C2FT: Enhancing Fine-Grained Perception in MLLMs via Confuse-then-Contrast Fine-Tuning

Shaoxuan He* **Benlei Cui*✉** **Shikai Qiu** **Yuwen Zhai**
Hui Xue **Longtao Huang** **Jingqun Tang** **Haiwen Hong✉**

Yuvion Team, Alibaba Group

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities, yet they frequently struggle with fine-grained visual perception, often suffering from hallucinations when faced with visually similar but semantically distinct instances. A major underlying cause is the conventional single-sample Supervised Fine-Tuning (SFT) paradigm, which isolates training instances and limits the model's ability to explicitly learn subtle discriminative boundaries. To address this, we propose **C2FT**, a novel **Confuse-then-Contrast Fine-Tuning** framework that enhances fine-grained perception by shifting from independent supervision to joint supervision. Specifically, C2FT assesses the MLLM's internal uncertainty to dynamically mine model-specific semantic confusions, subsequently constructing challenging multi-image groups comprising hard positive and hard negative samples. By interleaving these samples into a unified prompt, our approach forces the MLLM's internal attention mechanisms to cross-reference images and explicitly capture localized visual discrepancies. Extensive experiments on both Fine-Grained Visual Classification (FGVC) and visual question answering (VQA) benchmarks demonstrate the effectiveness of C2FT. For instance, on the Qwen3-VL-4B model, our approach achieves an average improvements of 3.87% and 4.35% points over standard SFT and GRPO baselines, respectively. Code is available at: `https://github.com/Alibaba-VELLDEPTH/C2FT`.

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved remarkable progress across a broad range of vision-language tasks through large-scale pre-training [1, 2, 3, 4, 5], demonstrating strong global semantic understanding capabilities. They can recognize common objects, interpret complex scenes, and generate fluent language responses conditioned on visual inputs. Despite these advances, their visual perception remains largely coarse-grained [6, 7, 8]. In tasks requiring precise visual discrimination [9, 10], such as spatial reasoning [11, 12, 13, 14], fine-grained instance recognition [15, 16], and localized detail attribution [17, 18], even advanced MLLMs often hallucinate or rely on superficial cues [19, 20, 21]. This limitation prevents them from capturing subtle structural, textural, or spatial differences, which are essential for distinguishing visually similar instances and resolving fine-grained visual ambiguities.

Recent studies have primarily sought to mitigate this issue through specialized supervised fine-tuning [22, 23, 24] or reinforcement learning methods [25, 26] such as DPO. Although effective to some extent, these approaches often require expensive fine-grained annotations [8, 27, 28] or manually crafted Chain-of-Thought (CoT) rationales [29, 30, 31]. More importantly, most existing approaches follow an independent supervision paradigm, in which each training sample is optimized

*Equal contribution. ✉Corresponding authors: cuibenlei.cbl@alibaba-inc.com, honghaiwen.hhw@alibaba-inc.com.

