, the second demonstration reintroduces only the female character and mentions 'Mia', allowing the model to infer that the male character in the first demonstration is 'Jack'—knowledge necessary to correctly answer the query as 'Jack'. Similarly, in the second and third rows, accurate visual understanding of each demonstration is required to produce the correct answer. Consequently, PICL evaluate true visual grounding in multimodal ICL and preventing shortcut solutions based solely on pre-trained semantic priors.

A.14 Details of PICL

A.14.1 Details of Benchmark Construction

PICL is constructed from visual storytelling datasets where the same characters appear across multiple frames, each with a corresponding caption, including PororoSV [27], FlintstonesSV [19], and StoryStream [61]. The construction consists of two steps: (i) data preprocessing and (ii) ICL data construction.

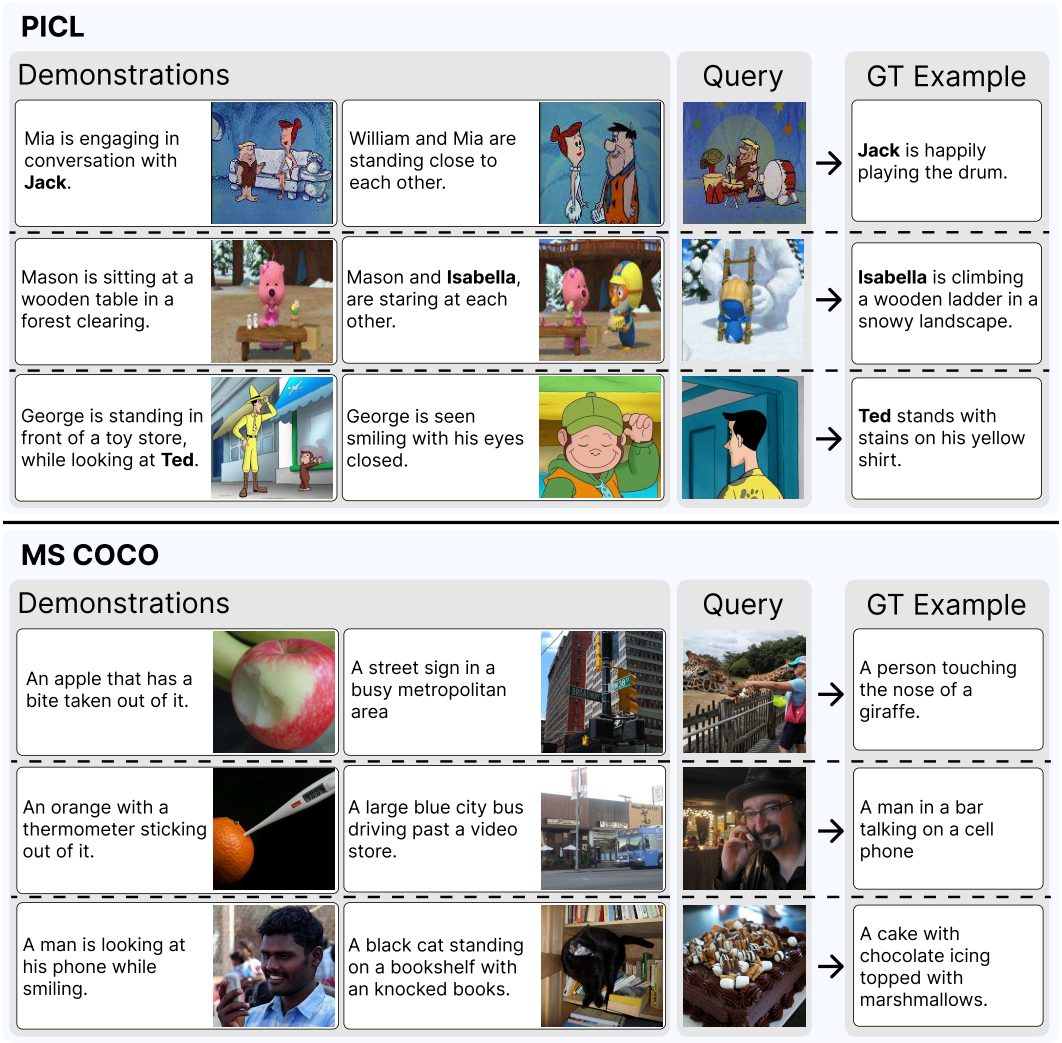


Figure 10: **Examples comparing PICL and MS COCO.** While PICL requires visual understanding all demonstrations to answer correctly (*e.g.*, resolving "Mason" in the first demo is necessary to correctly identify "Isabella" for Pororo in the second, enabling the query prediction), MS COCO ICL task can often be solved using pre-trained priors without understanding demonstrations.

