NOT JUST WHAT IT IS, BUT WHAT IT ISN'T: COMPARATIVE META IN-CONTEXT LEARNING FOR FINE-GRAINED CATEGORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual in-context learning enables pretrained multimodal large language models to perform downstream categorization from labeled demonstrations without further parameter updates. Fine-grained categorization requires recognizing not only what a query is, but also what it is not among visually similar alternatives. While substantial effort has focused on demonstration retrieval and context construction, strengthening this comparative use of context through limited-data meta-training remains less explored. Standard cross-entropy (CE) supervises the correct label but does not explicitly penalize incorrect continuations once generation departs from the target sequence. Our diagnostic analysis shows that CE tuning raises both target and distractor scores, partially offsetting gains in the discriminative margin. We propose *CLOUD* (**C**ontrastive **L**ikelihood **O**nly **U**pon **D**ivergence), which complements target-label learning with explicit negative supervision on in-context distractor labels. Across three backbone configurations, *CLOUD* improves accuracy over CE-based Meta-ICL by 1.74–2.21% on six held-out natural fine-grained benchmarks and 2.14–6.11% on three specialized benchmarks. With *CLOUD*, Qwen3-VL-2B further outperforms Qwen3-VL-32B with vanilla ICL by 4.63% on average across all nine benchmarks.

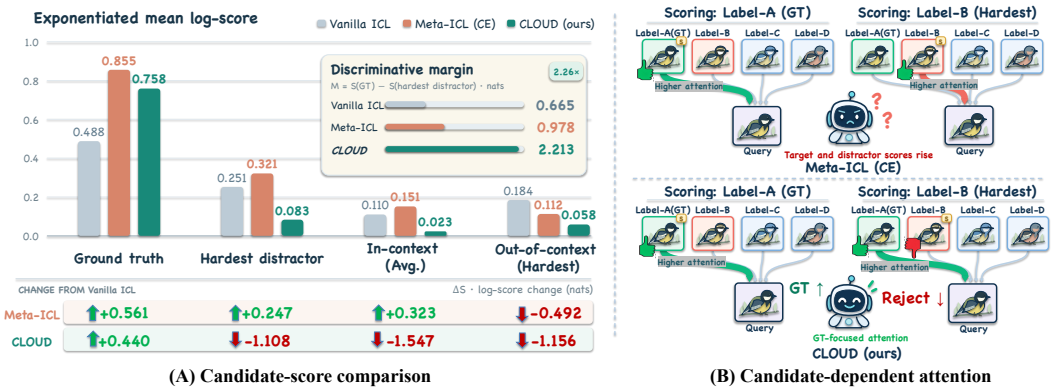
1 INTRODUCTION

Fine-grained visual categorization (FGVC) involves distinguishing closely related categories amid substantial intra-class variation Wei et al. (2021); Zhang et al. (2014); Fu et al. (2017); Dubey et al. (2018); Luo et al. (2019). In-context learning enables pretrained multimodal large language models (MLLMs) to use labeled demonstrations for downstream prediction without further parameter updates Alayrac et al. (2022); Zhao et al. (2024). In fine-grained settings, variation in exemplar selection and support–query appearance can make the relevant category distinctions less explicit in the context Chen et al. (2025a); Qin et al. (2024). This motivates improving how models extract and compare discriminative evidence from the available demonstrations Chen et al. (2025b).

Prior studies have explored demonstration selection and organization to improve context construction Chen et al. (2023; 2025a); Xiong et al. (2026); Fu et al. (2026); Qin et al. (2024). Complementing these efforts, we focus on how pretrained MLLMs utilize the available demonstrations for fine-grained discrimination. Specifically, we investigate candidate-discriminative meta-training with limited source-task data, examining its transfer to tasks excluded from meta-training under variable support–query visual alignment.

MetaICL-style meta-training improves context utilization through supervision across source tasks Min et al. (2022). Its standard cross-entropy (CE) objective directly supervises the target sequence, while rejection of distractor label sequences remains implicit. Our analysis reveals two related patterns: CE-based tuning raises both target and distractor scores, partially offsetting gains in the discriminative margin (Fig. 1(A)), and attention shifts toward the demonstration associated with the candidate being scored (Fig. 1(B)). Such candidate-dependent matching can support competing labels individually without adequately separating them, motivating explicit negative supervision for fine-grained candidate discrimination.

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



(A) Candidate-score comparison (B) Candidate-dependent attention

Figure 1: **Candidate scoring and discrimination under CE tuning and *CLOUD*. (A)** Scores are exponentiated macro-averaged log-scores; the outside-context group pools the highest-scoring outside-context label from each model (Appendix A). *CLOUD* reduces the hardest-distractor score and increases the target-to-distractor log-score margin from 0.978 to 2.213 relative to Meta-ICL. **(B)** Schematic summary of the attention analysis: when scoring $Label - A$ or $Label - B$, matching only to the corresponding demonstration can make both labels appear correct. *CLOUD* instead aims to use the demonstrations jointly to favor the correct $Label - A$ over the distractor $Label - B$.

To strengthen comparative discrimination, we propose *CLOUD* (**C**ontrastive **L**ikelihood **O**nly **U**pon **D**ivergence), which complements target-label supervision with explicit penalties on in-context distractor labels. Since fine-grained category names may share initial tokens, indiscriminate negative supervision can conflict with target learning at shared positions. *CLOUD* excludes the longest common prefix between each distractor label and the target, applying bounded negative penalties to the remaining suffix. This design targets distractor label continuations while retaining the baseline architecture and inference procedure.

Using only 5,549 query–context pairs from two source datasets, Oxford Pets Parkhi et al. (2012) and DTD Cimpoi et al. (2014), we meta-fine-tune *CLOUD* and evaluate its cross-domain generalization on nine datasets excluded from meta-training. The evaluation spans six natural fine-grained benchmarks and three specialized domains—agriculture Mohanty et al. (2016), industrial inspection Bergmann et al. (2019), and medical imaging Borgli et al. (2020)—with model parameters frozen throughout downstream evaluation. Across three backbone configurations, *CLOUD* improves mean accuracy over CE-based Meta-ICL by 1.74–2.21% on natural fine-grained benchmarks and 2.14–6.11% on specialized benchmarks. Its 2B variant further exceeds Vanilla ICL-32B by approximately 4.63% across all nine datasets. Further analyses demonstrate higher contextual fidelity in following the supplied demonstrations.

Our core contributions are summarized as follows:

- **Data-Efficient Context Utilization.** We study fine-grained visual in-context learning through limited-data meta-training, focusing on the transfer of candidate-discrimination capabilities under imperfect support–query visual alignment.
- **Candidate Discrimination Analysis.** Our score and attention analyses identify elevated distractor scores and candidate-dependent exemplar matching under CE-based tuning, motivating explicit supervision for rejecting in-context distractor labels.
- **Divergence-Gated Negative Supervision.** We introduce *CLOUD*, which complements target-label learning with negative supervision on distractor label suffixes, excluding shared-prefix tokens while preserving the baseline architecture and inference procedure.
- **Cross-Domain Generalization and Contextual Fidelity.** With only 5,549 meta-training pairs, *CLOUD* improves over matched CE baselines across nine held-out datasets and exhibits higher contextual fidelity, with its 2B variant surpassing Vanilla ICL-32B on average.

2 Related Work

2.1 Context Construction for Visual In-Context Classification

Visual in-context learning enables pretrained multimodal models to perform classification from labeled demonstrations without task-specific parameter updates Alayrac et al. (2022); Zhao et al. (2024). In fine-grained settings, these demonstrations must convey subtle category distinctions despite variation in appearance, viewpoint, and background Wei et al. (2021). The effectiveness of the context therefore depends on both the examples selected and how their category information is presented Chen et al. (2023); Qin et al. (2024). Label-space manipulation investigates how label representations affect in-context classification, while coreset optimization selects informative demonstration sets for multimodal image classification Chen et al. (2023; 2025a). CIRCLES uses attribute-guided retrieval to construct counterfactual-style demonstrations that expose discriminative visual differences Xiong et al. (2026). ContextNav extends context construction to an agentic workflow, combining retrieval, structural alignment, and feedback-driven refinement to improve robustness to noisy demonstrations Fu et al. (2026).

Together, these approaches improve the evidence available for prediction through context selection, representation, and organization. A complementary question is how effectively the model uses that evidence when the supplied examples are imperfectly aligned with the query. This distinction is particularly relevant to fine-grained classification, where recognizing shared category characteristics must be accompanied by discrimination among visually similar alternatives. Our work addresses this utilization problem through meta-training.

