Concept Mediation Enables Robust Fine-Grained Visual Understanding

Anonymous authors
Paper under double-blind review

Abstract

Large vision-language models exhibit strong general multimodal understanding, yet training-free prompting strategies often fail on fine-grained visual recognition, where correct predictions depend on subtle and localized visual attributes. Existing approaches such as chain-of-thought reasoning and in-context learning often produce fluent explanations or contextual cues without reliably grounding decisions in discriminative visual evidence. To address this issue, we introduce **Concept-Mediated In-Context Learning (CM-ICL)**, a training-free prompting strategy that first extracts visual attribute concepts from the input image and then uses them as structured context for classification. Without training the model, CM-ICL provides an explicit intermediate representation that re-expresses image-derived cues for fine-grained decision making. To evaluate the extracted concepts without manual concept annotations, we combine promptable-segmentation-based perceptual grounding metrics with task-coupled diagnostics that examine how visual localizability relates to downstream prediction behavior. Experiments on six fine-grained datasets show that CM-ICL improves accuracy over training-free approaches, produces more concise and visually localizable concepts, and substantially reduces generation failures. The results demonstrate that concept mediation provides an effective and interpretable route for training-free fine-grained visual recognition.

1 Introduction

Large vision-language models (LVLMs) have achieved remarkable progress in multimodal understanding and can generate coherent chain-of-thought style explanations for vision-language tasks without task-specific supervision (Caffagni et al., 2024; Liang et al., 2024; Achiam et al., 2023; Team et al., 2023; Bai et al., 2025b;a; Zhu et al., 2025). A *de facto* approach for applying LVLMs to downstream tasks is prompting. More specifically, multimodal in-context learning (ICL) has emerged as a training-free prompting paradigm in which LVLMs incorporate a small number of demonstrations into the input context to guide inference, without modifying model parameters (Zong et al., 2024; Chen et al., 2023a). This makes ICL an appealing way to adapt LVLMs to new tasks at inference time.

However, this apparent flexibility weakens sharply for *fine-grained visual recognition* (FGVR), where correct decisions hinge on subtle, localized attributes rather than coarse categories (Wei et al., 2021). FGVR is a practically important capability for reliable multimodal interaction, as users often require precise object identification, and downstream reasoning implicitly assumes that the predicted class is correct (Geigle et al., 2024). Empirically, LVLM performance can collapse when shifting from coarse to fine granularity. For example, Kim & Ji (2024) report that LLaVA-1.5 achieves 98.43% accuracy at the superordinate level on iNaturalist, but this drops to 46.91% at the coarse-grained level and further to just 1.56% at the fine-grained level. Concretely, a model may correctly recognize an object as a cat, yet confuse visually similar breeds such as Birman and Siamese, producing fluent chain-of-thought explanations that rely on generic breed stereotypes (e.g., "pointed ears" or "slender body") rather than the subtle visual attributes actually present in the image (Fig. 1). Prior analyses further indicate that multimodal ICL (Fig. 2b) is driven primarily by textual regularities rather than visual evidence (Zong et al., 2024; Chen et al., 2023a), and that chain-of-thought reasoning (Fig. 2a) can drift and overthink, yielding explanations that are not consistently grounded

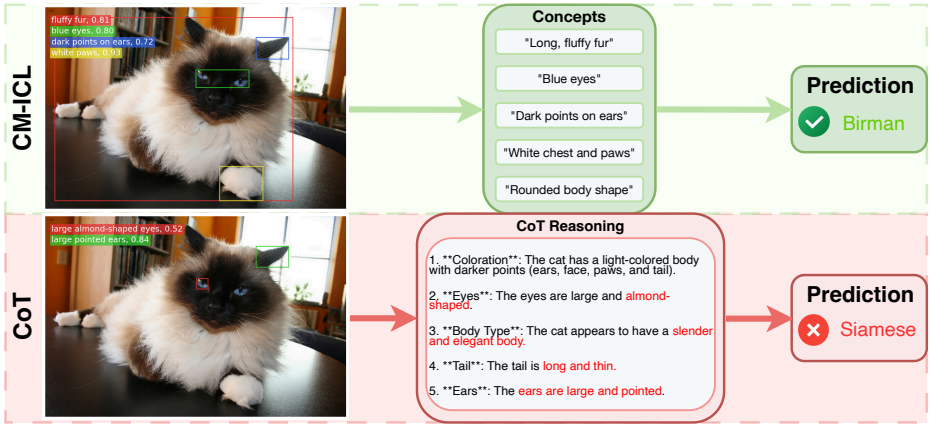


Figure 1: Illustrative comparison on fine-grained cat breed recognition. CM-ICL concepts (top) are more visually grounded and lead to the correct prediction *Birman*, whereas CoT-derived phrases (bottom) emphasize stereotypical cues and predict *Siamese*. Left: SAM 3 grounding; right: reasoning content.

in the visual input (Turpin et al., 2023; Barez et al., 2025; Arcuschin et al., 2025; Liu et al., 2024; Lee et al., 2025; Peng et al., 2025). Taken together, existing prompting methods lack an explicit mechanism to surface and consistently use the fine-grained visual cues required for FGVR.

To address this mismatch, we hypothesize that the missing ingredient is a controllable intermediate interface between image evidence and the final prediction. We propose **Concept-Mediated In-Context Learning (CM-ICL)** (Fig. 2c), a training-free prompting strategy that first elicits a compact set of image-derived attribute concepts from the query image and then re-injects them as structured context for classification. Rather than adding retrieved demonstrations or unconstrained long-form rationales, CM-ICL re-expresses fine-grained visual cues as concise language-space anchors that the frozen LVLM can condition on during the final decision, while preserving the original image input. To evaluate this intermediate representation without manual concept annotations, we combine promptable-segmentation-based grounding metrics with task-coupled diagnostics that connect visual localizability to prediction behavior. Experiments across six fine-grained datasets demonstrate that CM-ICL consistently improves accuracy over training-free baselines, produces more concise and visually localizable concepts, and substantially reduces generation failures. These results indicate that concept mediation offers an effective pathway for eliciting fine-grained sensitivity in current large vision-language models.

Our contributions are threefold: (i) we introduce a training-free prompting strategy that guides frozen LVLMs with image-derived attribute concepts for fine-grained recognition; (ii) we propose perceptual grounding metrics and task-coupled diagnostics for annotation-free concept evaluation; and (iii) we provide extensive empirical evidence of improved performance, stability, and interpretability across six FGVR datasets.

2 Related Work

2.1 Fine-Grained Visual Recognition

Fine-grained visual recognition (FGVR) focuses on distinguishing subordinate categories such as bird species, aircraft variants, dog breeds, or car models (Wei et al., 2021). Its central challenge is that correct predictions often depend on subtle, localized, and visually confusable attributes. Classical FGVR methods therefore explicitly model discriminative parts or attention regions (Shu et al., 2022). With the emergence of large vision–language models (LVLMs), a key question is whether general multimodal representations can support such fine-grained distinctions without task-specific training.

Recent studies show that LVLMs still struggle to capture the attribute-level distinctions required by FGVR (Yu et al., 2025a). Several methods improve fine-grained recognition by injecting additional se-

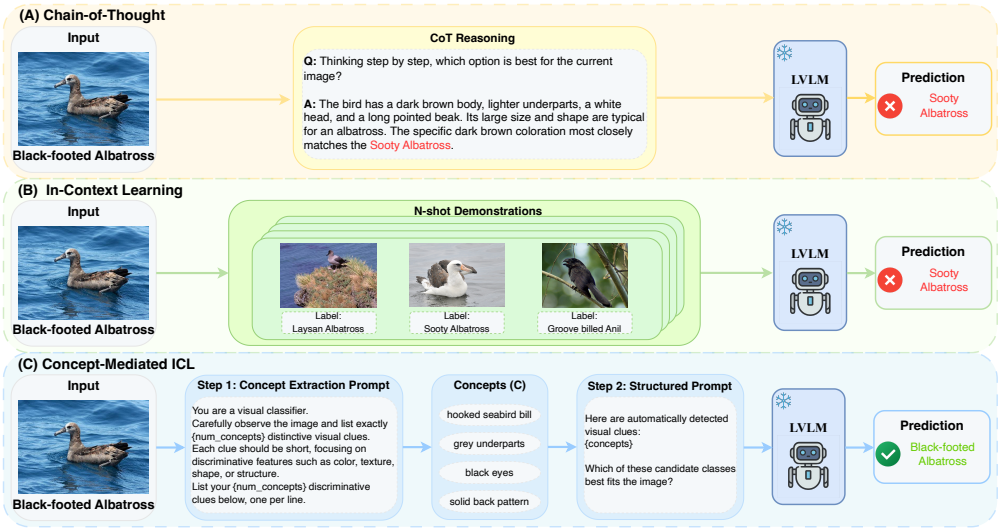


Figure 2: Comparison of inference strategies under a frozen LVLM. (a) CoT performs free-form reasoning, (b) ICL conditions on demonstrations, and (c) CM-ICL first extracts discriminative visual concepts before using them as structured context for final prediction.

mantic signals during training, such as attribute-enriched descriptions (He et al., 2025) or vocabulary-free categorization frameworks (Demidov et al., 2025). These approaches are effective but rely on specialized supervision, curated semantic resources, or auxiliary processing. In contrast, CM-ICL targets the strictly training-free setting: it does not assume a predefined attribute vocabulary or any parameter update, but dynamically extracts image-derived concepts and uses them as structured context for a frozen LVLM.

2.2 Multimodal In-Context Learning

In-context learning (ICL) enables language models to perform new tasks by conditioning on examples in the prompt, without parameter updates (Dong et al., 2024). Prior work suggests that ICL can resemble implicit inference-time adaptation (Dai et al., 2023; Zhou et al., 2024), and zero-shot variants further reduce the need for curated demonstrations by generating pseudo examples (Chen et al., 2023b; Lyu et al., 2023). However, multimodal ICL remains challenging. Recent benchmarks show that LVLMs often struggle to benefit from image–text demonstrations (Zong et al., 2024; Li et al., 2025), and mechanistic analyses suggest that adaptation can be dominated by textual patterns while image information contributes weakly (Chen et al., 2023a). This is especially problematic for FGVR, where success depends on subtle visual attributes rather than global semantic similarity or label priors. These findings suggest a utilization bottleneck: fine-grained visual cues may be encoded in the visual tokens, but not sufficiently emphasized by the language decoder during the final decision.

CM-ICL addresses this issue from a different angle than demonstration-based ICL. Instead of adding external image–label examples, it constructs a compact concept block from the query image itself. The resulting image-derived concepts serve as language-space anchors, allowing fine-grained visual cues to enter the textual context that LVLM decoders can more readily condition on. Thus, CM-ICL is an in-context strategy in the sense of contextual intervention, but it does not rely on retrieved or hand-crafted demonstrations.

2.3 Chain of Thought Prompting

Chain-of-thought (CoT) prompting encourages models to generate step-by-step explanations before answering (Wei et al., 2023). While useful for some reasoning tasks, recent work questions the faithfulness and stability of CoT rationales (Turpin et al., 2023; Barez et al., 2025; Arcuschin et al., 2025). Long reasoning chains may drift, overthink, or accumulate hallucinated details (Liu et al., 2024; Lee et al., 2025; Peng et al.,

2025). These issues are particularly harmful in FGVR, where predictions should depend on concrete local visual evidence rather than generic class stereotypes or speculative descriptions.