Data Preprocessing. Each dataset contains a finite set of characters $\mathcal{C}$ (*e.g.*, $\mathcal{C} = \{\text{Pororo, Loopy, Crong, Poby, Eddy}\}$ in PororoSV). We define a reference image for each character in $\mathcal{C}$. Then, using SigLIP [68] and the reference images, we determine which characters are present in each frame and group frames according to the set of characters present in each frame.

We then rewrite the corresponding caption for each frame using GPT-4o-mini. Although visual storytelling datasets provide captions for each frame, these are often misaligned with the frame or inaccurate, as they are derived from video-level descriptions and may not precisely reflect frame-level content. However, frame-only captioning cannot reliably identify character names, even when the set of present characters is known, as it cannot determine which name corresponds to which entity when multiple characters appear. To prevent this, we provide the reference images $\mathcal{C}$ as context and rewrite the captions accordingly, using the following prompt:

Contrastive Benchmark	TrueMICL					MS COCO				
	0-shot	1-shot	2-shot	4-shot	8-shot	0-shot	1-shot	2-shot	4-shot	8-shot
Winoground	28.74	48.12	52.03	**64.91**	66.77	36.84	41.72	43.03	42.11	43.88
VisMin	28.74	49.72	54.18	60.05	65.11	**40.07**	44.82	44.44	45.11	43.96
InpaintCOCO	**31.11**	49.63	**54.62**	63.80	**68.24**	38.21	45.02	**45.18**	43.77	**46.91**
EqBen	27.41	**52.88**	53.03	59.44	63.02	38.72	**47.03**	41.92	**48.21**	41.33

Table 15: **Quantitaive comparison of ProCoRe across different contrastive benchmarks.** For each dataset and shot, the top scores are highlighted in bold, and the second-best scores are underlined. For a fair comparison, we randomly sample 1,000 contrastive pairs from each benchmark. We use Qwen-VL-3-8B as the base MLLM for all contrastive benchmarks.

```
You are given an image and a set of reference images.  Write a caption
describing the image.


Rules:
- Generate a caption for the image using the reference images to identify
characters
- Add visual information based only on the image
- Focus on visible objects, interactions, positions, and background elements
- Prefer concrete scene details over adjectives or adverbs
- Do not invent unsupported details
- Do not make assumptions
```

Next, we replace original (*i.e.*, common) identifiers with personalized ones. Specifically, we generate English names using GPT-4o-mini [1] and substitute each identifier with a randomly selected name. To generate English names, we use following prompt:

```
Generate 100 general English first names.


Rules:
- Include a balanced mix of male and female names.
- Use common, natural-sounding English names.
- Avoid rare, fantasy, historical, or celebrity-specific names.
- Each name should be a single first name only.
- No duplicate names.
```

ICL Data Construction. After preprocessing, we manually select a subset of frames (*i.e.*, image–personalized caption pairs) from each sample in the visual storytelling datasets to construct ICL demonstrations and queries. Specifically, frames are selected based on the following criteria: (i) characters in the query image appear in at least one demonstration; (ii) demonstrations include at least two distinct characters; (iii) demonstrations provide sufficient context to distinguish character identities.

The first criterion is a necessary condition for solving personalized ICL, as demonstrations provide the only cues for identifying the correct identifier in the query image. The second criterion ensures evaluation of true visual grounding. If all demonstrations contain only a single personalized entity, the model may predict query caption by reusing the subject from the text demonstrations without understanding visual evidence. In such cases, it is unclear whether the correct identifier is derived from visual grounding or merely copied from the text. Therefore, this criterion prevents shortcut solutions and enables a more faithful evaluation of visual grounding. The final criterion prevents confusing ICL demonstrations. If multiple characters exhibit similar behaviors across all demonstrations, it becomes difficult to determine which entity corresponds to the character described in the text, potentially making the task unsolvable even with strong visual grounding. By eliminating such cases, this criterion ensures that performance reflects visual grounding rather than ambiguity.



Figure 11: **Comparison across ICL-shot settings.** Higher-shot settings require reasoning over increasingly complex relationships among demonstrations, rather than simply providing more context. For example, in the 3-shot case, the first demo lets the model infer that the dinosaur is "Henry", while the second remains ambiguous since "Ava" and "Mason" perform the same action; only after the third demo (where the fox is identified as "Henry") can the model resolve the ambiguity and infer that the pink character in the second demo is "Ava". In contrast, lower-shot settings can be easier due to fewer required inferences (e.g., in 1-shot, the pink character can be directly identified as "Amella").