2.2 Learning Transferable Context Utilization

Training can strengthen a model's ability to learn from demonstrations. Flamingo develops multimodal few-shot capabilities through large-scale training, while MMICL explicitly models multimodal contextual relationships Alayrac et al. (2022); Zhao et al. (2024). MetaICL establishes cross-task meta-training as a means of improving in-context adaptation in language models Min et al. (2022). More directly targeting visual context utilization, DARA introduces a fine-tuning strategy that rebalances attention across visual and textual tokens to encourage the use of visual demonstrations Chen et al. (2025b). In the related setting of fine-grained MLLM adaptation, Fine-R1 combines chain-of-thought supervision with triplet-augmented policy optimization. Its main fine-grained evaluation examines base-to-new category generalization within the datasets represented during training He et al. (2026). Our setting instead examines whether context-dependent discrimination learned from a small source-task corpus transfers across datasets and application domains, with model parameters frozen during downstream evaluation.

Building on these advances, we study data-efficient meta-training for fine-grained in-context discrimination, emphasizing effective use of imperfect demonstrations and generalization across domains without modifying the inference procedure.

3 Methodology

This section delineates our meta-training and evaluation framework for fine-grained visual in-context learning (Fig. 2): formalizing class-centroid support stream construction and length-normalized evaluation (§3.1); diagnosing optimization bottlenecks in standard autoregressive Cross-Entropy (§3.2); formulating the *CLOUD* objective (§3.3); and analyzing its theoretical properties and computational overhead (§3.4).

3.1 Task Formalization and Evaluation Protocol

Class-Centroid Random Support Stream. In standard fine-grained benchmarks, training and test images often share similar acquisition conditions and visual characteristics. Direct image-level nearest-neighbor retrieval can therefore favor demonstrations that closely resemble the query, making it less informative about context utilization when such visual alignment is unavailable. To examine this setting using existing benchmarks, we retrieve candidate classes through visual class centroids and randomly sample exemplars within each retrieved class (Fig. 2(A)). This retains category-relevant context while avoiding explicit selection for exemplar-level similarity, enabling us to study

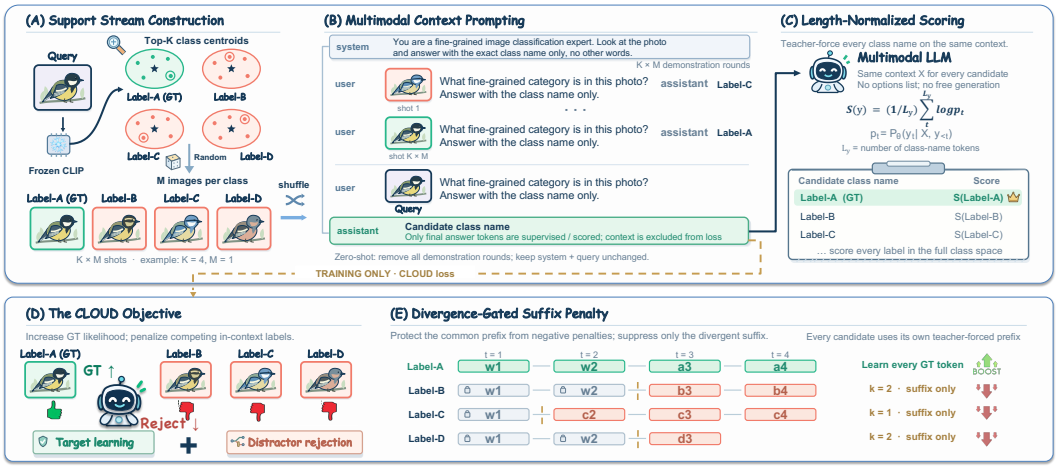


Figure 2: **Overview of the *CLOUD* framework. (A) Support Stream Construction:** CLIP centroids retrieve confounding classes $\mathcal{Y}_{\text{ctx}}$, randomly sampled into a context $\mathbf{X}$ of $K \times M$ demonstrations. **(B) Multimodal Context Prompting:** Multi-turn dialogue supervising only the final query answer tokens. **(C) Length-Normalized Scoring:** Candidates are evaluated via length-normalized log-likelihood $\mathcal{S}_\theta(y \mid \mathbf{X})$. **(D) The *CLOUD* Objective:** Training loss combining positive target alignment $\mathcal{L}_{\text{pos}}$ with distractor penalties $\mathcal{L}_{\text{neg}}$. **(E) Divergence-Gated Suffix Penalty:** Restricts $\mathcal{L}_{\text{neg}}$ to divergent suffix tokens ($t > k(c)$), preserving shared label prefixes.

whether meta-training improves fine-grained discrimination from demonstrations with variable visual alignment to the query.

Given a query image x_q with label $y_q \in \mathcal{Y}_{\text{all}}$, let $\mathcal{D}_k$ denote the exemplar pool for class k. Using a frozen OpenAI CLIP ViT-L/14 image encoder f_ϕ Radford et al. (2021), we compute each class centroid as the mean of its normalized image embeddings: $\mathbf{c}_k = \frac{1}{|\mathcal{D}_k|} \sum_{x \in \mathcal{D}_k} \frac{f_\phi(x)}{\|f_\phi(x)\|_2}$. We select the K classes with the highest cosine similarity to the query: $\mathcal{Y}_{\text{ctx}} = \text{TopK}_{k \in \mathcal{Y}_{\text{all}}} \frac{\langle f_\phi(x_q), \mathbf{c}_k \rangle}{\|f_\phi(x_q)\|_2 \|\mathbf{c}_k\|_2}$. Retrieval does not use the query's ground-truth label, and the ground-truth class is not forcibly included in $\mathcal{Y}_{\text{ctx}}$. All queries are retained for evaluation, including those whose ground-truth class is absent from the retrieved context. For each selected class k, we randomly sample M labeled exemplars from $\mathcal{D}_k$, without ranking individual images by their similarity to the query. We then randomly shuffle the resulting $N = KM$ demonstrations to obtain the support sequence $\mathcal{S} = \{(x_i, y_i)\}_{i=1}^N$. The multimodal input consists of these demonstrations followed by the query (Fig. 2(B)):

$$\mathbf{X} = \big[\mathcal{T}_{\text{vis}}(x_1), \mathcal{T}_{\text{txt}}(y_1), \ldots, \mathcal{T}_{\text{vis}}(x_N), \mathcal{T}_{\text{txt}}(y_N), \mathcal{T}_{\text{vis}}(x_q), \mathcal{T}_{\text{txt}}(\emptyset)\big], \tag{1}$$

where $\mathcal{T}_{\text{vis}}$ and $\mathcal{T}_{\text{txt}}$ denote the visual encoding and text formatting operations, respectively, and $\mathcal{T}_{\text{txt}}(\emptyset)$ marks the query answer position.

Length-Normalized Conditional Scoring. For a candidate label tokenized as $y = (y_1, \ldots, y_{L_y})$, where L_y is its token length, the model assigns the autoregressive conditional likelihood $P_\theta(y \mid \mathbf{X}) = \prod_{t=1}^{L_y} P_\theta(y_t \mid \mathbf{X}, y_{<t})$, where θ denotes the model parameters. Following prior work on candidate ranking Brown et al. (2020), we score each label by its mean token log-probability to reduce length effects (Fig. 2(C)):

$$\mathcal{S}_\theta(y \mid \mathbf{X}) = \frac{1}{L_y} \sum_{t=1}^{L_y} \log P_\theta(y_t \mid \mathbf{X}, y_{<t}). \tag{2}$$

Prediction is performed over the full task label space: $\hat{y} = \arg\max_{y \in \mathcal{Y}_{\text{all}}} \mathcal{S}_\theta(y \mid \mathbf{X})$. Thus, $\mathcal{Y}_{\text{ctx}}$ determines the classes represented in the demonstrations, while $\mathcal{Y}_{\text{all}}$ defines the prediction space.

3.2 Candidate Discrimination under Cross-Entropy Training

CE-based meta-training maximizes the likelihood of the correct label y^* of token length L_*: $\mathcal{L}_{\text{CE}} = -\frac{1}{L_*} \sum_{t=1}^{L_*} \log P_\theta(y_t^* \mid \mathbf{X}, y_{<t}^*)$. At each step, softmax normalization suppresses alternative tokens

under the correct prefix. However, CE does not directly supervise a distractor label's continuation once its prefix differs from the target.