A related line of work improves multimodal reasoning by explicitly grounding intermediate steps in regions, points, or verified visual evidence (Wu et al., 2025; Man et al., 2025; Yi & Shang, 2025). Many of these methods require grounded supervision, post-training, reinforcement learning, or specialized modules. Additional discussion is provided in Appendix E. CM-ICL is complementary: it remains fully training-free and does not claim that generated concepts are faithful internal explanations. Instead, it replaces free-form rationales with short image-derived concepts that act as structured context for the final prediction, reducing uncontrolled reasoning drift while preserving an inspectable intermediate representation.

3 Concept-Mediated In-Context Learning

We now present Concept-Mediated In-Context Learning (CM-ICL), a training-free structured inference strategy that guides a frozen LVLM through image-derived concept context. Given a query image and its associated question, CM-ICL uses the image in two stages: the model first elicits a compact set of short, discriminative visual concepts, and then conditions the final prediction on both the original image and the generated concept block. We use the term in-context learning in the broad sense that the model is guided solely by information placed in its input context; unlike conventional demonstration-based ICL, CM-ICL does not retrieve or supply labeled examples, but constructs the additional context from the query image itself. The generated concepts are not intended as free-form rationales or faithful explanations of the model's internal decision process. Instead, they serve as semantic anchor tokens that re-express selected visual cues in the language-token space of the LVLM. When re-injected into the prediction prompt, these anchors provide a compact interface through which the final answer query can interact with fine-grained image evidence under the model's standard attention mechanism. Below, we first describe the CM-ICL prompt construction and then analyze how concept insertion affects the final answer attention computation under exact scaled dot-product softmax attention.

3.1 Method Overview

We consider LVLMs as conditional language models that process a unified token sequence combining visual and textual tokens. In the prediction stage, the concept block should not be viewed as replacing visual evidence with text. Rather, it provides additional textual anchor tokens that interact with the original visual and textual context through the model's attention mechanism. This design is motivated by a utilization bottleneck: fine-grained visual cues may be encoded in the visual tokens, yet remain weakly used by the language decoder because inference is often shaped by text-space representations and linguistic priors. CM-ICL therefore exposes image-derived cues in a form that the decoder can more readily condition on, without removing the original image from the input. We analyze concept mediation as an inference-time contextual intervention on the final answer attention computation.

Base prompt without concepts. Consider a vision-language model that processes a multimodal prompt consisting of a task instruction, an image, and a question. After multimodal encoding, the model represents the original prompt as a sequence of contextual hidden states $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^{l} \in \mathbb{R}^{d\times l}$, which serve as the keys and values for an attention head when the model predicts the answer. Here, $\boldsymbol{X}$ may contain both visual and textual context available to the final answer query. Let $\boldsymbol{h} \in \mathbb{R}^d$ denote the hidden state of the query token at step t. The corresponding query vector is computed as $\boldsymbol{q} = \boldsymbol{W}_Q\boldsymbol{h} \in \mathbb{R}^{d_k}$, where $\boldsymbol{W}_Q \in \mathbb{R}^{d_k\times d}$ is the query projection matrix. For this attention head, the standard scaled dot-product attention output can be written as

$$\mathcal{F}_{\text{base}}(\boldsymbol{q}) = \boldsymbol{W}_V\boldsymbol{X}\,\text{softmax}\left(\frac{(\boldsymbol{W}_K\boldsymbol{X})^\top\boldsymbol{q}}{\sqrt{d_k}}\right), \tag{1}$$

where $\sqrt{d_k}$ denotes the scaling factor, and $\boldsymbol{W}_K, \boldsymbol{W}_V$ are the key and value projection matrices, respectively. To make the effect of concept insertion explicit while keeping the softmax normalization intact, define the base-token key and value vectors as $\boldsymbol{k}_j^x = \boldsymbol{W}_K\boldsymbol{x}_j$ and $\boldsymbol{v}_j^x = \boldsymbol{W}_V\boldsymbol{x}_j$, respectively. For a fixed query $\boldsymbol{q}$, let

$s_j(\boldsymbol{q}) = \frac{\langle \boldsymbol{q}, \boldsymbol{k}_j^x \rangle}{\sqrt{d_k}}$. Then Eqn. (1) can be equivalently written as

$$\mathcal{F}_{\text{base}}(\boldsymbol{q}) = \sum_{j=1}^{l} \frac{\exp(s_j(\boldsymbol{q}))}{S_X(\boldsymbol{q})} \boldsymbol{v}_j^x, \tag{2}$$

where $S_X(\boldsymbol{q}) = \sum_{j=1}^{l} \exp(s_j(\boldsymbol{q}))$. This expression is an exact rewriting of the scaled dot-product softmax attention output and will serve as the reference representation before inserting concept tokens.

Concept extraction from the image. Given the same query image I, we first run the frozen LVLM in a concept elicitation mode. The model receives a prompt[1] that contains the image and a short instruction such as "Describe the m distinctive visual cues in the image". It then generates a set of concise concept phrases $\mathcal{C}(I) = \{c_1, \ldots, c_m\}$, where each c_i is a short textual description of a visual attribute (e.g., "hooked seabird bill", "grey underparts", "black eyes"). Here, c_i denotes a concept phrase generated from the image, rather than a class label or a free-form reasoning step. These phrases are designed to capture localized or attribute-level visual evidence in a compact form. In the prediction stage, they are reintroduced as textual context, allowing image-derived cues to be represented as language-space anchors that can interact with the original multimodal context through attention.

Augmented prompt with concept tokens. In the prediction stage, we use the same frozen LVLM and insert the previously generated concepts back into the prompt as a structured concept block for fine-grained classification. Concretely, the concepts extracted in the first stage are prepended to the original task input in a fixed format. The resulting prompt contains a dedicated concept block followed by the task instruction and query, with a clear textual separation to preserve semantic boundaries between concept descriptions and task content. The original image remains part of the input, so the model predicts from both the visual evidence and the image-derived concept context.

After tokenization and encoding, this augmented prompt is processed by the LVLM in the prediction stage. At the representation level, this means that the attention head now attends over an augmented token sequence

$$[\boldsymbol{C}, \boldsymbol{X}] = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_l] \in \mathbb{R}^{d \times (m+l)}, \tag{3}$$

where each $\boldsymbol{c}_i \in \mathbb{R}^d$ is the hidden representation of the concept token c_i produced by the language model encoder. For notational simplicity, we represent each concept phrase by one concept token; if a phrase is split into multiple subword tokens, the same analysis applies by treating those subword hidden states as additional concept-token representations. The concept tokens do not introduce labeled demonstrations, external supervision, or parameter updates. They serve as language-space anchors that re-express image-derived visual cues within the model's textual context. As a result, concept mediation changes the final prediction context while preserving the frozen model and the original visual input.

Exact softmax displacement. For each concept token, define its key and value vectors as $\boldsymbol{k}_i^c = \boldsymbol{W}_K \boldsymbol{c}_i$ and $\boldsymbol{v}_i^c = \boldsymbol{W}_V \boldsymbol{c}_i$, respectively. The corresponding concept-token score is $r_i^c(\boldsymbol{q}) = \frac{\langle \boldsymbol{q}, \boldsymbol{k}_i^c \rangle}{\sqrt{d_k}}$. Under the exact scaled dot-product softmax attention, the attention output after inserting concept tokens is

$$\mathcal{F}_{\text{cm}}(\boldsymbol{q}; \boldsymbol{C}) = \sum_{j=1}^{l} \frac{\exp(s_j(\boldsymbol{q}))}{S_X(\boldsymbol{q}) + S_C(\boldsymbol{q})} \boldsymbol{v}_j^x + \sum_{i=1}^{m} \frac{\exp(r_i^c(\boldsymbol{q}))}{S_X(\boldsymbol{q}) + S_C(\boldsymbol{q})} \boldsymbol{v}_i^c, \tag{4}$$

where $S_C(\boldsymbol{q}) = \sum_{i=1}^{m} \exp(r_i^c(\boldsymbol{q}))$. Then the difference between the concept-mediated attention output and the base attention output admits the following exact decomposition:

$$\mathcal{F}_{\text{cm}}(\boldsymbol{q}; \boldsymbol{C}) - \mathcal{F}_{\text{base}}(\boldsymbol{q}) = \sum_{i=1}^{m} \alpha_i^c(\boldsymbol{q}; \boldsymbol{C}) \left(\boldsymbol{v}_i^c - \mathcal{F}_{\text{base}}(\boldsymbol{q}) \right), \tag{5}$$

where $\alpha_i^c(\boldsymbol{q}; \boldsymbol{C}) = \frac{\exp(r_i^c(\boldsymbol{q}))}{S_X(\boldsymbol{q}) + S_C(\boldsymbol{q})}$ is the softmax attention mass assigned to the i-th concept token after concept insertion. This identity follows directly from softmax normalization and does not replace the attention mechanism with an unnormalized linear-attention surrogate.

[1]Full prompts and implementation details are provided in Appendix B.

Interpretation. The exact softmax decomposition shows that concept mediation acts through attention-mass reallocation rather than a global additive update to the attention matrix. For a fixed answer query, inserted concepts shift the attention output from the base aggregate representation toward residual concept-value directions. CM-ICL therefore does not add new visual evidence; instead, it re-expresses image-derived cues as compact language-space anchors that can compete in the final decision context. This is useful when fine-grained visual cues are encoded but weakly used due to cross-modal alignment gaps, softmax competition, or linguistic priors in the decoder. A concept contributes to the decision only when it receives non-negligible attention and its residual direction aligns with the correct-class decision direction. We formalize the resulting margin improvement and robustness properties next.

3.2 Why Concept Mediation Works

We now connect the exact softmax displacement derived above to the model's final decision. Our goal is not to provide a training or generalization analysis, since CM-ICL does not update model parameters. Instead, we analyze concept mediation as a local inference-time intervention on the final answer attention computation. The analysis fixes the answer-query state $\boldsymbol{q}$ and keeps the scaled dot-product softmax attention introduced in Sec. 3.1 intact. This allows us to characterize two decision-level effects: how concept tokens can shift the pairwise class margin, and why the prediction can remain stable when the generated concepts are imperfect. All proofs are provided in Appendix A.

To connect the exact softmax displacement to the final prediction, suppose the class logit is locally read out from the attention output as

$$\ell_y(\boldsymbol{C}) = \boldsymbol{w}_y^\top \mathcal{F}_{\mathrm{cm}}(\boldsymbol{q}; \boldsymbol{C}) + b_y, \qquad \ell_y(\varnothing) = \boldsymbol{w}_y^\top \mathcal{F}_{\mathrm{base}}(\boldsymbol{q}) + b_y, \tag{6}$$

where $\ell_y(\boldsymbol{C})$ denotes the logit after inserting the concept block $\boldsymbol{C}$, while $\ell_y(\varnothing)$ denotes the base logit without concept tokens. This local readout isolates how the final attention output affects the relative preference among candidate classes. For the reference class y^* and any competing class $y \in \mathcal{Y} \setminus \{y^*\}$, define the pairwise margin

$$M_y(\boldsymbol{C}) = \ell_{y^*}(\boldsymbol{C}) - \ell_y(\boldsymbol{C}), \qquad M_y(\varnothing) = \ell_{y^*}(\varnothing) - \ell_y(\varnothing). \tag{7}$$

Using the exact softmax displacement in Eqn. (5), the concept-induced margin shift relative to the base prompt is

$$\Delta M_y(\boldsymbol{C}) = M_y(\boldsymbol{C}) - M_y(\varnothing) = \sum_{i=1}^{m} \alpha_i^c(\boldsymbol{q}; \boldsymbol{C}) \left\langle \boldsymbol{u}_y, \boldsymbol{v}_i^c - \mathcal{F}_{\mathrm{base}}(\boldsymbol{q}) \right\rangle, \tag{8}$$

where $\boldsymbol{u}_y = \boldsymbol{w}_{y^*} - \boldsymbol{w}_y$. Eqn. (8) gives a decision-level decomposition of concept mediation. Each concept token affects the pairwise margin through two factors: its softmax attention mass and the alignment between its residual value direction and the logit direction favoring the reference class over the competing class. Thus, writing down an image-derived concept is useful only when the concept both receives attention in the final decision context and moves the attention output in a direction that increases the correct-class margin.