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

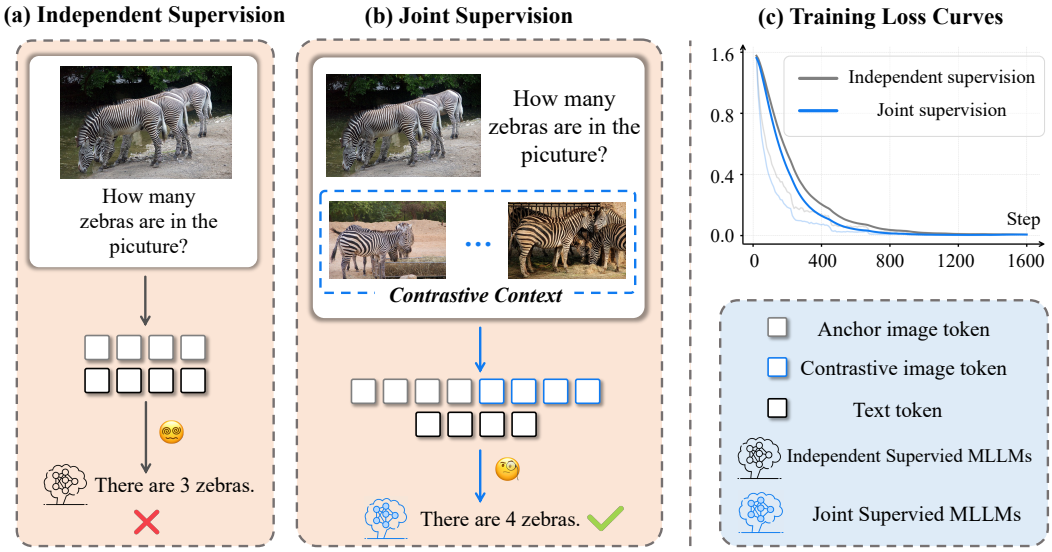


Figure 1: **Comparison between Independent and Joint Supervision paradigms for MLLMs.** (a) Under standard independent supervision, MLLMs often fail on complex queries due to a lack of comparative context. (b) By incorporating auxiliary images as a contrastive context pool, our joint supervision explicitly encourages the model to differentiate fine-grained features. (c) Joint supervision achieves faster convergence than the independent baseline.

in isolation on the input side. This isolated training scheme provides limited contrastive supervision and thus makes it difficult for the model to identify subtle instance-level differences. As shown in Fig. 1(a), models tend to form coarse semantic associations rather than sharp discriminative boundaries, leaving them vulnerable to confusing fine-grained details.

In this paper, we propose **C2FT**, a **Confuse-then-Contrast Fine-Tuning** framework to enhance fine-grained perception in MLLMs. Inspired by the success of contrastive approaches [32, 33] in traditional vision tasks, we integrate a comparative mechanism directly into the MLLM with input side joint supervision. Specifically, C2FT first identifies samples that are likely to confuse the model based on output likelihoods and visual feature similarities. It then constructs hard negative pairs with similar visual appearances but different answers, as well as hard positive pairs with diverse appearances but similar answers. These samples are combined into unified multi-image instructions, enabling cross-image attention and encouraging the model to explicitly compare instances. Moreover, to make multi-image training more robust, we randomly shuffle image orders to reduce positional bias and mix the constructed multi-image data with single-image samples to mitigate catastrophic forgetting [34, 35]. As shown in Fig. 1(b), this process encourages MLLMs to learn fine-grained details beyond standard independent supervision.

To validate the effectiveness of C2FT, we conduct extensive experiments across diverse fine-grained multimodal benchmarks, encompassing both FGVC and VQA tasks. Empirical results demonstrate that our approach outperforms standard SFT and GRPO baselines. Our method achieves an overall average score of 82.95% on Qwen3-VL-4B, exceeding the SFT and GRPO baselines by 3.87 and 4.35 points, respectively; on Gemma-4-E2B, the corresponding gains are 2.72 and 3.56 points. Furthermore, we observe a promising scaling trend: *expanding the pool of interleaved contrastive samples consistently enhances fine-grained perceptual performance*. These findings further support our core motivation that enabling MLLMs to explicitly learn from model-specific confusions and cross-image comparisons is crucial for resolving subtle visual ambiguities. Our contributions are threefold:

- We identify the limitations of isolated single supervised training for fine-grained multimodal perception, highlighting its lack of contrastive supervision for distinguishing visually similar but semantically different instances.

- We propose C2FT, a Confuse-then-Contrast Fine-Tuning framework. This approach dynamically mines model-specific hard positives and hard negatives and organizes them into multi-image instructions to enforce explicit cross-image attention and discriminative learning.
- We conduct extensive experiments on FGVC and VQA benchmarks, showing that C2FT outperforms standard SFT and GRPO baselines across different model scales, while also revealing a clear scaling trend as the number of interleaved contrastive samples increases.

2 Related Work

2.1 Fine-grained Perception of MLLMs

MLLMs have been widely deployed across a broad spectrum of real-world applications [36, 37, 38, 39, 40], motivating increasing attention to their fine-grained visual perception capabilities. One line of work adopts training-free strategies [41], such as prompt engineering [42], or attention-based feature extraction [43, 44], to better elicit fine-grained visual cues without additional model updates. While computationally efficient, these methods often lack sufficient discriminative supervision, making it difficult to reliably separate visually similar categories. Another line of work employs knowledge distillation (KD) [27, 45, 46] to transfer task-specific fine-grained skills from powerful teacher models to more efficient student networks. Although KD significantly reduces computational overhead [47, 48], it primarily aligns intermediate features or output logits, often missing the critical fine-grained alignment between localized visual objects and specific categories [27, 49].

Beyond direct perception enhancement, chain-of-thought (CoT) reasoning has been introduced to encourage MLLMs to decompose fine-grained recognition into intermediate visual evidence before predicting the final category [50, 30]. By providing explicit reasoning traces, CoT can improve interpretability and visual grounding. While explicit reasoning traces can improve interpretability and visual grounding, their effectiveness depends heavily on the quality of the reasoning data, which is difficult to obtain at scale for fine-grained tasks [51, 52]. Reinforcement learning [53, 54, 55] offers another way to improve visual alignment by rewarding image-grounded responses. However, their application to fine-grained visual perception usually requires specially designed reward signals and suffers from reward hacking [56, 57].