For in-context distractor labels $\mathcal{N} = \mathcal{Y}_{\text{ctx}} \setminus \{y^*\}$, we measure candidate separation by

$$\mathcal{M} = \mathcal{S}_\theta(y^* \mid \mathbf{X}) - \max_{c \in \mathcal{N}} \mathcal{S}_\theta(c \mid \mathbf{X}). \tag{3}$$

In Fig. 1(A), CE tuning raises both target and distractor scores: the margin improves from 0.665 to 0.978, but the distractor increase partially offsets the target-score gain. The attention analysis in Fig. 1(B) additionally shows candidate-dependent matching to demonstrations. These observations motivate *CLOUD*: an objective that explicitly trains the model to support the correct category and reject distractor label continuations. Diagnostic details appear in Appendix A.

3.3 CLOUD: Contrastive Likelihood Only Upon Divergence

We propose *CLOUD*, a candidate-discriminative objective for fine-grained visual in-context learning that promotes the correct label and penalizes distractor continuations beyond their shared prefixes with the target. Positive supervision uses the target-label cross-entropy defined in Section 3.2, $\mathcal{L}_{\text{pos}} = \mathcal{L}_{\text{CE}}$. Negative supervision uses a $-\log(1-p)$ penalty to suppress incorrect-token probabilities; its application must account for prefixes shared with the target.

Shared-Prefix Gradient Conflict. At shared-prefix positions, positive supervision and an ungated negative penalty act on the same conditional token probability p_θ. Their local contribution is $\mathcal{L}_{\text{shared}} = -a \log p_\theta - b \log(1 - p_\theta)$, giving

$$\nabla_\theta \mathcal{L}_{\text{shared}} = \left(-a + \frac{b p_\theta}{1 - p_\theta} \right) \nabla_\theta \log p_\theta, \tag{4}$$

where $a, b > 0$ collect the positive and aggregate negative weights, including normalization. In the unclipped region, these contributions cancel at $p_\theta = a/(a+b)$; above this threshold, the coefficient reverses sign relative to positive supervision alone. Excluding shared-prefix tokens removes this direct conflict, although parameter sharing can still couple updates across positions.

Divergence-Gated Negative Supervision. For each distractor label $c \in \mathcal{N}$ of token length L_c, let $k(c)$ denote the length of its longest common token prefix with y^*, and $p_{c,t} = P_\theta(c_t \mid \mathbf{X}, c_{<t})$. We penalize the remaining suffix (Fig. 2(E)):

$$\mathcal{L}_{\text{neg}}(c) = \begin{cases} -\dfrac{1}{L_c - k(c)} \displaystyle\sum_{t=k(c)+1}^{L_c} \log[1 - \min(p_{c,t}, 1-\epsilon)], & L_c > k(c), \\ 0, & L_c = k(c). \end{cases} \tag{5}$$

CLOUD couples context-derived negative supervision with prefix-aware gating for fine-grained discrimination. It reinforces the target label while penalizing competing continuations only beyond shared prefixes, using a negative log-complement penalty Kim et al. (2019); Welleck et al. (2020). We set $\epsilon = 10^{-6}$ to avoid numerical singularities in the log-complement computation. Tokens that coincide after divergence remain supervised because their conditioning prefixes differ. Combining target-label supervision with the average divergence-gated distractor penalty yields the *CLOUD* objective (Fig. 2(D)):

$$\mathcal{L}_{\text{CLOUD}} = \mathcal{L}_{\text{pos}} + \underbrace{\frac{1}{|\mathcal{N}|} \sum_{c \in \mathcal{N}} \mathcal{L}_{\text{neg}}(c)}_{\overline{\mathcal{L}}_{\text{neg}}}. \tag{6}$$

The negative term is defined as zero when $\mathcal{N} = \emptyset$. The objective suppresses competing labels collectively, rather than only the current highest-scoring distractor, since any of them may limit the target's discriminative margin.

3.4 Gradient Properties and Computational Cost

Probability-Adaptive Negative Gradients. Let $p = P_\theta(c_t \mid \mathbf{X}, c_{<t})$. In the unclipped region $p < 1 - \epsilon$, the suffix penalty gives:

$$\frac{\partial \ell}{\partial z_t(c_t)} = p, \qquad \frac{\partial \ell}{\partial z_t(v)} = -\frac{p}{1-p} P_\theta(v \mid \mathbf{X}, c_{<t}), \quad v \neq c_t. \tag{7}$$

In this region, higher-probability incorrect tokens receive stronger rejection signals. Since $P_\theta(v \mid \mathbf{X}, c_{<t}) \leq 1 - p$ for $v \neq c_t$, all logit derivatives have magnitude at most $p \leq 1$, independently of clipping. These bounds concern logits, not network parameters. Above the clipping threshold, this penalty contributes no gradient; the boundary is nondifferentiable. At a fixed prefix, the negative term distributes compensating logit updates across the remaining vocabulary in proportion to current token probabilities, rather than concentrating them on a designated alternative. This provides a distributed rejection signal without explicitly promoting another in-context label.

Training and Inference Cost. *CLOUD* adds no model parameters or inference stages and retains the CE baseline's candidate-scoring procedure. During training, shared-context KV states can be reused across candidates, following the shared-prefix reuse principle Juravsky et al. (2024), with the computation graph retained for backpropagation. Candidate continuations can be scored in parallel through teacher forcing; additional computation concerns these continuations and their gradients. Identifying each shared-prefix boundary costs $\mathcal{O}(\min(L_c, L_*))$.

4 Experiments

4.1 Experimental Setup

Meta-Generalization Protocol and Benchmarks. We meta-train on only 5,549 query–context pairs from Oxford Pets (37 classes; 3,669 pairs) Parkhi et al. (2012) and DTD (47 classes; 1,880 pairs) Cimpoi et al. (2014), and evaluate on nine datasets excluded from meta-training without downstream parameter updates. These comprise six natural fine-grained benchmarks—Flowers102 Nilsback & Zisserman (2008), Food-101 Bossard et al. (2014), FGVC-Aircraft Maji et al. (2013), CUB-200 Wah et al. (2011), Stanford Cars Krause et al. (2013), and NABirds Van Horn et al. (2015)—and three specialized classification benchmarks: CropDisease Mohanty et al. (2016), Hyper-Kvasir Borgli et al. (2020), and MVTec AD Bergmann et al. (2019), covering agriculture, medical imaging, and industrial inspection, respectively. These specialized domains test context-utilization transfer under larger shifts from the meta-training sources. Cross-dataset averages are unweighted means.

Unless otherwise stated, we use Centroid-Random (C-R; Section 3.1) with four classes and one randomly sampled exemplar per class. Evaluation contexts are shared across methods; predictions use full-label length-normalized scoring.

Baselines and Implementation. We evaluate Qwen3-VL-2B/4B Bai et al. (2025) and Gemma3-4B Kamath et al. (2025). Zero-Shot and Vanilla ICL use the pretrained models without and with demonstrations, respectively; Meta-ICL uses CE-based meta-training Min et al. (2022). To isolate the contribution of distractor supervision to meta-generalization, *CLOUD* and Meta-ICL share backbones, source data, LoRA configurations, and evaluation contexts. We further compare supervision objectives and test robustness across retrieval strategies.

We apply LoRA Hu et al. (2021) ($r = 16$, $\alpha = 32$, dropout 0.05) to all linear projections in the vision encoder, projector, and language model, keeping embeddings and output heads frozen. We train for three epochs on eight H20 GPUs with AdamW Loshchilov & Hutter (2017), with a peak learning rate of 10^{-5} and cosine decay with warmup. Dataset and prompt details appear in Appendix B.

4.2 Main Results on Held-Out Fine-Grained Datasets

Table 1 evaluates whether candidate-discriminative meta-training transfers to fine-grained datasets absent from training. *CLOUD* improves over matched CE-based Meta-ICL on every dataset across all three backbones, with mean gains of 2.21%, 1.74%, and 1.76%, respectively. Qwen3-VL-2B gains 4.25% on CUB-200 and 2.72% on Stanford Cars. With only 5,549 source training pairs, *CLOUD*-2B also surpasses Vanilla ICL-32B by 6.04% on average, highlighting the value of explicitly training candidate discrimination for transferable context utilization.