3.2.1 How Concepts Aid Inference

Eqn. (8) gives an exact additive decomposition of the margin shift over concept tokens under softmax attention. For each concept token, define its softmax-weighted contribution to the margin against class y as

$$T_{i,y}^{\mathrm{sm}} = \alpha_i^c(\boldsymbol{q}; \boldsymbol{C}) \left\langle \boldsymbol{u}_y, \boldsymbol{v}_i^c - \mathcal{F}_{\mathrm{base}}(\boldsymbol{q}) \right\rangle. \tag{9}$$

This contribution is positive when the concept token is assigned non-negligible attention mass and its residual value direction moves the representation toward the reference class relative to the competing class. Conversely, an incorrect or weakly relevant concept may contribute negatively. Its influence, however, is attenuated when the concept receives little attention or when its residual value direction is weakly aligned with the decision boundary. Thus, concept mediation does not require every generated concept to be correct or individually decisive; it only requires the aggregate softmax-weighted contribution of useful concepts to dominate the bounded contribution of noisy ones. Let $Z_i = \mathbb{1}(c_i \text{ is correct})$ indicate whether the predicted concept matches a correct visual attribute, and let $\mathbb{E}[Z_i] = a$ denote the expected concept accuracy.

Assumption 1 (Softmax-weighted concept contributions)**.** For every competing class $y \in \mathcal{Y} \setminus \{y^*\}$, there exist constants $\mu_y^+ > 0$ and $\mu_y^- \geq 0$ such that, for every concept token $i \in \{1, \ldots, m\}$,

$$\mathbb{E}\big[T_{i,y}^{\mathrm{sm}} \mid Z_i = 1\big] \geq \mu_y^+, \qquad \mathbb{E}\big[T_{i,y}^{\mathrm{sm}} \mid Z_i = 0\big] \geq -\mu_y^-. \tag{10}$$

Assumption 1 is a local relevance condition on the final answer query. It states that visually correct concepts have positive expected softmax-weighted residual alignment with the reference class, while incorrect concepts may hurt the margin only with bounded expected magnitude. Importantly, this assumption does not require every correct concept to be useful in every example. The condition is distributional and operates after accounting for both attention mass and residual-value alignment.

Theorem 1 (Concept-aided expected margin improvement)**.** *Under Assumption 1, if*

$$a > \max_{y \neq y^*} \frac{\mu_y^-}{\mu_y^+ + \mu_y^-}, \tag{11}$$

then $\mathbb{E}[\Delta M_y(\boldsymbol{C})] > 0$ *holds for every competing class* $y \in \mathcal{Y} \setminus \{y^*\}$.

Theorem 1 shows that concept mediation can improve the expected decision margin even when some generated concepts are imperfect. The key quantity is not raw concept correctness alone, but the softmax-weighted contribution in Eqn. (9). A concept helps only when it both receives attention in the final decision context and moves the attention output in a direction that favors the reference class over competing classes. When the positive contributions of correct concepts dominate the bounded negative contributions of incorrect concepts, the aggregate margin shift is positive in expectation.

This result yields an empirical prediction: concept mediation should induce positive decision shifts, especially when generated concepts are visually accurate and influential in the final decision context. Since our evaluation parses final answers from generative LVLM outputs, we do not rely on token-level logits or candidate margins that may vary across backbones, tokenizers, and answer formats; Sec. 4 instead evaluates the corresponding output-level behavior through task-coupled diagnostics.

3.2.2 Tolerance to Imperfect or Noisy Concepts

We next analyze why the prediction can remain stable when the generated concept set is imperfect or partially perturbed. Let $\boldsymbol{C}$ be the original concept block and let $\boldsymbol{C}' = \boldsymbol{C} + \boldsymbol{E}$ be an alternative concept block obtained by replacing a subset of concepts. For any competing class $y \neq y^*$, the exact softmax-induced margin $M_y(\boldsymbol{C})$ is treated as a function of the injected concept block. Our goal is to bound how much this margin can change when $\boldsymbol{C}$ is replaced by $\boldsymbol{C}'$ under the same exact softmax attention mechanism introduced above. This analysis captures the effect of imperfect concept generation at the representation level. A noisy or hallucinated concept can affect the final decision only through its induced change in the concept-token representations and the resulting change in the exact softmax attention output. Therefore, the relevant question is whether this perturbation is large enough to overcome the margin buffer of the original concept-mediated prediction.

Assumption 2 (Sparse and bounded concept perturbations)**.** Let $\boldsymbol{E} = [\boldsymbol{e}_i]_{i=1}^m$. The replacement from a fixed concept block $\boldsymbol{C}$ to the alternative block $\boldsymbol{C}' = \boldsymbol{C} + \boldsymbol{E}$ satisfies:

- (A1) **Bounded token and concept representations.** There exists a finite constant $R > 0$ such that, for all base tokens $\boldsymbol{x}_j$ and all concept indices i, $\|\boldsymbol{x}_j\|_2 \leq R$, $\|\boldsymbol{c}_i\|_2 \leq R$, and $\|\boldsymbol{c}_i'\|_2 \leq R$.
- (A2) **Sparse concept replacement.** At most m_e concept tokens are modified, i.e., $\boldsymbol{E}$ has at most m_e nonzero columns.
- (A3) **Bounded per-concept replacement.** For every modified concept token, $\|\boldsymbol{e}_i\|_2 \leq \varepsilon_c$, where $\|\cdot\|_2$ denotes the ℓ_2 norm.
- (A4) **Uniformly bounded margin sensitivity.** The exact-softmax margin sensitivity is uniformly bounded across candidate classes: there exists a finite constant $B > 0$ such that, for all $y \in \mathcal{Y}$,

$$\|\boldsymbol{u}_y\|_2 \left(\|\boldsymbol{W}_V\|_2 + 2R \frac{\|\boldsymbol{W}_K^\top \boldsymbol{q}\|_2}{\sqrt{d_k}} \|\boldsymbol{W}_V\|_2 \right) \leq B.$$

Assumption 2 formalizes a moderate concept-noise regime. Only a limited number of concept representations are replaced, each replacement has bounded magnitude, and the final margin is not arbitrarily sensitive to small changes in the exact softmax attention output. The assumption does not require every concept to be correct. Instead, it requires that concept errors remain sparse and bounded in the representation space used by the final attention computation.

Theorem 2 (Prediction invariance under exact softmax attention)**.** *Suppose the prediction obtained with concept block $\boldsymbol{C}$ is $\hat{y}(\boldsymbol{C}) = y^*$. Then for any alternative concept block $\boldsymbol{C}' = \boldsymbol{C} + \boldsymbol{E}$ satisfying Assumption 2, if*

$$\min_{y \neq y^*} M_y(\boldsymbol{C}) > B\sqrt{m_e}\varepsilon_c,$$

the pairwise margins under $\boldsymbol{C}'$ remain positive for all $y \in \mathcal{Y} \backslash \{y^*\}$, i.e., $M_y(\boldsymbol{C}') > 0$, and hence the prediction is invariant:

$$\hat{y}(\boldsymbol{C}') = \hat{y}(\boldsymbol{C}) = y^*.$$

Implication. Theorem 2 gives a margin-buffer view of robustness under exact softmax attention. As long as the original concept-mediated prediction has a sufficient margin buffer and the replacement from $\boldsymbol{C}$ to $\boldsymbol{C}'$ is sparse and bounded, the prediction remains unchanged. Thus, the final decision depends on the aggregate margin effect of the concept block rather than the correctness of every individual concept. This result also clarifies the effect of hallucinated concepts. A strong but incorrect concept can change the prediction only if its representation-level perturbation is large enough, or if enough concepts are replaced, to overcome the original margin buffer. Conversely, moderate concept errors may change some intermediate contributions without flipping the final decision. This provides a testable prediction: replacing more concepts or inserting stronger incorrect concepts should increase the prediction flip rate, while examples with larger concept-mediated margins should remain more stable.

The analysis above characterizes how concept tokens can affect the final decision through softmax attention and margin shifts. We next evaluate a complementary property of the intermediate concept block: whether the generated concepts are visually grounded in the input image. To this end, Section 4 introduces annotation-free grounding metrics based on a promptable segmentation model, which measure the localizability, reliability, and spatial expressiveness of concept prompts.

4 Concept Grounding and Task-Coupled Diagnostics

We evaluate the generated concept block from two complementary perspectives. First, we assess whether concepts are visually localizable using a promptable segmentation model. Second, we introduce task-coupled diagnostics that relate grounding quality to classification behavior. The first group of metrics characterizes the perceptual grounding of concept prompts, while the second group tests whether such grounding is associated with downstream recognition outcomes.

4.1 Perceptual Grounding Metrics

To quantify the visual localizability of generated concepts, we use a pre-trained promptable concept segmentation model, SAM 3 (Carion et al., 2025). SAM 3 takes a concept prompt as a textual query and predicts instance-level segmentation regions in the image that correspond to the queried concept, along with a grounding confidence score for each detected instance. These metrics evaluate whether generated concept prompts correspond to localizable image regions. They are not intended to prove that the LVLM internally attends to those regions, nor do they by themselves establish that a concept is task-discriminative.

Let the image dataset be denoted as $\{I_i\}_{i=1}^{n}$. For each image I_i, its associated concept prompts are denoted as $\boldsymbol{p}_i = \{p_{i,j}\}_{j=1}^{m_i}$, where m_i is the number of prompts (concepts) for I_i. For each prompt $p_{i,j}$, which is a short noun phrase describing a visual concept (e.g., "hooked-seabird bill", "wing-mounted engines"), the model returns the number of regions in image I_i that match concept $p_{i,j}$ denoted as $g_{i,j}$. When $g_{i,j} \geq 1$, the concept $p_{i,j}$ is considered visually groundable on I_i.

Based on these outputs, we define three complementary grounding metrics. Rather than collapsing grounding performance into a single score, these metrics are designed to capture different and diagnostically meaningful

aspects of perceptual concept quality, including image-level coverage, dataset-level reliability, and spatial expressiveness.

Mean Valid Ratio (Mean-VR). This metric measures the fraction of generated concepts that can be successfully grounded at the image level. For each image I_i, the valid ratio is defined as

$$\mathrm{VR}(I_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbb{1}(g_{i,j} \geq 1), \tag{12}$$

where m_i denotes the number of concept prompts for image I_i, and $\mathbb{1}(\cdot)$ is the indicator function. We then report the dataset-level average as

$$\text{Mean-VR} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{VR}(I_i). \tag{13}$$

Mean-VR captures per-image grounding coverage and is sensitive to hallucinated or weakly grounded concepts that lack visual support.

Prompt Success Rate (PSR). Prompt Success Rate measures the dataset-level grounding reliability of concept prompts. It is defined as

$$\mathrm{PSR} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathbb{1}(g_{i,j} \geq 1)}{\sum_{i=1}^{n} m_i}. \tag{14}$$

While Mean-VR focuses on per-image coverage, PSR reflects the overall stability of the concept generation process across images. Equivalently, Mean-VR is a macro average over images, whereas PSR is a micro average over all concept prompts. When each image has the same number of concept prompts, the two metrics become numerically close; we report both to support comparison with CoT-derived concepts, whose number of extracted phrases may vary across images.