We note that data for each shot setting is constructed independently following the above rules, rather than smaller-shot demonstrations being subsets of larger-shot ones. As a result, as discussed in Sec. 4 and illustrated in Fig. 11, higher-shot settings require reasoning over increasingly complex relationships among demonstrations, rather than simply providing more context, while lower-shot settings may be easier due to requiring fewer inferences.

A.15 Robustness across Random Seeds

We summarize the mean and standard deviation over three random seeds in Tab. 16. As shown, our proposed ProCoRe significantly outperforms the baselines, with performance gains that are substantially larger than the observed standard deviations. ICL ordering-based baselines (*e.g.*, RICES and MMICES) are deterministic, as the demonstrations and their order are uniquely determined by their ranking criterion. Consequently, their results are identical across random seeds and exhibit no variance.

Category	Method	PICL			TrueMICL		
		2-shot	3-shot	4-shot	2-shot	4-shot	8-shot
Off-the-shelf	-	24.97±1.55	29.18±1.50	29.37±1.52	49.37±1.29	55.58±2.60	62.47±2.40
Retrieval & Ordering	RICES	26.44±0.00	27.14±0.00	28.20±0.00	47.16±0.00	51.83±0.00	61.86±0.00
	VICL	28.11±1.39	26.10±1.75	28.48±1.89	50.06±1.59	55.11±1.74	61.66±0.53
	MMICES	21.70±0.00	23.82±0.00	24.79±0.00	48.40±0.00	58.07±0.00	63.04±0.00
Task Vector	MTV	13.85±2.10	15.09±1.84	13.24±1.25	42.94±1.78	45.39±0.32	55.55±0.81
	FV	25.86±1.92	28.02±2.10	29.62±2.26	48.97±1.22	56.24±1.52	60.74±1.10
ICL-Data Training	MiMIC	27.53±2.18	28.86±2.57	31.60±2.58	54.64±1.57	54.80±2.27	61.22±1.25
	STV	32.46±2.13	33.62±1.88	33.93±2.63	50.20±1.55	57.83±1.82	64.03±1.54
	MT-ICL	32.78±1.82	33.03±2.27	33.75±2.25	50.42±1.86	58.01±1.78	62.84±1.80
	DARA	25.35±2.27	28.35±2.03	29.78±1.86	50.94±1.78	56.24±1.02	63.58±2.13
Mitigating Hallucination	VOLCANO	24.92±2.26	27.42±2.13	29.53±2.34	50.90±1.64	57.66±1.76	63.46±2.22
	PAI	23.43±2.21	25.01±2.68	27.78±2.62	48.05±1.73	56.32±1.66	59.97±1.26
	IBD	22.92±2.69	23.74±1.21	26.34±2.39	50.47±1.74	55.51±1.28	62.62±2.10
	LRV	28.95±1.82	30.28±2.36	32.45±2.78	51.99±1.90	57.24±1.18	63.74±1.91
	SPARC	26.36±2.29	27.14±1.60	33.20±2.66	47.41±1.26	57.13±1.71	61.19±1.43
	LURE	25.30±2.15	29.78±2.10	30.40±2.24	46.08±1.46	53.08±1.56	60.38±2.46
	ProCoRe (Ours)	**36.11±1.59**	**36.88±1.31**	**36.80±1.74**	**56.07±1.87**	**62.56±1.77**	**68.04±0.92**

Table 16: **Comparison on PICL and TrueMICL across 3 seeds**. Task Vector and ICL-Data Training baselines use the training and evaluation splits from the same dataset.

A.16 Algorithm of ProCoRe

Algorithm 1 provides a pseudocode for the proposed ProCoRe.

A.17 Qualitative Analysis of Visual Grounding

We qualitatively compare visual grounding with ICL baselines in Fig. 12. As shown, when a generic concept appears in the demonstrations, both ICL baselines and ProCoRe correctly predict the concept for the query image and attend to the corresponding visual regions when generating the token. However, when the generic concept is replaced with a personalized identifier unseen during training, ICL baselines fail to attend to the relevant visual regions and instead output generic words (*i.e.*, red words). This indicates that they do not effectively leverage visual demonstrations and rely on pretrained semantic priors for prediction. In contrast, ProCoRe attends to the relevant visual regions

Algorithm 1 ProCoRe

Input: Contrastive pair data $(I_1, t_1), (I_2, t_2)$, MLLM f_θ, sigmoid function σ, reward weights $\lambda_I, \lambda_T, \lambda_s$, proposition decomposer PD, image-text similarity function sim_I, entailment score function entail