2.2 Visual Contrastive Training

The concept of visual contrastive training refers to learning visual representations by encouraging semantically related inputs to be closer in the feature space while pushing unrelated ones farther apart [58]. This paradigm has shown strong effectiveness in self-supervised learning [59, 60, 61] where different augmented views of the same image are treated as positive pairs and other samples serve as negatives. Beyond such view-level objectives [62], prior work has also investigated explicit cross-image contrast [63, 64, 65], which enables models to capture finer-grained distinctions and relational cues across instances. In multimodal settings, contrastive vision-language alignment [66, 67, 68] has significantly advanced MLLMs, enhancing visual discrimination and relational reasoning. Recently, contrastive principles have emerged as data construction strategies for instruction tuning [69, 70, 71], leveraging hard negatives to refine MLLM comparative perception and fine-grained judgment. However, how to effectively perform visual contrasting for MLLMs without external knowledge remains underexplored. To address this, C2FT introduces a confuse-then-contrast paradigm that dynamically mines hard contrastive samples, providing explicit joint supervision for fine-grained perception.

3 Method

Our method, C2FT, enhances fine-grained perception in MLLMs through a confuse-then-contrast training paradigm. We first introduce a mining module that leverages the model's own uncertainty to identify visually or semantically confusable categories, and uses a strong encoder [72, 73] to retrieve hard positive and negative samples. We then adopt a progressive anti-shortcut training strategy that combines shuffled multi-image contrastive training with mixed-format supervision. This progressive design encourages the model to capture discriminative fine-grained visual cues, while mitigating

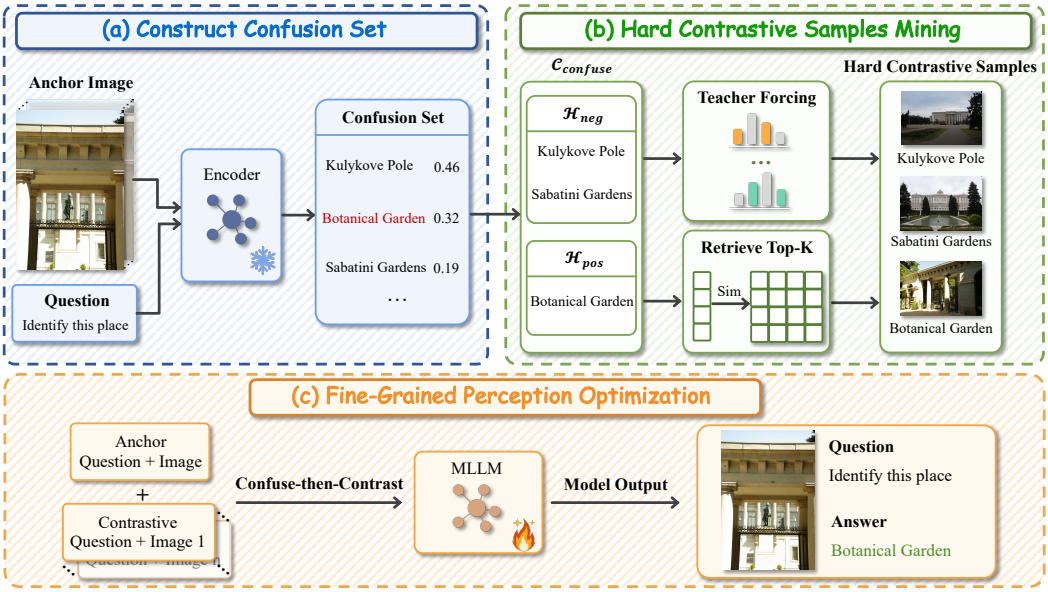


Figure 2: **Overview of C2FT.** (a) C2FT first constructs a dynamic semantic confusion set for the given anchor samples. (b) Then hard contrastive samples are mined: positives are retrieved directly through feature similarities, whereas negatives are verified via Teacher Forcing generation probabilities. (c) The assembled multi-image context is joint optimized fine-grained visual perception.

positional shortcuts and preserving its basic recognition capability. Fig. 2 provides an overview of the proposed C2FT framework.