Average Instances per Valid Prompt (AIC). To characterize the spatial expressiveness of each grounded concept, we further compute the average number of detected instances per valid prompt:

$$\mathrm{AIC} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} g_{i,j} \cdot \mathbb{1}(g_{i,j} \geq 1)}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} \mathbb{1}(g_{i,j} \geq 1)}. \tag{15}$$

AIC reflects whether a grounded concept typically corresponds to a single localized region or multiple coherent spatial instances, indicating the spatial extent of the concept once grounding succeeds rather than its mere existence. It should be interpreted as a descriptive grounding statistic rather than a direct measure of task relevance.

Collectively, Mean-VR, PSR, and AIC provide a compact evaluation protocol for the perceptual grounding of generated concepts. Mean-VR diagnoses image-level concept coverage, PSR captures dataset-level grounding reliability, and AIC measures spatial expressiveness. This decomposition enables fine-grained analysis of concept grounding behavior without requiring manual annotations and is applicable to any method that produces visual concepts. However, visual localizability alone does not imply that a concept is useful for classification. We therefore introduce task-coupled diagnostics below.

4.2 Task-Coupled Grounding Diagnostics

The perceptual metrics above measure visual localizability, but localizability alone does not imply task usefulness. Since ground-truth concept annotations are unavailable for most datasets, we treat grounding as an annotation-free proxy for visual support rather than direct concept accuracy. Following the margin view in Sec. 3.2, useful concepts should lead to favorable output changes only when they receive attention and align with the correct class direction. Because our LVLMs directly generate final answers and candidate-level logits are not uniformly comparable, we use output-level diagnostics based on per-image correctness. These diagnostics require no additional LVLM inference.

Let $\gamma_i^A = \mathbb{1}(\hat{y}_i^A = y_i)$ indicate whether method A correctly predicts image I_i. For concept-producing methods, $\mathrm{VR}^A(I_i)$ is computed from their concept prompts using Eqn. (12); for baselines without concepts, we use only their correctness indicators.

Fix-Harm Decomposition (FHD). As the primary output-level diagnostic, we decompose correctness changes into beneficial and harmful transitions:

$$\mathrm{Fix}^{\mathrm{CM}|B} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\gamma_i^{\mathrm{CM}} = 1,\ \gamma_i^{B} = 0), \quad \mathrm{Harm}^{\mathrm{CM}|B} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\gamma_i^{\mathrm{CM}} = 0,\ \gamma_i^{B} = 1),$$

$$\mathrm{Net}^{\mathrm{CM}|B} = \mathrm{Fix}^{\mathrm{CM}|B} - \mathrm{Harm}^{\mathrm{CM}|B}.$$

Net equals the accuracy gain of CM-ICL over baseline B, while Fix and Harm reveal whether the gain comes from correcting mistakes or introducing new errors. A low Harm rate is consistent with the prediction-invariance view in Theorem 2.

Localizability-only association checks. As supplementary diagnostics, we compute two Spearman associations to test whether image-level grounding coverage alone tracks prediction behavior. Grounding-Correctness Association (GCA) relates grounding coverage to prediction correctness:

$$\rho^{A}_{\mathrm{GCA}} = \mathrm{Spearman}\left(\{\mathrm{VR}^{A}(I_i)\}_{i=1}^{n}, \{\gamma_i^{A}\}_{i=1}^{n}\right).$$

Grounding-Gain Association (GGA) relates CM-ICL grounding coverage to the correctness change over a baseline B:

$$\Delta\gamma_i^{\mathrm{CM}|B} = \gamma_i^{\mathrm{CM}} - \gamma_i^{B}, \qquad \rho^{\mathrm{CM}|B}_{\mathrm{GGA}} = \mathrm{Spearman}\left(\{\mathrm{VR}^{\mathrm{CM}}(I_i)\}_{i=1}^{n}, \{\Delta\gamma_i^{\mathrm{CM}|B}\}_{i=1}^{n}\right).$$

These associations are not used as causal evidence or for concept selection; they only evaluate whether the localizability proxy by itself is predictive of downstream behavior.

These diagnostics are used only for post-hoc evaluation and do not use ground-truth labels for concept selection at inference time. Together, they assess favorable output changes, limited harmful side effects, and the extent to which visual localizability is associated with prediction behavior.

5 Experiments

5.1 Implementation Details

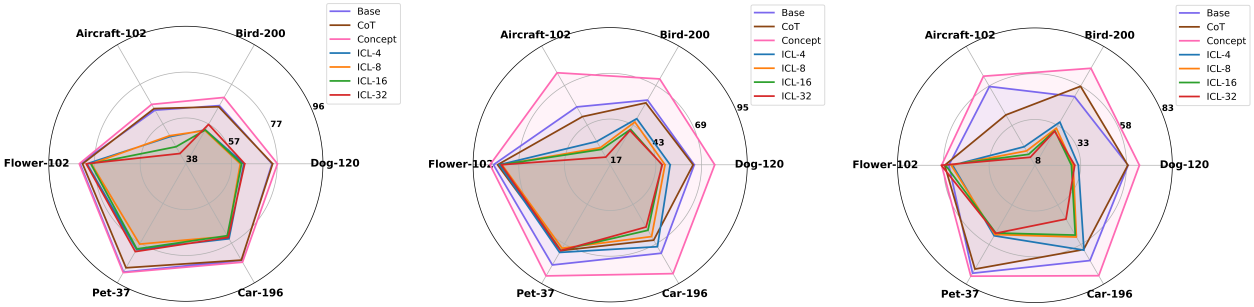
Datasets. We evaluate our method on six widely used fine-grained visual recognition datasets. The FOCI (Fine-grained Object Classification) benchmark (Geigle et al., 2024) formulates fine-grained object classification as multiple-choice problems to avoid ambiguous answers and includes FGVC-Aircraft (Maji et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Oxford-IIIT Pet (Parkhi et al., 2012), and Stanford Cars (Krause et al., 2013). Finedefics (He et al., 2025) further extends the benchmark by incorporating Birds-200 (Welinder et al., 2010) and Stanford Dogs-120 (Khosla et al., 2011). All datasets are publicly available research benchmarks and are used solely for non-commercial academic evaluation on the official test splits, without redistributing the original data.

Baselines. We intentionally restrict our comparisons to training-free reasoning methods that operate purely at inference time and do not rely on additional supervision or task-specific tuning. These include direct zero-shot classification, the traditional chain-of-thought (CoT) (Wei et al., 2023) prompting strategy (e.g., "Let's think step by step"), and in-context learning (ICL) (Zong et al., 2024; Li et al., 2025) with 4/8/16/32 image-label demonstration pairs as prompts. We also include a training-free prompt optimization baseline (Xiao et al., 2024; Yuksekgonul et al., 2024) in Appendix C Fig. 4.

LVLM Backbones. We select three open-source LVLMs with comparable model sizes as backbones, including Qwen2.5-VL-7B (Bai et al., 2025b), Qwen3-VL-8B (Bai et al., 2025a), and InternVL3-8B (Zhu et al., 2025). Since each dataset contains several thousand test instances and our primary goal is to analyze the general effectiveness of the proposed training-free methods, we do not include closed-source LVLMs in our evaluation. Evaluating such models would incur substantial computational and financial costs, which would hinder large-scale and systematic analysis. All experiments are conducted on NVIDIA workstation GPUs with 48 GB of memory per card.

Model	Dog-120	Bird-200	Aircraft-102	Flower-102	Pet-37	Car-196	Avg
Qwen2.5-VL	74.57	66.47	64.24	82.89	90.87	85.14	77.36
Qwen2.5-VL-CoT	74.84	65.91	65.02	81.83	88.78	85.03	76.9
Qwen2.5-VL-CM-ICL	**76.91**	**70.43**	**67.06**	**83.22**	**91.11**	**86.05**	**79.13**
Qwen3-VL	64.60	59.23	54.76	83.10	82.47	74.80	69.83
Qwen3-VL-CoT	63.93	57.40	48.39	80.77	73.07	66.24	64.97
Qwen3-VL-CM-ICL	**76.18**	**73.08**	**77.11**	**85.05**	**89.64**	**88.14**	**81.53**
InternVL	58.87	51.33	57.67	55.02	75.96	68.04	61.15
InternVL-CoT	58.73	57.80	39.87	57.26	73.37	61.15	58.03
InternVL-CM-ICL	**64.99**	**69.07**	**64.12**	**58.94**	**77.81**	**77.52**	**68.74**

Table 1: Performance comparison of LVLMs across six fine-grained datasets. Best scores per column are highlighted in bold. Light-blue rows denote our concept-mediated prompting methods. All results are top-1 accuracy (%).



(a) Qwen2.5-VL (b) Qwen3-VL (c) InternVL

Figure 3: Radar plot comparison across six fine-grained datasets for Zero-shot (Base), Chain-of-Thought (CoT), **CM-ICL (Concept)**, and In-Context Learning (ICL-4/8/16/32 shots) under three models: Qwen2.5-VL, Qwen3-VL, and InternVL. Higher values indicate better classification accuracy.

5.2 Main Results

Overall Comparison. Table 1 compares CM-ICL with CoT, while Fig. 3 summarizes performance across all training-free baselines. Overall, concept-mediated prompting consistently improves over zero-shot, conventional CoT and ICL 4/8/16/32 shot prompting on most datasets and backbones, demonstrating the effectiveness of introducing explicit visual concepts as an intermediate representation for training-free fine-grained recognition. Additional results with standard deviations are reported in Appendix C Table 3.

Comparison with In-Context Learning. Fig. 3 further compares zero-shot, CoT, CM-ICL, and ICL with different numbers of demonstrations across three LVLM backbones. In our ICL setting, demonstrations are constructed by randomly sampling n image–label pairs from the training set (excluding the query image), without any retrieval or semantic matching, thereby reflecting a strictly training-free and retrieval-free evaluation protocol. Additional kNN-ICL ablation results are reported in Appendix C Table 5. A clear and consistent trend across all three backbones is that ICL fails to outperform the zero-shot baseline on fine-grained classification and often leads to noticeable performance degradation, particularly on visually subtle datasets such as Bird-200 and Aircraft-102. This behavior indicates that random ICL demonstrations introduce substantial semantic noise and visual mismatch, which distract the model from the subtle, class-discriminative cues required for fine-grained recognition. Unlike coarse-grained tasks where exemplar-based reasoning may be beneficial, fine-grained recognition is highly sensitive to part-level and attribute-level visual cues, making it especially vulnerable to irrelevant contextual interference from unrelated demonstrations. In contrast, concept-mediated prompting consistently forms the outer envelope of the radar plots across all datasets and backbones, demonstrating both superior performance and stronger stability. Furthermore, the degradation becomes more pronounced as the number of random shots increases (ICL-16/32), confirming

Dataset	Mean-VR ↑	PSR ↑	AIC ↑	Fix ↑	Harm ↓	Net ↑
Dog-120	0.384 / **0.457**	0.377 / **0.451**	1.586 / **1.604**	0.086	0.065	0.021
Bird-200	0.456 / **0.462**	0.453 / **0.462**	1.152 / **1.161**	0.118	0.073	0.045
Aircraft-102	0.490 / **0.521**	0.495 / **0.527**	1.513 / **1.722**	0.119	0.098	0.020
Flower-102	0.585 / **0.613**	0.576 / **0.608**	6.093 / **6.631**	0.059	0.045	0.014
Pet-37	0.377 / **0.500**	0.372 / **0.493**	1.565 / **1.603**	0.046	0.023	0.023
Car-196	**0.696** / 0.633	**0.684** / 0.633	1.889 / **2.139**	0.056	0.046	0.010
Avg.	0.498 / **0.531**	0.493 / **0.529**	2.300 / **2.477**	0.081	0.058	0.022

Table 2: Perceptual grounding metrics and task-coupled diagnostics comparing CoT-derived prompts and CM-ICL concepts. For Mean-VR, PSR, and AIC, each cell shows *CoT / CM-ICL*; the better value is highlighted in bold. Fix, Harm, and Net report the fractions of examples corrected, harmed, and net improved by CM-ICL relative to CoT.

that longer random demonstration contexts exacerbate contextual distraction rather than providing useful inductive bias in fine-grained settings.