1: Generate captions for contrastive images using f

$$\hat{t}_1 \leftarrow f_\theta(I_1), \quad \hat{t}_2 \leftarrow f_\theta(I_2) \tag{8}$$

2: Decompose captions into propositions $(\hat{P}_1, \hat{P}_2, \hat{P}_s) \leftarrow \mathrm{PD}(\hat{t}_1, \hat{t}_2)$

3: Compute image contrastive reward

$$R_I^{(P)} \leftarrow \sigma\Big(\mathrm{sim}_I(I_1, \hat{P}_1) - \mathrm{sim}_I(I_1, \hat{P}_2) + \mathrm{sim}_I(I_2, \hat{P}_2) - \mathrm{sim}_I(I_2, \hat{P}_1)\Big)$$

4: Compute text contrastive reward

$$R_T^{(P)} \leftarrow \sigma\Big(\mathrm{entail}(t_1, \hat{P}_1) - \mathrm{entail}(t_2, \hat{P}_1) + \mathrm{entail}(t_2, \hat{P}_2) - \mathrm{entail}(t_1, \hat{P}_2)\Big)$$

5: Compute shared-content reward

$$R_s^{(P)} \leftarrow \sigma\Big(\mathrm{sim}_I(I_1, \hat{P}_s) + \mathrm{sim}_I(I_2, \hat{P}_s) + \mathrm{entail}(t_1, \hat{P}_s) + \mathrm{entail}(t_2, \hat{P}_s)\Big)$$

6: Combine rewards $R^{(P)} \leftarrow \lambda_I R_I^{(P)} + \lambda_T R_T^{(P)} + \lambda_s R_s^{(P)}$

7: **return** $R^{(P)}$

in the query image and correctly predicts the personalized identifier (*i.e.*, green word) introduced in the demonstrations, showing effective grounding and understanding of visual cues.

A.18 Limitations and Future Work

ProCoRe improves visual grounding using standard multimodal contrastive benchmarks and shows consistent gains across datasets. However, reliance on such curated benchmarks may limit scalability to larger-scale training and to data with more diverse contrastive visual attributes. Combining with generative frameworks that can scalably produce high-quality multimodal contrastive data across diverse visual attributes (*e.g.*, GenOL [48]) could further enhance its scalability and real-world applicability. Moreover, while PICL evaluates visual grounding in multimodal ICL via captioning, real-world multimodal tasks extend beyond captioning to more complex reasoning. Extending evaluation to such tasks would better reflect real-world settings.

A.19 Impact Statements

This work focuses on multimodal in-context learning, enabling MLLMs adapt to downstream tasks without parameter updates using a few demonstrations. However, since ProCoRe improves visual grounding via post-training with a contrastive reward, there is a potential for the unintended amplification of issues such as model bias and ethical misalignment from bias in training data. We are committed to taking all necessary measures to address these risks, although tackling these concerns is not the primary focus of this work.

A.20 Reproducibility Statement

To support reproducibility, we will release the code used in our experiments, along with instructions for setting up the environment and running the main experiments, at https://github.com/snumprlab/procore. The repository will include the relevant configuration files and evaluation scripts. Any additional details needed to reproduce the reported results are provided in the paper or repository documentation.