Table 1: Accuracy (%) on six held-out natural FGVC benchmarks under C-R. Δ_{meta} is the absolute mean gain over same-backbone Meta-ICL. Bold marks the best result within each trained-backbone group. The 32B baselines use no meta-training.

Model	Method	Flowers	Food	Aircraft	CUB-200	Cars	NABirds	Mean	Δ_{meta}
Qwen3-VL-32B	Zero-Shot	68.40	73.30	23.50	51.40	23.70	36.50	46.13	–
	Vanilla ICL	89.70	78.00	48.50	62.10	79.10	61.50	69.82	–
Qwen3-VL-2B	Zero-Shot	53.46	69.80	15.72	12.20	6.39	11.23	28.14	–
	Vanilla ICL	72.13	76.13	42.03	30.41	79.16	38.10	56.33	–
	Meta-ICL	96.44	88.56	54.37	60.82	84.42	57.31	73.65	–
	CLOUD	**96.94**	**90.22**	**56.20**	**65.07**	**87.14**	**59.58**	**75.86**	**+2.21**
Qwen3-VL-4B	Zero-Shot	46.07	64.82	10.08	6.04	3.77	4.81	22.60	–
	Vanilla ICL	88.96	84.27	44.22	47.72	85.56	49.38	66.69	–
	Meta-ICL	96.10	90.53	52.36	67.43	88.40	62.13	76.16	–
	CLOUD	**97.43**	**90.93**	**56.26**	**69.18**	**88.78**	**64.80**	**77.90**	**+1.74**
Gemma3-4B	Zero-Shot	35.58	67.26	15.72	16.85	26.35	10.62	28.73	–
	Vanilla ICL	74.70	73.27	31.32	25.06	71.83	26.56	50.46	–
	Meta-ICL	92.06	86.90	46.77	53.31	80.75	47.86	67.94	–
	CLOUD	**93.17**	**87.85**	**50.44**	**54.35**	**83.86**	**48.55**	**69.70**	**+1.76**

Table 2: Accuracy (%) on three held-out specialized domains under C-R. Δ_{meta} is the absolute mean gain over same-backbone Meta-ICL. Bold marks the best result within each trained-backbone group. The 32B baselines use no meta-training.

Model	Method	CropDisease	Hyper-Kvasir	MVTec	Mean	Δ_{meta}
Qwen3-VL-32B	Zero-Shot	16.50	14.40	14.10	15.00	–
	Vanilla ICL	46.90	46.50	63.90	52.43	–
Qwen3-VL-2B	Zero-Shot	4.71	11.94	3.97	6.88	–
	Vanilla ICL	42.46	30.18	39.11	37.25	–
	Meta-ICL	54.37	43.09	47.69	48.39	–
	CLOUD	**62.89**	**47.81**	**51.99**	**54.23**	**+5.84**
Qwen3-VL-4B	Zero-Shot	2.52	16.07	6.20	8.26	–
	Vanilla ICL	44.18	48.84	43.72	45.58	–
	Meta-ICL	60.79	52.69	55.17	56.22	–
	CLOUD	**69.01**	**55.97**	**62.00**	**62.33**	**+6.11**
Gemma3-4B	Zero-Shot	10.40	4.32	10.97	8.56	–
	Vanilla ICL	40.54	21.67	34.34	32.18	–
	Meta-ICL	57.58	31.80	39.90	43.09	–
	CLOUD	**59.19**	**35.33**	**41.18**	**45.23**	**+2.14**

4.3 Generalization to Specialized Domains

Table 2 extends the evaluation from natural fine-grained recognition to industrial, agricultural, and medical classification, with models trained only on pets and textures and no downstream parameter updates. Notably, *CLOUD*'s advantage over Meta-ICL increases on all three backbones: mean gains rise from 2.21%/1.74%/1.76% on natural benchmarks to 5.84%/6.11%/2.14% in these specialized domains. This consistent widening supports the value of candidate-discriminative training for cross-domain meta-generalization. *CLOUD*-2B also surpasses Vanilla ICL-32B by 1.80% across these domains and 4.63% across all nine datasets. Appendix D presents qualitative successes and failures across all nine datasets.

5 Robustness and Ablation Analyses

We assess *CLOUD*'s robustness to changes in context composition (K and M), retrieval strategy, and prediction protocol on Qwen3-VL-2B across six held-out natural fine-grained datasets. These analyses test whether its gains over Meta-ICL persist across evaluation conditions. Unless otherwise stated, we reuse the checkpoints trained under the default setting and modify only the evaluation conditions. We further assess robustness to label scrambling and contextual fidelity under counterfactual label binding, before examining objective design and prefix gating through separately trained ablations. Reported accuracies are unweighted dataset means; additional settings are described in Appendix C.

5.1 Robustness to Context Composition

Table 3 evaluates robustness to changes in context composition while keeping prediction over the full task label space. Increasing K broadens ground-truth coverage but also introduces more competing demonstrations. Accuracy nevertheless decreases for both trained models, suggesting that additional context is not automatically beneficial when it introduces further competition. Our four-class default provides multiple competing categories within a limited demonstration budget. As K increases

Table 3: Accuracy (%) under varying context structures on 6 natural benchmarks.

Scaling Axis	Configuration	Demos ($\lvert\mathcal{S}\rvert$)	Vanilla ICL	Meta-ICL	CLOUD	Δ_{meta}
Context Class Scaling (*Scaling K-way*, 1-shot)	2-way 1-shot	2	61.50	78.12	**79.71**	**+1.59**
	4-way 1-shot	4	56.33	73.65	**75.86**	**+2.21**
	6-way 1-shot	6	55.04	71.17	**74.05**	**+2.88**
	8-way 1-shot	8	54.31	69.28	**73.08**	**+3.80**
Evidence Scaling (*Scaling M-shot*, 4-way)	4-way 1-shot (1 Support / Class)	4	56.33	73.65	**75.86**	**+2.21**
	4-way 2-shot (2 Supports / Class)	8	68.45	76.61	**77.59**	**+0.98**
	4-way 3-shot (3 Supports / Class)	12	70.26	77.70	**78.47**	**+0.78**

Table 4: Accuracy (%) across context retrieval protocols on 6 natural FGVC benchmarks.

Retrieval Protocol	Support Sampling Mechanism	Vanilla ICL	Meta-ICL	CLOUD	Δ_{meta}
C-R (Default Anchor)	Centroid class retrieval + random sampling	56.33	73.65	**75.86**	**+2.21**
C-N (Centroid-Nearest)	Centroid class retrieval + feature-nearest exemplar	57.85	75.26	**77.89**	**+2.63**
I-kNN (Instance-kNN)	Exhaustive instance-level kNN gallery search	57.31	74.65	**77.26**	**+2.61**

from 2 to 8, *CLOUD*'s advantage over Meta-ICL grows from 1.59% to 3.80%, indicating greater robustness to this change in context composition.

With four classes fixed, increasing exemplars per class improves all methods. *CLOUD* remains consistently stronger than Meta-ICL, with gains of 2.21%, 0.98%, and 0.78% for one, two, and three exemplars, respectively. Its largest gain occurs in the one-shot setting. Together, these trends support the value of candidate-discriminative training when models must distinguish competing categories from limited evidence, improving context utilization without requiring additional demonstrations.

5.2 Robustness across Retrieval Strategies

Table 4 compares three context construction protocols: C-R retrieves classes by centroid similarity and randomly samples one exemplar per class; C-N uses the same class retrieval but selects the nearest exemplar within each class; and I-kNN retrieves the nearest exemplars from distinct classes across the support gallery. Within each protocol, all methods use identical evaluation contexts.

Nearest-neighbor selection improves accuracy over C-R for all methods, while *CLOUD* retains gains of 2.21%, 2.63%, and 2.61% over Meta-ICL under C-R, C-N, and I-kNN, respectively. Its advantage therefore extends beyond random exemplar selection and remains evident with more closely matched demonstrations. These results support improving context construction and candidate discrimination as complementary approaches to fine-grained in-context classification.

5.3 Robustness across Prediction Protocols

Table 5 tests whether *CLOUD*'s improvements extend beyond likelihood-based label ranking. In addition to the default full-label scoring, we evaluate greedy free generation and four-way multiple-choice prediction. *CLOUD* outperforms Meta-ICL by 2.21%, 1.00%, and 2.25%, respectively, showing that its advantage persists when predictions are produced through generation or option selection. Full-label scoring remains our default because it evaluates complete candidate labels with length normalization. Greedy generation makes locally optimal token choices conditioned on its own preceding outputs, so early decisions can constrain the final answer without a comparison of complete labels. Multiple-choice prediction is restricted to the four retrieved classes; queries with missing ground-truth classes are counted as incorrect. The alternative protocols thus test robustness under different prediction constraints, rather than replace the default evaluation.