5.3 Additional Analysis

Backbone Sensitivity and Generation Failure. Concept prompting is closely related to the backbone's reasoning and grounding capacity. Stronger LVLMs such as Qwen2.5-VL and Qwen3-VL benefit more consistently, while weaker models show higher variability across datasets. This indicates that concept-based prompting is broadly effective, but ultimately constrained by the backbone's intrinsic capacity. Additional results on more LVLM backbones and generation failure rate are reported in Appendix C Table 4 and Fig. 5.

Grounding quality, task-coupled diagnostics, and cost. For comparison with CoT rationales, we use the OpenAI GPT API (Achiam et al., 2023) to extract short visual concept phrases from CoT reasoning outputs, yielding about 3–5 phrases per image on average. Table 2 compares CoT-derived prompts and CM-ICL concepts on Qwen2.5-VL. CM-ICL concepts show stronger perceptual grounding than CoT-derived phrases: they achieve higher Mean-VR and PSR on five out of six datasets and higher AIC on all six datasets, indicating better grounding coverage, prompt-level reliability, and spatial expressiveness. The task-coupled results further show favorable output behavior: CM-ICL consistently fixes more CoT errors than it harms, yielding positive Net improvement on every dataset. Additional localizability-only association checks are reported in Appendix C Table 6. Table 7 further reports latency and token profiling on Qwen2.5-VL: CM-ICL improves average accuracy over CoT from 76.90% to 79.13% (+2.23% absolute accuracy gain) while reducing mean latency from 6.95s to 2.08s per example, and uses far fewer tokens than ICL-32 on average (407.3 vs. 1949.5). Together, these results suggest that CM-ICL provides a more structured, visually grounded, and cost-effective interface for fine-grained recognition. Fig. 1 provides a qualitative example where CM-ICL produces concise, localized concepts, whereas CoT relies on more generic or stereotypical cues. Additional case studies are provided in Appendix D.

6 Conclusion

We presented CM-ICL, a training-free prompting strategy that guides LVLMs toward fine-grained visual cues using image-derived concepts. Our experiments show that concept mediation improves accuracy, stability, and grounding quality across diverse LVLMs, highlighting the value of intermediate representations in fine-grained recognition. Future work includes extending concept mediation to open-vocabulary settings, refining concept extraction, and exploring its interaction with broader multimodal reasoning frameworks.

References

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.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful. *arXiv preprint arXiv:2503.08679*, 2025.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. Chain-of-thought is not explainability. *Preprint, alphaXiv*, pp. v1, 2025.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: a survey. *arXiv preprint arXiv:2402.12451*, 2024.

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.

Shuo Chen, Zhen Han, Bailan He, Mark Buckley, Philip Torr, Volker Tresp, and Jindong Gu. Understanding and improving in-context learning on vision-language models. *arXiv preprint arXiv:2311.18021*, 2023a.

Wei-Lin Chen, Cheng-Kuang Wu, Yun-Nung Chen, and Hsin-Hsi Chen. Self-icl: Zero-shot in-context learning with self-generated demonstrations. *arXiv preprint arXiv:2305.15035*, 2023b.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4005–4019, 2023.

Dmitry Demidov, Muhammad Zaigham Zaheer, Omkar Thawakar, Salman Khan, and Fahad Shahbaz Khan. Vocabulary-free fine-grained visual recognition via enriched contextually grounded vision-language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4216–4225, 2025.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/.

Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. *arXiv preprint arXiv:2406.14496*, 2024.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. *arXiv preprint arXiv:2501.15140*, 2025.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

Jeonghwan Kim and Heng Ji. Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. *arXiv preprint arXiv:2402.16315*, 2024.

Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.

Dohyeon Lee, Yeonseok Jeong, and Seung-won Hwang. From token to action: State machine reasoning to mitigate overthinking in information retrieval. *arXiv preprint arXiv:2505.23059*, 2025.

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. In *Second Conference on Language Modeling*, 2025.

Chia Xin Liang, Pu Tian, Caitlyn Heqi Yin, Yao Yua, Wei An-Hou, Li Ming, Tianyang Wang, Ziqian Bi, and Ming Liu. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd workshop on knowledge extraction and integration for deep learning architectures*, pp. 100–114, 2022.

Ryan Liu, Jiayi Geng, Addison J Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. *arXiv preprint arXiv:2410.21333*, 2024.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. Z-icl: Zero-shot in-context learning with pseudo-demonstrations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2304–2317, 2023.

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

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-centric reasoning with grounded chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14268–14280, 2025.

Minheng Ni, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Kevin Lin, Wangmeng Zuo, and Lijuan Wang. Point-rft: Improving multimodal reasoning with visually grounded reinforcement finetuning. *arXiv preprint arXiv:2505.19702*, 2025.

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.

Keqin Peng, Liang Ding, Yuanxin Ouyang, Meng Fang, and Dacheng Tao. Revisiting overthinking in long chain-of-thought from the perspective of self-doubt. *arXiv preprint arXiv:2505.23480*, 2025.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7441–7455, 2024.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025.

Yangyang Shu, Baosheng Yu, Haiming Xu, and Lingqiao Liu. Improving fine-grained visual recognition in low data regimes via self-boosting attention mechanism. In *European Conference on Computer Vision*, pp. 449–465. Springer, 2022.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners, 2022. URL `https://arxiv.org/abs/2209.01975`.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and Jun Xiao. Vgr: Visual grounded reasoning, 2025. URL `https://arxiv.org/abs/2506.11991`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL `https://arxiv.org/abs/2201.11903`.

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.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2503.12799*, 2025.

Tim Z Xiao, Robert Bamler, Bernhard Schölkopf, and Weiyang Liu. Verbalized machine learning: Revisiting machine learning with language models. *arXiv preprint arXiv:2406.04344*, 2024.

Shixin Yi and Lin Shang. Corgi: Verified chain-of-thought reasoning with post-hoc visual grounding. *arXiv preprint arXiv:2508.00378*, 2025.

Hong-Tao Yu, Xiu-Shen Wei, Yuxin Peng, and Serge Belongie. Benchmarking large vision-language models on fine-grained image tasks: A comprehensive evaluation. *arXiv preprint arXiv:2504.14988*, 2025a.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025b.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic" differentiation" via text. *arXiv preprint arXiv:2406.07496*, 2024.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1662, 2025.

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

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

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

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

Appendix

A Proof

Theorem 1 (Concept-aided expected margin improvement)**.** *Under Assumption 1, if*

$$a > \max_{y \neq y^*} \frac{\mu_y^-}{\mu_y^+ + \mu_y^-}, \tag{11}$$

then $\mathbb{E}[\Delta M_y(\boldsymbol{C})] > 0$ *holds for every competing class* $y \in \mathcal{Y} \setminus \{y^*\}$.

Proof. Fix any competing class $y \in \mathcal{Y} \setminus \{y^*\}$. By the exact margin decomposition in Eqn. (8), we have $\Delta M_y(\boldsymbol{C}) = \sum_{i=1}^m T_{i,y}^{\text{sm}}$. Taking expectation on both sides gives $\mathbb{E}[\Delta M_y(\boldsymbol{C})] = \sum_{i=1}^m \mathbb{E}\big[T_{i,y}^{\text{sm}}\big]$. For each concept token i, by the law of total expectation,

$$\begin{aligned}\mathbb{E}\big[T_{i,y}^{\text{sm}}\big] &= \mathbb{P}(Z_i = 1)\mathbb{E}\big[T_{i,y}^{\text{sm}} \mid Z_i = 1\big] + \mathbb{P}(Z_i = 0)\mathbb{E}\big[T_{i,y}^{\text{sm}} \mid Z_i = 0\big] \\ &= a\,\mathbb{E}\big[T_{i,y}^{\text{sm}} \mid Z_i = 1\big] + (1-a)\,\mathbb{E}\big[T_{i,y}^{\text{sm}} \mid Z_i = 0\big].\end{aligned}$$

Using Assumption 1, we obtain $\mathbb{E}\big[T_{i,y}^{\text{sm}}\big] \geq a\mu_y^+ - (1-a)\mu_y^-$. Therefore,

$$\mathbb{E}[\Delta M_y(\boldsymbol{C})] \geq m\left(a\mu_y^+ - (1-a)\mu_y^-\right).$$

The right-hand side is positive whenever $a\mu_y^+ - (1-a)\mu_y^- > 0$, which is equivalent to $a > \frac{\mu_y^-}{\mu_y^+ + \mu_y^-}$. Since we assume

$$a > \max_{y \neq y^*} \frac{\mu_y^-}{\mu_y^+ + \mu_y^-},$$

the above condition holds for every competing class $y \in \mathcal{Y} \setminus \{y^*\}$. Thus, $\mathbb{E}[\Delta M_y(\boldsymbol{C})] > 0$ for all competing classes, completing the proof. □

Theorem 2 (Prediction invariance under exact softmax attention)**.** *Suppose the prediction obtained with concept block* $\boldsymbol{C}$ *is* $\hat{y}(\boldsymbol{C}) = y^*$*. Then for any alternative concept block* $\boldsymbol{C}' = \boldsymbol{C} + \boldsymbol{E}$ *satisfying Assumption 2, if*

$$\min_{y \neq y^*} M_y(\boldsymbol{C}) > B\sqrt{m_e}\varepsilon_c,$$

the pairwise margins under $\boldsymbol{C}'$ *remain positive for all* $y \in \mathcal{Y} \setminus \{y^*\}$*, i.e.,* $M_y(\boldsymbol{C}') > 0$*, and hence the prediction is invariant:*

$$\hat{y}(\boldsymbol{C}') = \hat{y}(\boldsymbol{C}) = y^*.$$

Proof. Fix a concept block $\boldsymbol{C}$ such that $\hat{y}(\boldsymbol{C}) = y^*$, and let $\boldsymbol{C}' = \boldsymbol{C} + \boldsymbol{E}$ be any alternative concept block satisfying Assumption 2. For each concept token, define the interpolation $\boldsymbol{c}_i(\tau) = (1-\tau)\boldsymbol{c}_i + \tau\boldsymbol{c}_i' = \boldsymbol{c}_i + \tau\boldsymbol{e}_i$ for $\tau \in [0,1]$, and let $\mathcal{F}_\tau$ denote the exact softmax attention output under the concept block induced by $\boldsymbol{C}_\tau = [\boldsymbol{c}_i(\tau)]_{i=1}^m$, i.e., $\mathcal{F}_\tau = \mathcal{F}_{\text{cm}}(\boldsymbol{q}; \boldsymbol{C}_\tau)$. Thus, $\mathcal{F}_0 = \mathcal{F}_{\text{cm}}(\boldsymbol{q}; \boldsymbol{C})$ and $\mathcal{F}_1 = \mathcal{F}_{\text{cm}}(\boldsymbol{q}; \boldsymbol{C}')$.