3.1 Preliminary

Joint Attention over Multi-Image. Modern MLLMs typically encode each input image into a sequence of visual tokens and concatenate them with textual tokens before feeding them into a Transformer-based language model. Given multiple images, their visual tokens are placed in a shared token sequence, allowing the self-attention layers to compute dependencies not only between text and image tokens, but also across visual tokens from different images. We refer to this mechanism as joint attention over multiple images.

Formally, let $\mathbf{V}_i = \{v_i^1, \ldots, v_i^{n_i}\}$ denote the visual tokens of the i-th image, and let $\mathbf{T}$ denote the textual tokens. The input sequence can be written as

$$\mathbf{X} = [\mathbf{V}_1; \mathbf{V}_2; \ldots; \mathbf{V}_K; \mathbf{T}],$$

where K is the number of images. Standard self-attention is then computed over the entire sequence:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right)\mathbf{V}.$$

Since all image tokens participate in the same attention operation, tokens from one image can directly attend to tokens from another image. This provides a structural basis for cross-image comparison and enables the model to relate visually similar regions, objects, or attributes across different images.

Comparative Supervision for Discrimination. Fine-grained perception is inherently comparative, as visually similar samples often differ only in subtle attributes [32]. Standard single-sample SFT maximizes $p_\theta(y \mid I, q)$ for each image-text pair independently, which provides limited discriminative signal. Comparative supervision addresses this limitation by introducing a positive image I^+ and a set of hard negatives $\mathcal{I}^- = \{I_1^-, \ldots, I_m^-\}$. The model is encouraged to rank the matched pair above mismatched ones:

$$s_\theta(I^+, q, y) > s_\theta(I_j^-, q, y), \quad \forall I_j^- \in \mathcal{I}^-,$$

where s_θ denotes the model's compatibility score. By contrasting visually similar candidates, this supervision makes discriminative cues more salient and encourages the model to exploit multi-image joint attention for fine-grained reasoning.

3.2 Confuse-then-Contrast Mining

The Confuse-then-Contrast Mining (C2M) module dynamically constructs challenging multi-image exemplars by coupling the MLLM's intrinsic semantic uncertainty with fine-grained visual retrieval. Operating on-the-fly at the end of each training epoch, C2M proceeds in two sequential steps: model-centric confusion estimation and fine-grained visual contrastive retrieval.

Model-Centric Confusion Estimation. Traditional hard negative mining typically relies on static visual similarities, which often fails to capture the actual semantic weaknesses of MLLMs dynamically. To address this, we propose a model-centric confusion estimation strategy based on feature-level semantic retrieval.

Prior to training, we pre-compute and construct an offline feature bank. This includes extracting the text embeddings for all ground-truth (GT) answers and query texts, as well as the visual embeddings for all images using a frozen pre-trained encoder. At the end of each epoch, we then activate the dynamic estimation mechanism.

Given an anchor sample x_a with its ground-truth y_a, if the model yields a wrong answer, we extract the text embedding of this generated incorrect prediction, denoted as e_{pred}. To quantify the model's current semantic confusion, we calculate the similarity (e.g., cosine similarity) between e_{pred} and the pre-computed GT text embeddings e^c_{gt} of all other candidate categories $c \in \mathcal{C}$:

$$S_{confuse}(c \mid x_a) = \text{Sim}(e_{pred}, e^c_{gt}). \tag{1}$$

By sorting these candidate categories in descending order based on their similarity scores $S_{confuse}$, we obtain a ranked confusion set $\mathcal{C}_{confuse}$. The categories ranked at the top of $\mathcal{C}_{confuse}$ represent the concepts that the current model is most heavily confusing with the anchor, thereby serving as highly targeted semantic anchors for the subsequent fine-grained visual retrieval.

Fine-Grained Visual Contrastive Retrieval. Having identified the dynamic semantic confusion set $\mathcal{C}_{confuse}$, we proceed to construct a highly challenging multi-image context. To ensure the contextual instances provide optimal contrastive difficulty, we apply distinct retrieval and verification strategies for hard negative and hard positive samples.