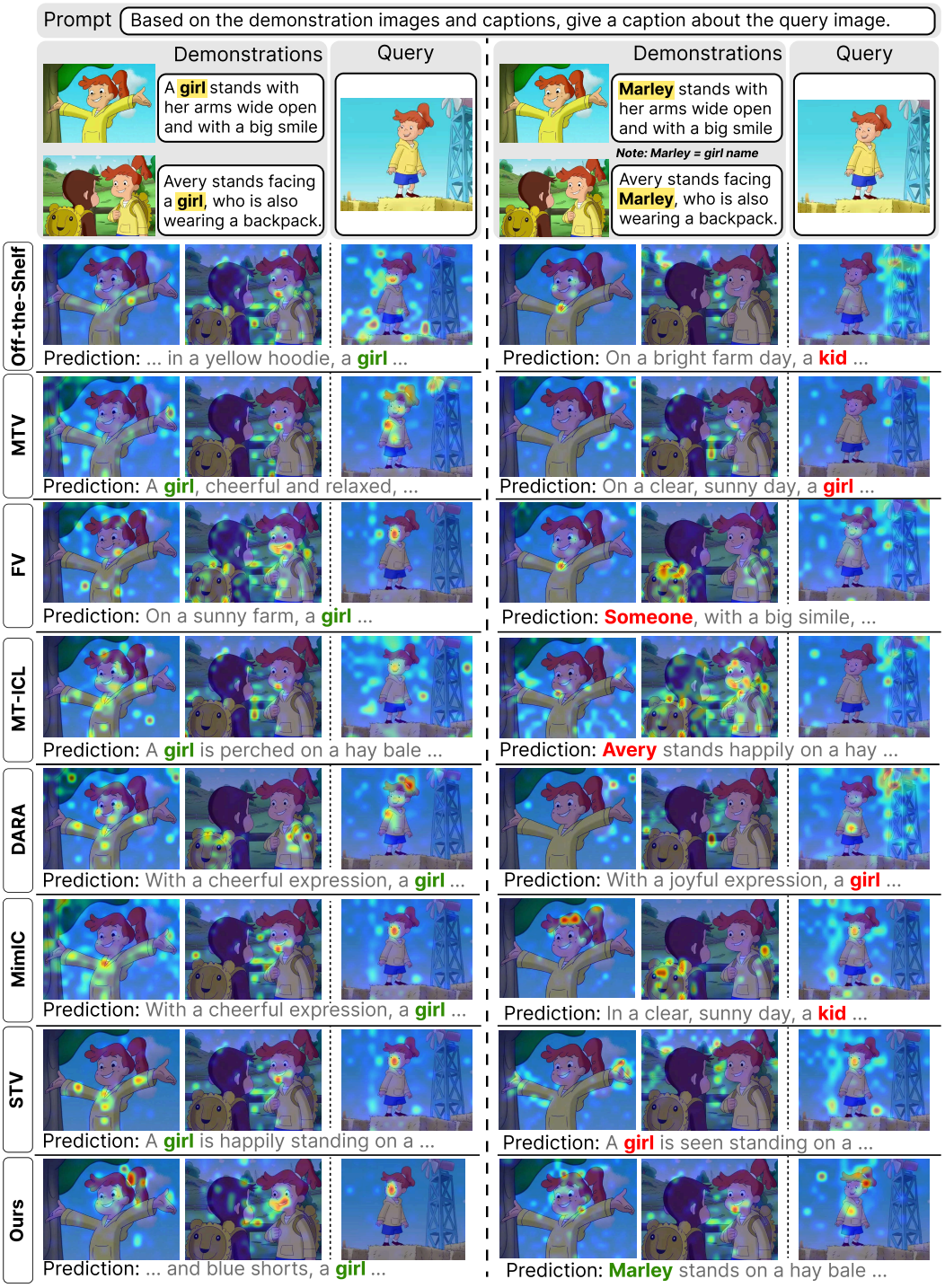


Figure 12: **Visual grounding in multimodal ICL with generic vs. personalized identifiers.** Attention maps show visual-token attention when predicting the highlighted token using Qwen-VL-3-8B. (a) When a generic identifiers (*e.g.*, "**girl**") seen during pre-training appears in the ICL demonstrations, both baselines and ProCoRe attend to visual regions relevant to the predicted token (*i.e.*, **girl**). (b) In contrast, when the generic identifiers is replaced with a personalized identifier (*e.g.*, "**Marley**"), baselines fail to attend to the relevant region and thereby fall back to generic terms (*e.g.*, "**kid**", "**girl**" "**someone**"), or incorrectly predict other names (*e.g.*, "**Avery**" in MT-ICL), whereas ProCoRe correctly predicts the personalized identifier (*i.e.*, "**Marley**") with strong visual grounding.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction consistently describe the main contributions and scope of the paper, and these contributions are clearly summarized at the end of the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations and future research direction in Sec. A.18.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This work does not include theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the experimental setup in Sec. 5.1 and include more detailed explanations and implementation details in Sec. A.2.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the link to our data and code in Sec. A.20.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the experimental setups and test details in Sec. A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide details in Sec. A.2, and prior work shows that variance in ICL experiments is typically low [2, 38].

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details of computational resources we used in Sec. A.2

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We checked the the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed impact statement of our work in Sec. A.19

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: Although we proposed a new benchmark, *i.e.*, PICL, it is composed of a combination of existing publicly deployed datasets that have already undergone safeguards, and no newly scraped datasets are included

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We followed the implementations of methods proposed by their respective authors and cited them accordingly.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We provide the details of our proposed PICL and will prepare a documentation and license before releasing the dataset after publication.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing and research with human subjects

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:his paper does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use MLLMs as a model, and the training details and the model versions are provided in the appendix.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.