For concept token i, recall that $\boldsymbol{v}_i^c(\tau) = \boldsymbol{W}_V\boldsymbol{c}_i(\tau)$ and $r_i^c(\boldsymbol{q}, \tau) = \frac{\langle \boldsymbol{q}, \boldsymbol{W}_K\boldsymbol{c}_i(\tau)\rangle}{\sqrt{d_k}}$. The base-token scores $s_j(\boldsymbol{q})$ and values $\boldsymbol{v}_j^x$ remain fixed along the path. Let $Z(\boldsymbol{q}, \tau) = S_X(\boldsymbol{q}) + \sum_{h=1}^m \exp(r_h^c(\boldsymbol{q}, \tau))$ be the softmax denominator along the interpolation path. The corresponding base-token and concept-token attention weights are $\beta_j(\boldsymbol{q}, \tau) = \frac{\exp(s_j(\boldsymbol{q}))}{Z(\boldsymbol{q}, \tau)}$ and $\alpha_i(\boldsymbol{q}, \tau) = \frac{\exp(r_i^c(\boldsymbol{q}, \tau))}{Z(\boldsymbol{q}, \tau)}$, respectively. Thus the exact softmax attention output can be written as

$$\mathcal{F}_\tau = \sum_{j=1}^l \beta_j(\boldsymbol{q}, \tau)\boldsymbol{v}_j^x + \sum_{i=1}^m \alpha_i(\boldsymbol{q}, \tau)\boldsymbol{v}_i^c(\tau).$$

We now compute its derivative with respect to τ. Since $\boldsymbol{v}_j^x$ is fixed and $d\boldsymbol{v}_i^c(\tau)/d\tau = \boldsymbol{W}_V\boldsymbol{e}_i$, we have

$$\frac{d\mathcal{F}_\tau}{d\tau} = \sum_{j=1}^l \frac{d\beta_j(\boldsymbol{q}, \tau)}{d\tau}\boldsymbol{v}_j^x + \sum_{i=1}^m \frac{d\alpha_i(\boldsymbol{q}, \tau)}{d\tau}\boldsymbol{v}_i^c(\tau) + \sum_{i=1}^m \alpha_i(\boldsymbol{q}, \tau)\boldsymbol{W}_V\boldsymbol{e}_i.$$

Note that

$$\frac{dZ(\boldsymbol{q},\tau)}{d\tau} = \sum_{h=1}^{m} \exp(r_h^c(\boldsymbol{q},\tau)) \frac{dr_h^c(\boldsymbol{q},\tau)}{d\tau}.$$

Therefore,

$$\frac{d\beta_j(\boldsymbol{q},\tau)}{d\tau} = -\frac{\exp(s_j(\boldsymbol{q}))}{Z(\boldsymbol{q},\tau)^2} \frac{dZ(\boldsymbol{q},\tau)}{d\tau} = -\beta_j(\boldsymbol{q},\tau) \sum_{h=1}^{m} \alpha_h(\boldsymbol{q},\tau) \frac{dr_h^c(\boldsymbol{q},\tau)}{d\tau}.$$

Similarly, for the concept-token weight,

$$\begin{aligned}\frac{d\alpha_i(\boldsymbol{q},\tau)}{d\tau} &= \frac{\exp(r_i^c(\boldsymbol{q},\tau)) \frac{dr_i^c(\boldsymbol{q},\tau)}{d\tau} Z(\boldsymbol{q},\tau) - \exp(r_i^c(\boldsymbol{q},\tau)) \frac{dZ(\boldsymbol{q},\tau)}{d\tau}}{Z(\boldsymbol{q},\tau)^2} \\ &= \alpha_i(\boldsymbol{q},\tau) \frac{dr_i^c(\boldsymbol{q},\tau)}{d\tau} - \alpha_i(\boldsymbol{q},\tau) \sum_{h=1}^{m} \alpha_h(\boldsymbol{q},\tau) \frac{dr_h^c(\boldsymbol{q},\tau)}{d\tau}.\end{aligned}$$

Substituting these two identities into the derivative of $\mathcal{F}_\tau$ gives

$$\begin{aligned}\frac{d\mathcal{F}_\tau}{d\tau} =& \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \boldsymbol{W}_V \boldsymbol{e}_i + \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \frac{dr_i^c(\boldsymbol{q},\tau)}{d\tau} \boldsymbol{v}_i^c(\boldsymbol{q},\tau) \\ &- \left(\sum_{h=1}^{m} \alpha_h(\boldsymbol{q},\tau) \frac{dr_h^c(\boldsymbol{q},\tau)}{d\tau} \right) \left(\sum_{j=1}^{l} \beta_j(\boldsymbol{q},\tau) \boldsymbol{v}_j^x + \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \boldsymbol{v}_i^c(\boldsymbol{q},\tau) \right) \\ =& \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \boldsymbol{W}_V \boldsymbol{e}_i + \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \frac{dr_i^c(\boldsymbol{q},\tau)}{d\tau} \left(\boldsymbol{v}_i^c(\boldsymbol{q},\tau) - \mathcal{F}_\tau \right) \\ =& \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \left[\boldsymbol{W}_V \boldsymbol{e}_i + \frac{dr_i^c(\boldsymbol{q},\tau)}{d\tau} \left(\boldsymbol{v}_i^c(\boldsymbol{q},\tau) - \mathcal{F}_\tau \right) \right].\end{aligned}$$

We first bound the value-change term. Since the concept attention weights are nonnegative and their sum is at most one,

$$\left\| \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \boldsymbol{W}_V \boldsymbol{e}_i \right\|_2 \leq \|\boldsymbol{W}_V\|_2 \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \|\boldsymbol{e}_i\|_2 \leq \|\boldsymbol{W}_V\|_2 \|\boldsymbol{E}\|_F.$$

We next bound the softmax-weight-change term. The score derivative satisfies

$$\left| \frac{dr_i^c(\boldsymbol{q},\tau)}{d\tau} \right| = \left| \frac{\langle \boldsymbol{q}, \boldsymbol{W}_K \boldsymbol{e}_i \rangle}{\sqrt{d_k}} \right| \leq \frac{\|\boldsymbol{W}_K^\top \boldsymbol{q}\|_2}{\sqrt{d_k}} \|\boldsymbol{e}_i\|_2.$$

By Assumption 2 (A1), $\|\boldsymbol{x}_j\|_2 \leq R$, $\|\boldsymbol{c}_i\|_2 \leq R$, and $\|\boldsymbol{c}_i'\|_2 \leq R$. Hence $\|\boldsymbol{c}_i(\tau)\|_2 \leq R$ for all $\tau \in [0,1]$. Therefore every value vector along the path has norm at most $\|\boldsymbol{W}_V\|_2 R$. Since $\mathcal{F}_\tau$ is a convex combination of these value vectors, $\|\mathcal{F}_\tau\|_2 \leq \|\boldsymbol{W}_V\|_2 R$, and consequently

$$\|\boldsymbol{v}_i^c(\boldsymbol{q},\tau) - \mathcal{F}_\tau\|_2 \leq 2\|\boldsymbol{W}_V\|_2 R.$$

Combining the last two bounds gives

$$\begin{aligned}\left\| \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \frac{dr_i^c(\boldsymbol{q},\tau)}{d\tau} \left(\boldsymbol{v}_i^c(\boldsymbol{q},\tau) - \mathcal{F}_\tau \right) \right\|_2 &\leq 2R \frac{\|\boldsymbol{W}_K^\top \boldsymbol{q}\|_2}{\sqrt{d_k}} \|\boldsymbol{W}_V\|_2 \sum_{i=1}^{m} \alpha_i(\boldsymbol{q},\tau) \|\boldsymbol{e}_i\|_2 \\ &\leq 2R \frac{\|\boldsymbol{W}_K^\top \boldsymbol{q}\|_2}{\sqrt{d_k}} \|\boldsymbol{W}_V\|_2 \|\boldsymbol{E}\|_F.\end{aligned}$$

Thus, for all $\tau \in [0,1]$,

$$\left\| \frac{d\mathcal{F}_\tau}{d\tau} \right\|_2 \leq \left(\|\boldsymbol{W}_V\|_2 + 2R \frac{\|\boldsymbol{W}_K^\top \boldsymbol{q}\|_2}{\sqrt{d_k}} \|\boldsymbol{W}_V\|_2 \right) \|\boldsymbol{E}\|_F.$$

Integrating over $\tau \in [0,1]$ yields

$$\|\mathcal{F}_{\mathrm{cm}}(\boldsymbol{q};\boldsymbol{C}') - \mathcal{F}_{\mathrm{cm}}(\boldsymbol{q};\boldsymbol{C})\|_2 \le \left(\|\boldsymbol{W}_V\|_2 + 2R\frac{\|\boldsymbol{W}_K^\top \boldsymbol{q}\|_2}{\sqrt{d_k}}\|\boldsymbol{W}_V\|_2\right)\|\boldsymbol{E}\|_F.$$

For any competing class $y \neq y^*$,

$$\begin{aligned}|M_y(\boldsymbol{C}') - M_y(\boldsymbol{C})| &= \left|\boldsymbol{u}_y^\top\left(\mathcal{F}_{\mathrm{cm}}(\boldsymbol{q};\boldsymbol{C}') - \mathcal{F}_{\mathrm{cm}}(\boldsymbol{q};\boldsymbol{C})\right)\right| \\ &\le \|\boldsymbol{u}_y\|_2\,\|\mathcal{F}_{\mathrm{cm}}(\boldsymbol{q};\boldsymbol{C}') - \mathcal{F}_{\mathrm{cm}}(\boldsymbol{q};\boldsymbol{C})\|_2 \\ &\le \|\boldsymbol{u}_y\|_2\left(\|\boldsymbol{W}_V\|_2 + 2R\frac{\|\boldsymbol{W}_K^\top \boldsymbol{q}\|_2}{\sqrt{d_k}}\|\boldsymbol{W}_V\|_2\right)\|\boldsymbol{E}\|_F.\end{aligned}$$

By Assumption 2 (A4), this implies

$$|M_y(\boldsymbol{C}') - M_y(\boldsymbol{C})| \le B\|\boldsymbol{E}\|_F.$$

Moreover, by Assumption 2 (A2) and (A3), at most m_e columns of $\boldsymbol{E}$ are nonzero and each nonzero column has norm at most ε_c, so $\|\boldsymbol{E}\|_F \le \sqrt{m_e}\varepsilon_c$. Therefore,

$$|M_y(\boldsymbol{C}') - M_y(\boldsymbol{C})| \le B\sqrt{m_e}\varepsilon_c.$$

If $\min_{y\neq y^*} M_y(\boldsymbol{C}) > B\sqrt{m_e}\varepsilon_c$, then for every competing class $y \neq y^*$,

$$M_y(\boldsymbol{C}') \ge M_y(\boldsymbol{C}) - |M_y(\boldsymbol{C}') - M_y(\boldsymbol{C})| > 0.$$

Thus, under $\boldsymbol{C}'$, the logit of y^* remains larger than that of every competing class. Hence

$$\hat{y}(\boldsymbol{C}') = \hat{y}(\boldsymbol{C}) = y^*,$$

completing the proof. □

B Prompt Template

To facilitate reproducibility, we provide full prompt templates and concept extraction instructions. Our method is training-free and does not rely on any parameter updates or external supervision. All experiments can be reproduced using standard LVLM inference with the provided prompts and settings.

ICL Prompt. We adopt a standard in-context learning template that provides a task description and a set of image-label demonstrations, followed by the test query.