5.4 Label Semantics and Contextual Fidelity

Table 6 stress-tests context utilization through two inference-time perturbations. **Scrambled Labels** permutes characters within category names to disrupt familiar lexical cues. **Counterfactual Binding** cyclically reassigns labels across support classes. If the query's ground-truth class is among the retrieved support classes, its target label is remapped using the same permutation; otherwise, its target label remains unchanged. All queries are included in evaluation. This creates a particularly demanding setting: the model must follow demonstration-defined mappings that directly contradict familiar category meanings.

Under scrambled labels, *CLOUD* loses only 2.60%, compared with 5.88% for Meta-ICL and 10.85% for Vanilla ICL, expanding its advantage over Meta-ICL from 2.21% to 5.49%. Counterfactual

Table 5: Accuracy (%) across evaluation readout protocols on 6 natural FGVC benchmarks.

Readout Mechanism	Operational Strategy	Vanilla ICL	Meta-ICL	CLOUD	Δ_{meta}
Full-label Likelihood (Ours)	Length-normalized log-likelihood ranking	56.33	73.65	**75.86**	+2.21
Free Generation	Greedy autoregressive string decoding	49.24	73.06	**74.06**	+1.00
4-way MCQ	Option selection (A/B/C/D) over retrieved candidate classes	64.04	71.87	**74.11**	+2.25

Table 6: Counterfactual perturbation and label semantic decoupling across 6 natural FGVC benchmarks. All perturbations are applied only at inference.

Method	Standard Test (%)	Scrambled Labels (%)	Counterfactual Binding (%)
Vanilla ICL	56.33	45.48	14.04
Meta-ICL	73.65	67.77	31.07
CLOUD	**75.86**	**73.26**	**43.62**
Δ_{meta}	**+2.21**	**+5.49**	**+12.55**

Table 7: Accuracy (%) across loss formulations and divergence gating on 6 natural FGVC benchmarks. Models are independently re-trained, differing solely in loss formulations.

Objective Formulation	Negative Loss Formulation (ℓ_{neg})	Average Accuracy (%)
Standard Meta-ICL	None ($\mathcal{L}_{\text{pos}}$ only)	73.65
Unbounded Negative Loss	$+\log P_\theta(c_t \mid \cdot)$	66.81
Whole-Label InfoNCE	Sequence-level normalized contrast	69.00
CLOUD (Standard Training)	Divergence-gated bounded log-barrier	**75.86**
Stress Test on Hierarchical Shared Prefixes (Controlled Multi-level Prefix Training)		
CLOUD w/ Prefix Gate	Bounded penalty restricted strictly after divergence index	75.53
CLOUD w/o Prefix Gate	Naive whole-string suppression across all tokens	70.07

binding causes a substantial accuracy drop for all methods, underscoring the difficulty of overriding familiar associations through context alone. Despite this direct conflict, *CLOUD* achieves 43.62% accuracy versus 31.07% for Meta-ICL, widening its advantage to 12.55%. These results demonstrate higher contextual fidelity under counterfactual bindings: *CLOUD* more effectively follows demonstration-defined mappings even when they directly contradict familiar category meanings.

5.5 Objective Design and Prefix Gating

Table 7 examines the loss formulation and shared-prefix gating through separately trained variants under matched optimization settings. *CLOUD* achieves 75.86% accuracy, compared with 73.65% for CE-only Meta-ICL and 69.00% for whole-label InfoNCE. These results show that adding candidate-level competition alone does not ensure improved classification; the formulation and placement of negative supervision matter.

To isolate the role of prefix gating, we introduce multi-level shared prefixes into meta-training labels while retaining the original category names during downstream evaluation. Under this stress test, gated training achieves 75.53%, compared with 70.07% when all negative-label tokens are penalized—a 5.46% absolute accuracy difference. This result supports the shared-prefix conflict analysis in Section 3.3: excluding predictions shared with the target is important for learning effective rejection of distractor continuations. The benefit transfers to downstream labels without the synthetic prefixes. The label-scrambling results in Section 5.4 further suggest that meaningful category names provide useful semantic cues, which could be weakened by renaming labels to eliminate shared prefixes. Prefix gating retains the original label structure and its positive supervision while restricting negative supervision to continuations beyond the shared prefix.

6 Conclusion

We presented *CLOUD*, a candidate-discriminative training objective that strengthens context utilization for fine-grained visual classification. By jointly promoting correct labels and penalizing distractor continuations beyond shared prefixes, *CLOUD* explicitly trains models to recognize both what a query is and what it is not. Meta-trained on only 5,549 query–context pairs, it improves over CE-based Meta-ICL across nine held-out datasets, with larger gains in specialized domains and higher contextual fidelity under label perturbations. Its 2B variant also surpasses Vanilla ICL-32B on average while preserving the baseline inference procedure. These findings highlight data-efficient training of context utilization as an effective route to cross-domain meta-generalization.

AI USE STATEMENT

We used generative AI tools to assist with literature search, manuscript drafting and editing, translation, figure preparation, and code development for training, evaluation, and supporting scripts. These tools also supported refinement of methodological explanations and interpretation of experimental results. The authors reviewed the AI-assisted text and figures, checked references against the original sources, and verified and tested the AI-assisted code used in this work. We take full responsibility for the final content of the paper, including its claims, results, code, and figures.

REFERENCES

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

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.

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592. IEEE, 2019.

Hanna Borgli, Vajira Thambawita, Pia H Smedsrud, Steven Hicks, Debesh Jha, Sigrun L Eskeland, Kristin Ranheim Randel, Konstantin Pogorelov, Mathias Lux, Duc Tien Dang Nguyen, et al. Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data*, 7(1):283, 2020.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pp. 446–461. Springer, 2014.

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.

Haokun Chen, Xu Yang, Yuhang Huang, Zihan Wu, Jing Wang, and Xin Geng. Manipulating the Label Space for In-Context Classification, 2023.

Huiyi Chen, Jiawei Peng, Kaihua Tang, Xin Geng, and Xu Yang. Enhancing multimodal in-context learning for image classification through coreset optimization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5130–5139, 2025a.

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*, 2025b.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Abhimanyu Dubey, Otkrist Gupta, Pei Guo, Ramesh Raskar, Ryan Farrell, and Nikhil Naik. Pairwise confusion for fine-grained visual classification. In *European conference on computer vision*, pp. 71–88. Springer, 2018.

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

Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 4476–4484. IEEE, 2017.

Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-R1: Make Multi-modal LLMs Excel in Fine-Grained Visual Recognition by Chain-of-Thought Reasoning. In *ICLR*, 2026.

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

Jordan Juravsky, Bradley Brown, Ryan Ehrlich, Daniel Y. Fu, Christopher Ré, and Azalia Mirhoseini. Hydragen: High-Throughput LLM Inference with Shared Prefixes, 2024.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 4, 2025.

Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

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

Wei Luo, Xitong Yang, Xianjie Mo, Yuheng Lu, Larry S Davis, Jun Li, Jian Yang, and Ser-Nam Lim. Cross-x learning for fine-grained visual categorization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8242–8251, 2019.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. In *Proceedings of the 2022 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2791–2809, 2022.

Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:215232, 2016.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

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. In *NeurIPS*, 2024.

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*, pp. 8748–8763. PmLR, 2021.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 595–604, 2015.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Xiu-Shen Wei, Yi-Zhe Song, Oisin Mac Aodha, Jianxin Wu, Yuxin Peng, Jinhui Tang, Jian Yang, and Serge Belongie. Fine-grained image analysis with deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8927–8948, 2021.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural Text Generation with Unlikelihood Training. In *ICLR*, 2020.

Guangzhi Xiong, Sanchit Sinha, Zhenghao He, and Aidong Zhang. Retrieving counterfactuals improves visual in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24352–24362, 2026.

Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *European conference on computer vision*, pp. 834–849. Springer, 2014.

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 *International Conference on Learning Representations*, 2024.

A Details of the Motivational Experiment

The diagnostic compares Qwen3-VL-2B under Vanilla ICL, Meta-ICL, and *CLOUD* on 70,223 retrieval-HIT queries from the six natural datasets, using matched contexts (Fig. 3). Each context contains one ground-truth (GT) class demonstration and three incorrect-class demonstrations. Statistics are averaged within each dataset and then across datasets with equal weight.