[**Task Description**]

N-shot Demonstration: {[Image][Label]}

Query: [Image][Question]

Prediction: [Answer]

[**Task Description**]: Given an image and a multiple-choice question, select the single best answer from the provided options.

[**Question**]: Now, based on the demonstrations above, decide which option best matches the NEW image shown.

Options:

(A) Option A

(B) Option B

(C) Option C

(D) Option D

Answer with only the letter (A, B, C, or D) on the LAST line.

CM-ICL Prompt. CM-ICL uses a two-stage prompting procedure consisting of concept extraction and concept-conditioned prediction. The two stages define the logical structure of concept mediation: the model first produces a compact concept block from the image, and the final prediction is then conditioned on both the image and this concept block.

Stage 1: Concept extraction. The model receives the query image and is instructed to list a fixed number of short discriminative visual clues. The prompt is:

Carefully observe the image and list exactly {N concepts} distinctive visual clues on the main object. Each clue should be short, focusing on discriminative features such as color, texture, shape, or structure. Avoid background details or context words. Do not mention any class names directly. List one clue per line.

Stage 2: Concept-conditioned prediction. The generated concept block is inserted into the final prediction prompt together with the original image and question:

Here are image-derived visual clues: [Concepts]

Which candidate class best fits the image?

Options: (A) Option A (B) Option B (C) Option C (D) Option D

Answer with only the letter (A, B, C, or D) on the last line.

C Additional Experimental Results

This section provides experimental results to complement the main evaluation in Section 5.

Model	Dog-120	Bird-200	Aircraft-102	Flower-102	Pet-37	Car-196
Qwen2.5-VL	74.50 ± 0.03	66.37 ± 0.05	63.86 ± 0.05	82.89 ± 0.13	90.74 ± 0.02	85.12 ± 0.01
Qwen2.5-VL-CoT	74.77 ± 0.09	65.66 ± 0.04	64.72 ± 0.11	81.82 ± 0.19	88.86 ± 0.04	85.03 ± 0.10
Qwen2.5-VL-CM-ICL	**76.75 ± 0.06**	**70.48 ± 0.07**	**66.94 ± 0.08**	**83.11 ± 0.01**	**91.01 ± 0.02**	**85.98 ± 0.03**
Qwen3-VL	64.46 ± 0.14	58.87 ± 0.23	54.74 ± 0.20	82.90 ± 0.07	82.60 ± 0.21	74.86 ± 0.11
Qwen3-VL-CoT	64.59 ± 0.26	54.59 ± 1.63	45.62 ± 0.41	79.93 ± 0.39	73.68 ± 0.15	67.04 ± 0.37
Qwen3-VL-CM-ICL	**76.34 ± 0.06**	**73.06 ± 0.14**	**76.61 ± 0.46**	**85.10 ± 0.10**	**89.75 ± 0.14**	**88.09 ± 0.10**
InternVL	54.15 ± 0.30	47.00 ± 0.48	49.63 ± 0.49	52.96 ± 0.28	73.63 ± 0.07	62.80 ± 0.12
InternVL-CoT	55.77 ± 0.10	56.60 ± 0.66	38.46 ± 1.08	56.97 ± 0.46	72.78 ± 0.41	62.05 ± 0.22
InternVL-CM-ICL	**62.62 ± 0.33**	**66.54 ± 0.21**	**60.41 ± 0.09**	**58.12 ± 0.08**	**75.98 ± 0.31**	**75.57 ± 0.28**

Table 3: Accuracy of LVLMs across six fine-grained datasets with sampling-enabled decoding. Results are reported as mean ± standard deviation over multiple runs. Best results in each column are highlighted in bold. Light-blue rows indicate concept-mediated prompting.

Stability under repeated runs. The main results reported in Section 5.2 use deterministic decoding (sampling disabled), which yields fixed outputs and therefore does not introduce variance from decoding. To complement these deterministic results, Table 3 reports top-1 accuracy under decoding with sampling enabled, summarized as mean ± standard deviation over three independent runs for the three primary LVLM backbones (Qwen2.5-VL, Qwen3-VL, and InternVL). Across datasets, concept-mediated prompting maintains competitive performance while exhibiting relatively low variance compared with baseline prompting strategies, indicating stable behavior even when decoding randomness is introduced.

Extended backbone evaluation. To assess robustness beyond the primary backbones, Table 4 includes additional results on LLaVA-OneVision-1.5 (An et al., 2025) and MiniCPM-V 4.5 (Yu et al., 2025b). Us-

Model	Dog-120	Bird-200	Aircraft-102	Flower-102	Pet-37	Car-196	Avg
Qwen2.5-VL	74.57	66.47	64.24	82.89	90.87	85.14	77.36
Qwen2.5-VL-CoT	74.84	65.91	65.02	81.83	88.78	85.03	76.9
Qwen2.5-VL-CM-ICL	**76.91**	**70.43**	**67.06**	**83.22**	**91.11**	**86.05**	**79.13**
Qwen3-VL	64.60	59.23	54.76	83.10	82.47	74.80	69.83
Qwen3-VL-CoT	63.93	57.40	48.39	80.77	73.07	66.24	64.97
Qwen3-VL-CM-ICL	**76.18**	**73.08**	**77.11**	**85.05**	**89.64**	**88.14**	**81.53**
InternVL	58.87	51.33	57.67	55.02	75.96	68.04	61.15
InternVL-CoT	58.73	57.80	39.87	57.26	73.37	61.15	58.03
InternVL-CM-ICL	**64.99**	**69.07**	**64.12**	**58.94**	**77.81**	**77.52**	**68.74**
LLaVA	53.44	52.81	55.69	66.87	73.18	71.78	62.30
LLaVA-CoT	50.98	46.58	49.38	60.22	64.62	65.02	56.13
LLaVA-CM-ICL	**53.83**	51.83	**55.84**	**67.93**	**73.29**	70.33	62.18
MiniCPM	54.21	54.71	40.47	66.03	82.39	79.44	62.88
MiniCPM-CoT	78.72	85.61	70.51	87.59	93.40	88.96	84.13
MiniCPM-CM-ICL	**84.44**	**87.25**	**74.47**	87.48	**94.47**	**93.98**	**87.02**

Table 4: Performance comparison of LVLMs across six fine-grained datasets. Best scores per column are highlighted in bold. Light-blue rows denote our concept-mediated prompting methods. All results are top-1 accuracy (%).

Shots	**Dog-120**	**Bird-200**	**Aircraft-102**	**Flower-102**	**Pet-37**	**Car-196**
4	64.72	55.35	36.60	92.50	86.54	74.10
8	66.49	58.72	38.25	95.56	88.06	76.99
16	69.10	64.36	42.33	96.91	89.51	79.38
32	70.83	68.35	44.10	97.58	90.19	81.08
CM-ICL	**76.91**	**70.43**	**67.06**	83.22	**91.11**	**86.05**

Table 5: Retrieval-based kNN-ICL results on Qwen2.5-VL across different shot numbers. Demonstrations are selected using CLIP nearest-neighbor retrieval. We report top-1 accuracy (%) on six fine-grained datasets.

ing the same evaluation protocol, the observed performance trends remain consistent, suggesting that the effectiveness of concept-mediated prompting generalizes across a broader range of LVLM architectures.

Retrieval-Based kNN-ICL Baseline. We further evaluate a retrieval-based kNN-ICL baseline to provide a stronger comparison than random demonstrations. For each query image, we retrieve nearest labeled examples using CLIP image embeddings and use them as in-context demonstrations, without any task-specific training. As shown in Table 5, kNN-ICL improves with more shots on several datasets, confirming that demonstration quality matters. Nevertheless, CM-ICL outperforms kNN-ICL on five out of six datasets. The only exception is Flower-102, where retrieved examples are particularly effective, likely due to stronger visual clustering among flower categories. Overall, these results show that CM-ICL remains competitive with a stronger retrieval-based ICL baseline while requiring no labeled demonstration pool and no long retrieved context.

Comparison with Prompt Optimization. Fig. 4 reports test accuracy on six fine-grained datasets when using Qwen2.5-VL-7B as the inference model and Qwen2.5-VL-32B as the optimizer to iteratively refine the classification prompt. Since this class of methods requires separate training, validation, and test splits for prompt selection, we further randomly split the original test set into train/validation/test subsets with a ratio of 1:1:8. This split is applied only to the prompt optimization baseline, while all other methods are evaluated on the full original test set. We observe that the blue bars (prompt optimization) are consistently below the orange baseline across all datasets, indicating that the iterative prompt-search procedure actually hurts performance in this setting. This suggests that, for fine-grained recognition, the optimizer tends to overfit the small validation split and drifts the prompt toward idiosyncratic cues that do not transfer to the

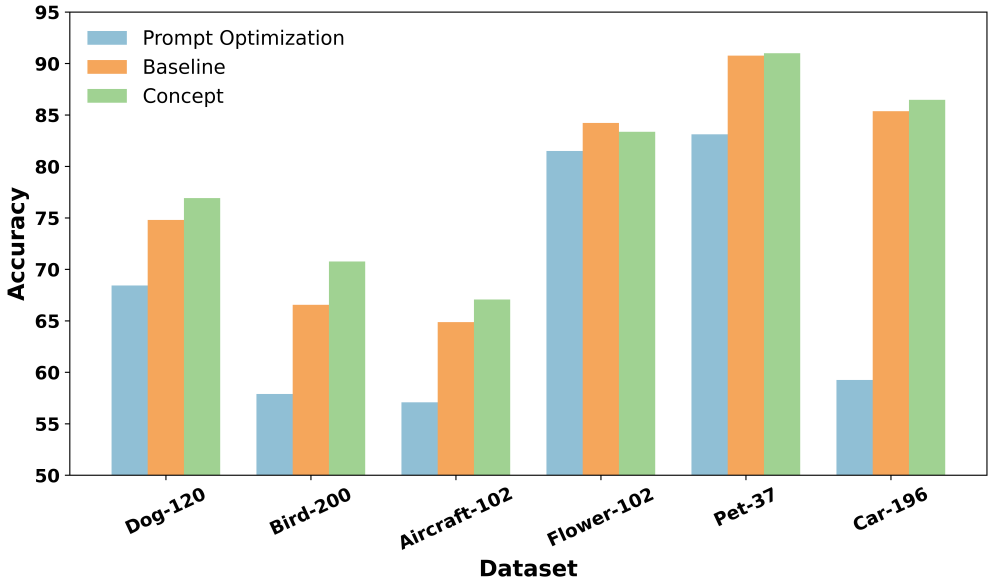


Figure 4: Test subset accuracy on six datasets using Qwen-7b as the inference model and Qwen-32b as the optimizer. Orange indicates the baseline, blue denotes prompt optimization, and green represents our method.

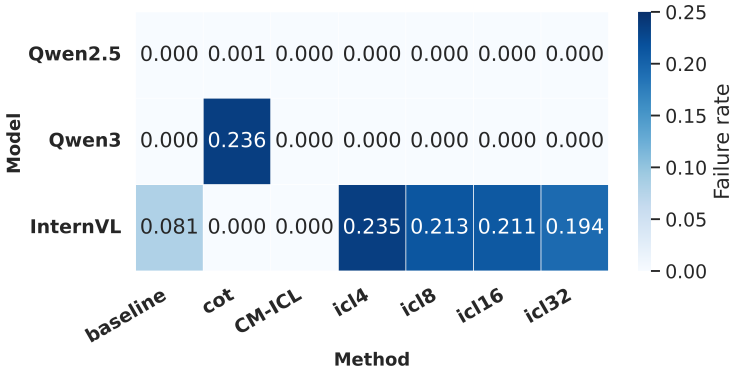


Figure 5: Generation failure rate heatmap across prompting methods (x-axis) and LVLM backbones (y-axis). Each cell shows the average fraction of invalid predictions; darker colors correspond to higher failure rates.

held-out test images. In contrast, our concept-mediated method (green) either matches or improves upon the baseline on all datasets, despite using a single, concept-structured prompt without iterative search. This highlights that, rather than investing additional computation into prompt tuning with a second LVLM, it is more effective to inject structured, part-level semantic concepts that directly align with the visual evidence required for fine-grained discrimination.

Generation Failure Rate. Fig. 5 shows the generation failure rates across prompting strategies and LVLMs. Qwen2.5-VL and Qwen3-VL exhibit near-zero failure rates, indicating stable output behavior. InternVL, however, is more sensitive to prompt format: ICL prompts cause a marked increase in invalid generations (about $19 \sim 23\%$), suggesting that long or demonstration-heavy inputs introduce unstable reasoning in fine-grained settings. In contrast, concept-mediated prompts maintain consistently low failure rates across all models, reflecting that their concise, structured form mitigates over-generation and helps control the overthinking behavior documented in related work.

Additional Localizability-Only Association Checks. In addition to the main perceptual grounding and fix-harm results, we report supplementary checks on whether image-level visual localizability alone tracks

Dataset	GCA	GGA
Dog-120	-0.058	-0.030
Bird-200	0.015	-0.028
Aircraft-102	-0.016	0.070
Flower-102	-0.098	-0.054
Pet-37	0.003	0.010
Car-196	0.036	-0.028

Table 6: Supplementary localizability-only association checks on Qwen2.5-VL. GCA is the Spearman association between image-level grounding coverage and CM-ICL correctness. GGA is the Spearman association between CM-ICL grounding coverage and correctness change relative to CoT.

downstream prediction behavior. Specifically, GCA measures the Spearman association between CM-ICL image-level valid ratio and prediction correctness, while GGA measures the association between CM-ICL valid ratio and correctness change relative to CoT. These coefficients are not used as causal evidence or for concept selection. Instead, they test whether the localizability proxy by itself is sufficient to explain task behavior. As shown in Table 6, the associations are generally modest, indicating that visual localizability alone is not a sufficient surrogate for discriminative usefulness. This is consistent with the mechanism in Sec. 3.2: a concept affects the final decision only when it receives attention and its residual value direction aligns with the class-margin direction. The primary output-level evidence is therefore the fix–harm decomposition in Table 2.

Inference Cost and Accuracy-Latency Trade-off. We report inference cost on Qwen2.5-VL across all six fine-grained datasets. All methods use the same frozen LVLM and require no parameter updates. For accuracy, we use the main-result values for all methods; for latency/token statistics, we use the cost-profiling runs. This separates the main accuracy comparison from implementation-level cost measurement. Latency should be interpreted as an implementation-level measurement, since absolute runtime can vary with hardware, inference backend, batching, and system load.

Dataset	Acc. (%)	Lat. (s)	Tokens
Dog-120	74.57 / 74.84 / 62.67 / 63.04 / **76.91**	0.13 / 4.52 / 0.22 / 1.45 / 1.89	80.7 / 297.2 / 292.2 / 1726.1 / 391.5
Bird-200	66.47 / 65.91 / 54.64 / 57.32 / **70.43**	0.13 / 5.75 / 0.24 / 1.54 / 1.95	85.1 / 362.6 / 319.1 / 1911.1 / 399.7
Aircraft-102	64.24 / 65.02 / 51.40 / 43.14 / **67.06**	0.29 / 8.29 / 0.22 / 1.45 / 2.56	89.9 / 297.6 / 347.1 / 2101.0 / 418.8
Flower-102	82.89 / 81.83 / 79.96 / 80.31 / **83.22**	0.35 / 8.31 / 0.36 / 1.52 / 2.21	76.8 / 294.1 / 270.8 / 1581.5 / 411.0
Pet-37	90.87 / 88.78 / 79.86 / 80.89 / **91.11**	0.14 / 7.04 / 0.24 / 1.35 / 1.85	76.9 / 279.7 / 270.0 / 1575.5 / 382.2
Car-196	85.14 / 85.03 / 74.63 / 73.98 / **86.05**	0.25 / 7.77 / 0.23 / 1.40 / 2.04	106.7 / 326.4 / 449.5 / 2801.5 / 440.6
Avg.	77.36 / 76.90 / 67.19 / 66.45 / **79.13**	0.22 / 6.95 / 0.25 / 1.45 / 2.08	86.0 / 309.6 / 324.8 / 1949.5 / 407.3

Table 7: Inference cost and accuracy-latency trade-off on Qwen2.5-VL. Each cell follows the order *Base / CoT / ICL-4 / ICL-32 / CM-ICL*. Accuracy values follow the main results, while latency/token statistics are from cost-profiling runs. Latency is measured in seconds per example, and Tokens denotes the average text-token count reported by the profiling run.

Table 7 shows three trends. First, CM-ICL improves average accuracy over CoT from 76.90% to 79.13% (+2.23% absolute accuracy gain) while reducing mean latency from 6.95s to 2.08s. Second, although ICL-32 has lower decoding latency than CM-ICL due to short answer-only generation, it uses much longer textual contexts in the profiling runs: 1949.5 tokens on average, compared with 407.3 for CM-ICL. Third, ICL-32 performs substantially worse than CM-ICL on all six datasets, with a 12.68% absolute accuracy gap. These results indicate that CM-ICL provides a better accuracy-cost trade-off than long free-form CoT reasoning and avoids the long demonstration context required by high-shot ICL.

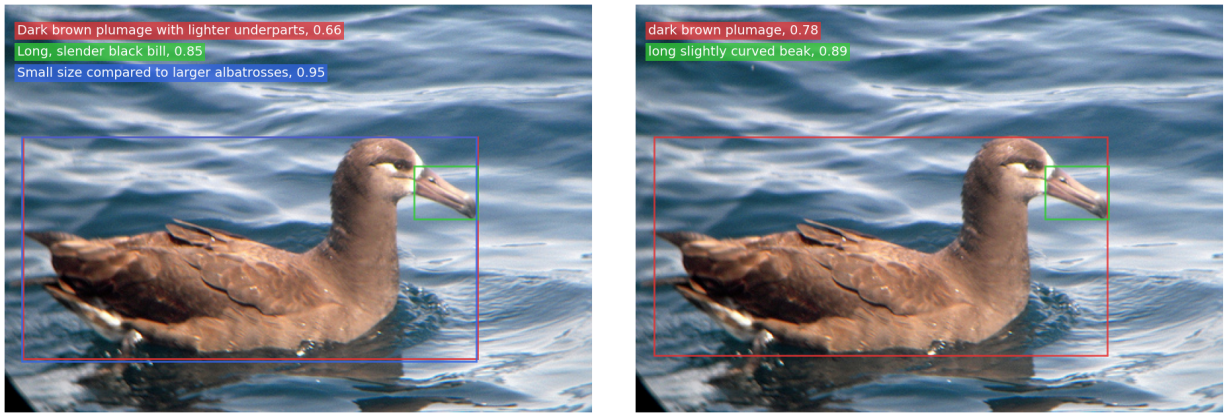
D Case Study Visualization

Figures 6 and 7 present additional qualitative case studies, serving as supplementary visual evidence for the results discussed in Section 5.3.

(a) CM-ICL correctly predicts: **DC-10**.

(b) CoT incorrectly predicts: **767-200**.

Figure 6: Qualitative comparison of SAM 3 grounding results for CM-ICL concepts and CoT-derived phrases in fine-grained aircraft recognition.



(a) CM-ICL correctly predicts: **Black-footed Albatross**.

(b) CoT incorrectly predicts: **Sooty Albatross**.

Figure 7: Qualitative comparison of SAM 3 grounding results for CM-ICL concepts and CoT-derived phrases in fine-grained bird recognition.

E Additional Related Work

E.1 Training-based Grounded Reasoning

A rapidly growing line of work seeks to improve multimodal reasoning by explicitly grounding intermediate reasoning steps in visual evidence, such as image regions, coordinates, or verified object-centric cues. Compared to standard CoT prompting, these approaches require additional training data, specialized supervision, or reinforcement learning objectives, and thus operate in a setting fundamentally different from inference-only prompting.

One representative direction is *grounded chain-of-thought* learning, where models are trained to align each reasoning step with spatial evidence such as bounding boxes or points, and to produce grounded rationales during inference (Wu et al., 2025; Man et al., 2025). Another line of work leverages visually grounded reinforcement learning or fine-tuning to reduce hallucination and enforce step-wise grounding in multi-step reasoning (Sarch et al., 2025; Ni et al., 2025). Recent efforts also construct large-scale datasets for supervised multimodal CoT training, demonstrating that post-training with grounded or preference-based supervision can substantially alter a model's reasoning behavior (Zhang et al., 2025). In addition, some approaches

introduce architectural components that first detect or propose relevant visual regions and then perform reasoning over the selected evidence, which again involves task-specific training or auxiliary modules beyond pure prompting (Wang et al., 2025). Verification-oriented methods aim to assess whether each intermediate reasoning step is supported by visual evidence, relying on learned verifiers or additional optimization mechanisms (Yi & Shang, 2025).

These grounded reasoning approaches are complementary to our work but are not directly comparable. They assume access to grounded supervision, curated reasoning datasets, reinforcement learning signals, or specialized architectures, whereas our focus is on a strictly training-free setting with frozen multimodal models. Moreover, most grounded reasoning methods are developed for general-purpose multimodal reasoning or visual question answering, while our work targets fine-grained visual recognition, where success hinges on capturing subtle attribute-level distinctions under minimal adaptation. Our goal is to study how far one can push fine-grained recognition using only inference-time structured guidance, without any form of grounded supervision or post-training.

E.2 Optimized In-Context Learning

Beyond vanilla prompting, a substantial body of work has explored how in-context learning (ICL) can be enhanced through optimized demonstration selection, annotation strategies, or iterative refinement. These studies provide important insights into the mechanics of ICL, but typically introduce additional assumptions that go beyond a strictly training-free paradigm.

Early analyses demonstrate that the effectiveness of ICL is highly sensitive to the choice of demonstration examples. Liu et al. (2022) show that selecting semantically similar or task-relevant examples can significantly improve performance, motivating retrieval-based or selection-aware ICL pipelines. Building on this observation, Zhang et al. (2023) extend the analysis to visual in-context learning and show that carefully chosen image–label demonstrations substantially affect recognition accuracy, highlighting the role of learned visual representations in demonstration quality. Subsequent work formulates demonstration selection as an explicit optimization problem. Selective annotation methods (Su et al., 2022) actively choose which examples to label in order to maximize few-shot performance, effectively trading annotation effort for improved in-context generalization. More recent approaches introduce iterative or adaptive selection mechanisms, where candidate demonstrations are repeatedly evaluated and refined during inference (Qin et al., 2024). While effective, these methods rely on annotated pools, auxiliary selection procedures, or validation feedback, thereby introducing additional supervision or optimization beyond pure prompting. Comprehensive surveys (Dong et al., 2024) summarize these developments and emphasize that many practical gains in ICL arise from refined context construction rather than from the base model alone. As a result, such approaches blur the boundary between inference-time conditioning and lightweight forms of learning or optimization.

In contrast, our work focuses on a minimal and complementary setting. We study how frozen LVLMs can be guided for fine-grained visual recognition using only inference-time, image-derived structured concepts, without any additional training, annotation, or demonstration optimization. This design is particularly suitable for fine-grained scenarios with a large number of categories, where selecting representative in-context demonstrations becomes increasingly challenging.