A.1 Candidate-Score Analysis

For each query, GT denotes the correct-label score, Hardest the maximum score among the three incorrect context labels, and In-context their mean score. To define Outside, each model $m \in \{\text{Vanilla ICL}, \text{Meta-ICL}, \textit{CLOUD}\}$ selects its highest-scoring label outside both the context and GT:

$$c_m^{\text{out}} = \operatorname*{arg\,max}_{c \in \mathcal{Y}_{\text{all}} \setminus (\mathcal{Y}_{\text{ctx}} \cup \{y^*\})} \mathcal{S}_m(c \mid \mathbf{X}), \qquad \mathcal{O}_q = \bigcup_m \{c_m^{\text{out}}\}. \tag{8}$$

Ties are resolved by class index. The deduplicated set $\mathcal{O}_q$ contains one to three labels and is frozen for paired comparison: every model is scored on the same set, with $S_m^{\text{out}}(q) = |\mathcal{O}_q|^{-1} \sum_{c \in \mathcal{O}_q} \mathcal{S}_m(c \mid \mathbf{X})$.

These query-level log-scores are averaged within each dataset and then equally across the six datasets; the figure displays $\exp(\overline{\mathcal{S}})$. The reported margin is the corresponding average of the GT-minus-Hardest log-score difference, $\mathcal{M}$ (Section 3.2), rather than a difference between the exponentiated bars. Outside measures scores assigned to strong outside-context alternatives, which need not be the models' actual predictions. On the diagnostic HIT queries, predictions fall outside the context for 16.8% of Vanilla ICL, 0.10% of Meta-ICL, and 0.44% of *CLOUD* cases. These scores are not normalized probabilities over candidate labels.

CE meta-training lowers outside-context scores while raising both correct-label and in-context distractor scores. This shows that a stronger preference for context labels does not necessarily yield better separation among them: competing labels can benefit alongside the target, partially offsetting gains in the discriminative margin. This distinction matters for fine-grained classification, where prediction depends on relative candidate scores rather than the target score alone.

Relative to Meta-ICL, *CLOUD* lowers both in-context and outside-context distractor scores and enlarges the margin between the correct label and the hardest in-context distractor. This reinforces the main-paper argument that explicit negative supervision can suppress the score support received by competing context labels, giving the correct class a clearer relative advantage among similar alternatives. The following attention analysis examines the demonstration-use patterns accompanying this scoring difference.

A.2 Attention Allocation to Demonstrations

Let $a_{\ell h}(r, j)$ denote the attention weight from scoring position r to key position j in layer ℓ and head h. For demonstration i, let $\mathcal{I}_i$ and $\mathcal{T}_i$ be its image and label token spans. Its combined attention mass is

$$m_{\ell h i}(r) = \sum_{j \in \mathcal{I}_i \cup \mathcal{T}_i} a_{\ell h}(r, j). \tag{9}$$

This treats the image and label as one demonstration and measures its share of attention allocated to the four demonstrations.

Decision-position attention. Fig. 3(B) uses the last prompt position r_0, which predicts the first answer token. At each layer, we normalize within each head and then average over the $H = 16$ heads:

$$A_{\ell i}^{\text{dec}} = \frac{1}{H} \sum_{h=1}^{H} \frac{m_{\ell h i}(r_0)}{\sum_{u=1}^{4} m_{\ell h u}(r_0)}. \tag{10}$$

We select the GT demonstration's share, average over queries within each dataset, and then equally across the six datasets. The curves cover all 28 layers; mean GT shares over layers 17–26 are 0.397 for Vanilla ICL, 0.449 for Meta-ICL, and 0.515 for *CLOUD*. Meta-ICL increases attention to the GT

Table 8: Attention allocation while scoring each model's strongest incorrect in-context label. Own support denotes the demonstration carrying that incorrect label; GT support depicts the correct class.

Model	Own support	GT support	Other two supports
Meta-ICL	0.4102	0.3074	0.2824
CLOUD	0.2861	0.4669	0.2471

Table 9: Change in mean token log-likelihood after attention blocking (nats). The scored label is specified separately from the blocked demonstration.

Scored label	Blocked support	Meta-ICL	*CLOUD*
Incorrect candidate	Its own demonstration	-1.7151	-0.3040
Incorrect candidate	Control demonstration	+0.0249	+0.0041
Correct label	GT demonstration	-1.3893	-1.3457

demonstration relative to Vanilla ICL, and *CLOUD* further increases this share, indicating a stronger focus on correct-class evidence before predicting the first answer token. Yet Meta-ICL's increased GT attention coexists with the elevated distractor scores observed in the main paper, suggesting that greater initial attention to the correct demonstration alone does not ensure effective discrimination among competing labels. We therefore examine whether attention remains focused on GT evidence during token-by-token scoring of an incorrect candidate or shifts toward that candidate's own demonstration.

Candidate-scoring attention. Table 8 instead uses the positions that predict the tokens of each model's highest-scoring incorrect context label. These include r_0 for the first token and the preceding candidate-token positions for subsequent tokens; the position after the complete label is excluded. For each scoring position, attention is averaged over heads and deep layers before normalization:

$$\bar{m}_i(r) = \frac{1}{10H} \sum_{\ell=17}^{26} \sum_{h=1}^{H} m_{\ell h i}(r), \qquad A_i^{\text{score}}(r) = \frac{\bar{m}_i(r)}{\sum_{u=1}^{4} \bar{m}_u(r)}. \tag{11}$$

Own support denotes the demonstration carrying the scored incorrect label; GT support denotes the correct-class demonstration. GT support is identified by class membership for analysis and is not marked in the prompt. The incorrect candidate is selected separately for each model, so these statistics characterize attention while scoring each model's hardest in-context alternative.

Candidate-scoring attention reveals a further distinction: Meta-ICL attends more to the incorrect candidate's own demonstration, whereas *CLOUD* attends more to the GT demonstration. This echoes Fig. 1(B): if attention follows the candidate being scored, competing labels may each find support in their corresponding demonstrations without being adequately distinguished.

The greater attention to GT evidence under an incorrect candidate is consistent with more comparative use of demonstrations by *CLOUD*. It also aligns with the objective design: direct supervision of incorrect-label continuations trains rejection under the incorrect candidate's own conditioning history. The next experiment tests whether blocking access to these demonstrations changes candidate scores.

A.3 Dependence on Individual Demonstrations

We block attention from candidate-scoring positions to a selected demonstration's image and label tokens, then renormalize the remaining attention weights. For candidate c and blocked demonstration b, the score change is

$$\Delta\mathcal{S}(c; b) = \mathcal{S}_\theta^{\text{block}(b)}(c \mid \mathbf{X}) - \mathcal{S}_\theta(c \mid \mathbf{X}). \tag{12}$$

A negative value indicates that blocking the demonstration lowers the candidate score.

In Table 9, blocking the incorrect candidate's own demonstration lowers its score by 1.7151 for Meta-ICL but only 0.3040 for *CLOUD*; control blocking has little effect on either model. The contrast with control blocking indicates that Meta-ICL's dependence is candidate-specific: the demon-

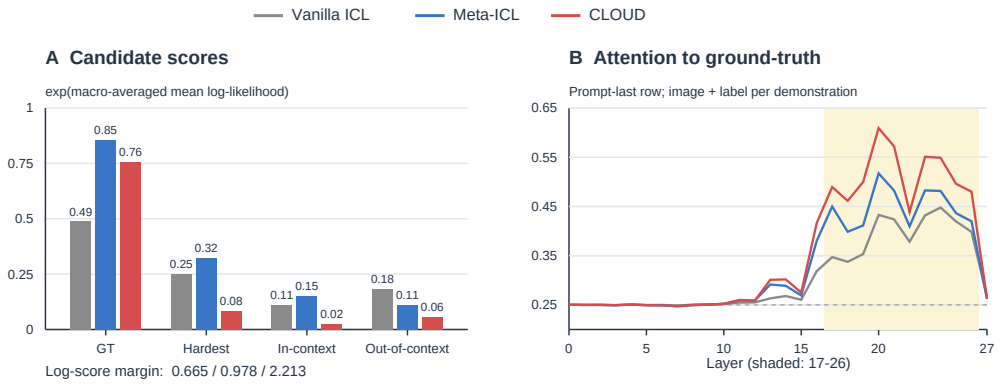


Figure 3: Measurements supporting the main-paper motivation. (A) Candidate scores and margins. (B) Attention from the last prompt position to the ground-truth demonstration, combining its image and label spans. The shaded band denotes layers 17-26; 0.25 indicates equal allocation across four demonstrations.

stration matching the incorrect label provides substantially more support than the control demonstration. This connects the preceding observations—candidate-dependent attention is accompanied by score support from the corresponding demonstration, helping explain the elevated in-context distractor scores.

Meanwhile, blocking the GT demonstration lowers correct-label scores by similar amounts for *CLOUD* and Meta-ICL. Thus, the weaker dependence of incorrect scores on their matching demonstrations is not accompanied by an equally reduced contribution from GT evidence. Together with the greater attention to GT evidence during incorrect-candidate scoring, these results support an interpretation in which *CLOUD* compares competing candidates using evidence across demonstrations, rather than allowing each candidate to draw support primarily from its own match. This more comparative pattern of context use is consistent with the lower distractor scores and greater relative advantage of the correct label observed in the main paper.

B Datasets and Evaluation Settings

B.1 Datasets and Retrieval Coverage

We meta-train on 5,549 query-context pairs from Oxford Pets and DTD and evaluate on nine datasets excluded from meta-training. Support and query images are disjoint, while both pools cover the task classes. Source queries are used for meta-training; downstream queries are used for evaluation without parameter updates. Classification results use the complete query splits. Under the default C-R protocol, the six natural datasets contain 73,200 queries: 70,223 HIT and 2,977 MISS. The motivational diagnostics in Appendix A use the HIT subset.

CropDisease uses 38 PlantVillage crop-disease categories. Hyper-Kvasir uses its labeled images. The original MVTec AD dataset comprises 15 object or texture categories for anomaly detection. We reformulate it as a 73-class classification task by distinguishing object or texture categories and defect types, with disjoint support and query pools drawn from the official test split.

Centroid-Random (C-R) retrieves K classes by CLIP centroid similarity and randomly samples M exemplars per class (Section 3.1). This uses existing benchmarks without directly selecting the most appearance-similar exemplar. Sampling and demonstration order are frozen with seed 42. A retrieval HIT means the correct class is among the K retrieved classes; otherwise the query is a MISS.

Under the default C-R protocol, retrieval hit rates on the source datasets are 99.02% for Oxford Pets and 89.79% for DTD. Table 12 reports coverage on all nine downstream datasets.

C-R retrieves the top four CLIP class centroids and randomly selects a support image within each class; C-N retrieves the same classes but selects the nearest support image within each class by cosine similarity. Their hit rates are therefore identical. I-kNN ranks individual support images,

Table 10: Image counts in the train (support) and test (query) partitions used in this work. Source queries contribute to meta-training; downstream queries are held out. †MVTec uses disjoint support and query subsets of the official test split.

Dataset	Classes	Train images (support)	Test images (query)
Oxford Pets (source)	37	3,680	3,669
DTD (source)	47	1,880	1,880
Flowers102	102	1,020	6,149
Food-101	101	75,750	25,250
FGVC-Aircraft	100	3,334	3,333
CUB-200	200	5,994	5,794
Stanford Cars	196	8,144	8,041
NABirds	555	23,929	24,633
CropDisease	38	43,596	10,709
Hyper-Kvasir	23	7,464	3,198
MVTec†	73	629	629

Table 11: C-R retrieval hit rates (%) on the complete query splits of the six natural datasets. $K \times M$ denotes K classes with M exemplars per class. Coverage is unchanged when varying M for fixed retrieved classes.

Dataset	2×1	4×1	6×1	8×1	4×2	4×3
Flowers102	99.07	99.64	99.72	99.77	99.64	99.64
Food-101	96.76	98.58	99.09	99.34	98.58	98.58
FGVC-Aircraft	72.07	85.15	90.40	93.58	85.15	85.15
CUB-200	91.96	96.76	97.98	98.64	96.76	96.76
Stanford Cars	94.33	98.68	99.30	99.61	98.68	98.68
NABirds	85.25	92.66	95.47	96.78	92.66	92.66

Table 12: Retrieval hit rates (%) for three four-shot protocols ($K = 4$, $M = 1$), evaluated on the complete query splits. A hit means that the correct class is among the four retrieved classes.

Dataset	Queries	C-R	C-N	I-kNN
Flowers102	6,149	99.64	99.64	99.59
Food-101	25,250	98.58	98.58	98.73
FGVC-Aircraft	3,333	85.15	85.15	74.47
CUB-200	5,794	96.76	96.76	94.60
Stanford Cars	8,041	98.68	98.68	98.36
NABirds	24,633	92.66	92.66	88.69
CropDisease	10,709	96.56	96.56	99.61
Hyper-Kvasir	3,198	95.12	95.12	97.75
MVTec	629	96.66	96.66	96.03

selects the first four distinct classes, and retains the corresponding nearest images. It changes class coverage as well as exemplar selection, with dataset-dependent effects (Table 12).

B.2 Prompt Format

Each demonstration is a user-assistant exchange, followed by the query. The following instruction text and message structure are shared by training and evaluation, using each backbone's chat template. Instruction lines are wrapped below for readability.

Classification Instruction

```
SYSTEM:
You are a fine-grained image classification expert.
Look at the photo and answer with the exact class name only,
no other words.

QUESTION:
What fine-grained category is in this photo?
```

```
Answer with the class name only.
```

Demonstration and Query Format

```
system:    {SYSTEM}
user:      [support image 1] {QUESTION}
assistant: {class name 1}
...
user:      [support image KM] {QUESTION}
assistant: {class name KM}
user:      [query image] {QUESTION}
assistant: [candidate answer]
```

Demonstration answers are conditioning context. Positive supervision applies only to the final correct answer. *CLOUD* additionally scores incorrect support labels at this answer position and penalizes their divergent suffixes. Zero-Shot evaluation removes the demonstration exchanges. Full-label likelihood scoring ranks all task labels by length-normalized log-likelihood, without an option list; labels absent from the context remain eligible.

Candidate scoring reuses the shared-context KV cache and has been verified against full-sequence scoring.

B.3 Multiple-Choice Prompt

For four-way multiple-choice (MCQ) evaluation, support exchanges and the system instruction are unchanged. The query text becomes:

Multiple-Choice Query

```
What fine-grained category is in this photo?
Options:
A. {label_1}
B. {label_2}
C. {label_3}
D. {label_4}
Answer with the single letter only.
```

Options are the four support labels, shuffled with Random(shuffle_seed*100003 + query_index). Greedy decoding generates at most two tokens, and the first A-D character is read. Only HIT queries require option prediction; MISS queries are counted as incorrect. Full-query accuracy is $n_{\text{correct,HIT}}/(n_{\text{HIT}} + n_{\text{MISS}})$, computed per dataset before macro-averaging.

C Additional Details of Robustness and Ablation Settings

C.1 Context Composition

We vary $K \in \{2, 4, 6, 8\}$ at $M = 1$, or $M \in \{1, 2, 3\}$ at $K = 4$. HIT contexts contain $K - 1$ incorrect labels and MISS contexts contain K. Increasing K therefore changes both coverage and the number of alternatives.

C.2 Retrieval Strategies

Retrieval follows the C-R, C-N, and I-kNN protocols defined in Appendix B, with identical contexts across methods within each protocol.

C.3 Prediction Protocols

Free generation uses greedy decoding; full-label scoring and MCQ follow the evaluation procedures in Appendix B.

C.4 Label Semantics and Contextual Fidelity

At inference time, we perturb category names using a fixed word-level letter-permutation mapping that preserves separators and capitalization and maps distinct words to distinct outputs. The same mapping is applied to support labels, query targets, and candidate names; images and their order remain unchanged.

Counterfactual binding rotates the four support labels left by one position. For a HIT query, the evaluation target becomes the new label attached to the demonstration from its original class; for a MISS query, the original target is retained. The model must therefore follow the supplied mapping despite its possible conflict with familiar category meanings. This tests contextual fidelity rather than recovery of the original semantic label.

C.5 Objective Variants and Prefix Gating

CE-only Meta-ICL supervises the correct label. Whole-label InfoNCE contrasts the normalized correct-label score with incorrect in-context labels at unit temperature:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathcal{S}_\theta(y^* \mid \mathbf{X}) + \log\left[\exp \mathcal{S}_\theta(y^* \mid \mathbf{X}) + \sum_{c \in \mathcal{N}} \exp \mathcal{S}_\theta(c \mid \mathbf{X})\right]. \tag{13}$$

CLOUD averages clipped $-\log(1-p)$ penalties over each divergent suffix and then over incorrect candidates, with weight one. Probabilities are capped at $1 - 10^{-6}$. A negative label that is a strict prefix of the target contributes zero but remains in the candidate-average denominator. The unbounded alternative uses $+\log p$ for incorrect tokens, whose minimization has no finite lower bound as p approaches zero.

Shared-Prefix Stress Test. We construct a synthetic label space with a fixed hierarchy of shared prefixes, to test divergence-gated negative supervision under controlled shared-prefix conflict (Section 3.3). This label space is used only for meta-training on Oxford Pets (37 classes; 3,669 queries) and DTD (47 classes; 1,880 queries). Downstream evaluation retains the original category names.

Each hierarchical label concatenates four space-separated components: a root, a group identifier, a subgroup identifier, and a leaf. The root is `xqkztv` for every class. Group and subgroup identifiers are randomly generated consonant strings of five and four characters, respectively. The leaf is obtained by deterministic letter permutation within each word of the original class name. For example, `Abyssinian` becomes `xqkztv nbpqb mjng Ybssainnia`, `american bulldog` becomes `xqkztv bhtfh rlmb meacrnia guobdll`, and `banded` becomes `xqkztv dtrvz brxk dednba`. All labels therefore share the root; labels in the same group additionally share the group identifier, and labels in the same subgroup share the first three components.

For a source dataset with C classes, the numbers of groups and candidate subgroups per group are

$$G = \max\Big(2, \text{round}(\sqrt{C})\Big), \qquad U = \max(2, \lfloor C/G \rfloor). \tag{14}$$

Classes are randomly assigned to group–subgroup pairs with replacement, and the resulting mapping remains fixed. These assignments are unrelated to semantic or visual similarity and do not enforce a balanced tree. Oxford Pets has $G = 6$ and $U = 6$, with 22 occupied group–subgroup pairs and sorted group sizes of 3, 4, 4, 5, 10, and 11. DTD has $G = 7$ and $U = 6$, with 27 occupied pairs and sorted group sizes of 2, 3, 6, 7, 8, 9, and 12. Here, U counts candidate subgroups within each group; the occupied-pair counts refer to the entire dataset.

Using the same hierarchical label mapping, the stress test compares divergence-gated, suffix-only negative supervision with whole-label negative supervision (Table 7). For each distractor c, the gate uses the longest common *token* prefix with the target, $k(c)$, and applies the negative penalty only at positions $t > k(c)$. Thus, the boundary is determined after tokenization rather than by character counts or the four textual components. Whole-label negative supervision also penalizes shared-prefix positions. This comparison tests whether excluding those positions avoids the direct conflict with positive target supervision while preserving transfer to downstream label spaces.

D CASE STUDIES OF IN-CONTEXT CANDIDATE DISCRIMINATION

The figures compare the classification results of Vanilla ICL, Meta-ICL, and *CLOUD* across nine datasets, showing query images, contextual demonstrations, candidate-label scores, and each model's prediction. Candidate scores are length-normalized log-likelihoods; higher is better.

Fig. 4 highlights improved separation of competing labels. On Aircraft and MVTec, *CLOUD* predicts correctly despite assigning the target a lower score than Meta-ICL, because competing labels receive still lower scores. On Hyper-Kvasir, the margin over the strongest in-context distractor increases from -0.052 under Meta-ICL to 0.093 under *CLOUD*. These cases illustrate the role of relative candidate separation, beyond increasing target scores alone.

The failure cases in Fig. 5 also reveal challenges that remain for *CLOUD*. In MVTec, all three methods select an incorrect class despite the presence of a correct-class demonstration, illustrating that more effective context utilization does not necessarily translate into accurate recognition of subtle visual differences. The outside-context predictions on Food-101 further show that supervision against in-context distractors does not ensure that the correct class ranks highest across the full label space. While *CLOUD* improves discrimination among competing classes, how best to combine MLLMs' visual recognition, semantic knowledge, and in-context learning capabilities for reliable fine-grained classification remains an open research question.

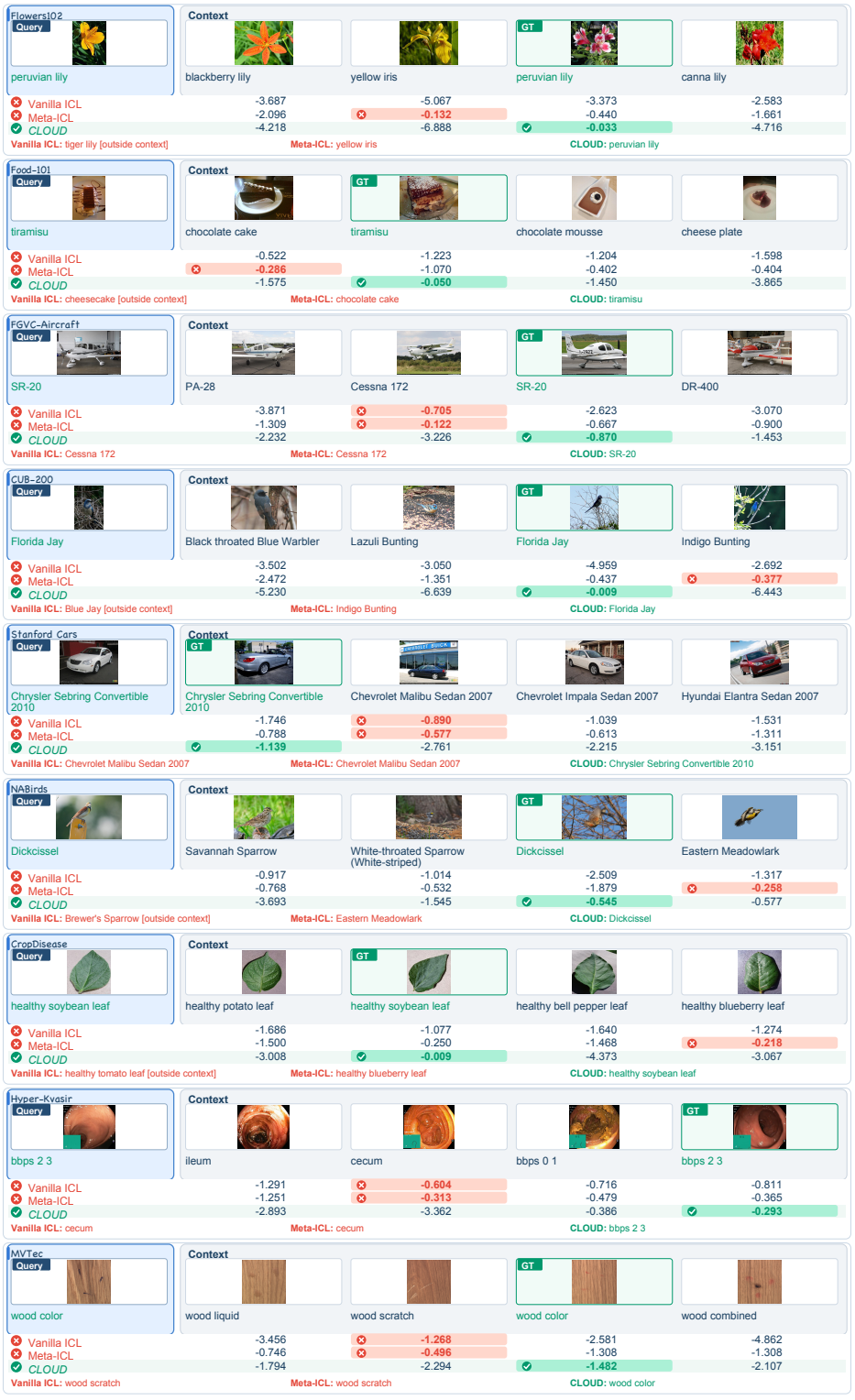


Figure 4: **Qualitative comparisons across nine datasets.** Green checks/red crosses indicate correct/incorrect predictions. Highlighted score cells mark predicted labels present in the context. Each model's full answer appears below its case; outside-context answers are marked explicitly. GT denotes ground truth.



Figure 5: **Failure cases across nine datasets.** Full answers are shown for every model, including labels outside the context. Rows without a highlighted candidate score indicate outside-context predictions; candidate scores and symbols follow Fig. 4.