For **hard negative samples**, we draw candidate images from the classes in $\mathcal{C}_{confuse}$. To verify whether a specific candidate sample genuinely induces cognitive confusion for the MLLM, we employ a Teacher Forcing verification step. For a candidate negative sample x_{neg} belonging to a confused class $c \in \mathcal{C}_{confuse}$, we compute the model's generation probability of the anchor's ground-truth label y_a, denoted as $P_{TF}(y_a \mid x_{neg})$. We evaluate the margin between this confusion probability and the probability of predicting its true label $P_{TF}(c \mid x_{neg})$. If the confusion probability is critically high, which indicates a small margin with the ground truth, then x_{neg} is explicitly verified and collected into the hard negative set $\mathcal{H}_{neg}$:

$$\mathcal{H}_{neg} = \{x_{neg} \in \mathcal{D}_c \mid c \in \mathcal{C}_{confuse},\ |P_{TF}(y_a \mid x_{neg}) - P_{TF}(c \mid x_{neg})| < \tau\}, \tag{2}$$

where τ is a predefined threshold (see ablation study in Sec. 4.3).

For **hard positive samples**, which are distinct instances belonging to the anchor's ground-truth class y_a, we bypass the computationally expensive Teacher Forcing step. Instead, we perform an efficient multi-modal feature retrieval. We retrieve candidate sample x_{pos} using the anchor's pre-computed image and text representation, denoted as $e^i_{x_a}$ and $e^t_{x_a}$ respectively. To ensure the selected positives are visually and semantically relevant, we enforce a constraint that the retrieved samples x_{pos} must be highly aligned with the anchor's image and text embedding. The hard positive set $\mathcal{H}_{pos}$ is thus constructed by selecting the Top-K samples that maximize this joint similarity:

$$\mathcal{H}_{pos} = \underset{x_{pos} \in \mathcal{D}_{y_a} \setminus \{x_a\}}{\text{Top-}K} \left(\frac{\text{Sim}(e^i_{x_{pos}}, e^i_{x_a}) + \text{Sim}(e^t_{x_{pos}}, e^t_{x_a})}{2} \right) \tag{3}$$

Finally, the anchor x_a, the verified hard positives $\mathcal{H}_{pos}$, and the hard negatives $\mathcal{H}_{neg}$ are concatenated into a unified contrastive group. By feeding this assembled context into the MLLM for joint supervised training, we explicitly implement the cross-image attention mechanisms. This paradigm forces the model to perform fine-grained visual comparisons within the context, comprehensively bridging the attention across instances to overcome its previous semantic bottlenecks.

3.3 Progressive Anti-Shortcut Training

In practice, we found that training the MLLM only with a fixed multi-image format can introduce undesirable shortcuts (See Sec. 4.3). Specifically, the model tends to generate multiple semantically similar answers in a single response, thereby increasing the chance that the correct answer is included without truly resolving the fine-grained distinction. Moreover, excessive exposure to multi-image instructions weakens the model's instruction-following ability. To address these issues, we propose a progressive anti-shortcut training strategy that combines order-agnostic multi-image contrastive learning with mixed-format supervision.

Given an anchor image x_a, we first collect the hard positives and hard negatives mined by the C2M module. To prevent the model from relying on fixed positions or formatting patterns, we merge these samples into a unified context pool and randomly shuffle their order:

$$X_{\text{context}} = \text{Shuffle}(\mathcal{H}_{\text{pos}} \cup \mathcal{H}_{\text{neg}}). \tag{4}$$

The shuffled context images are then concatenated with the anchor image in the multimodal prompt. The model is optimized with the cross-entropy loss:

$$\mathcal{L}_{\text{contrast}} = \mathbb{E}\left[\mathcal{L}_{\text{CE}}\left(\theta; [x_a, X_{\text{context}}], [y_a, Y_{\text{context}}]\right)\right], \tag{5}$$

where Y_{context} denotes the ground-truth labels corresponding to the contextual images. Since the positions of hard positives and negatives are randomized, the model cannot infer the answer from image order or template structure, and is instead encouraged to compare fine-grained visual evidence.

To further preserve the model's single-image recognition and instruction-following capabilities, we adopt a mixed-format training objective. Each training batch contains a proportion ρ of standard single-image samples and $1-\rho$ of mined multi-image contrastive samples. The overall objective is:

$$\mathcal{L}_{\text{total}} = \rho\, \mathbb{E}_{(x,y)\sim\mathcal{B}_{\text{single}}}\left[\mathcal{L}_{\text{CE}}(\theta; x, y)\right] + (1-\rho)\, \mathbb{E}_{(X,Y)\sim\mathcal{B}_{\text{multi}}}\left[\mathcal{L}_{\text{contrast}}\right]. \tag{6}$$

In practice, we set $\rho = 0.3$ to balance the instruction following ability and the multi-image contrastive objective. By combining shuffled hard exemplars with mixed-format supervision, the proposed training strategy encourages discriminative fine-grained reasoning while maintaining robust general instruction-following behavior.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate our proposed C2FT across two primary categories requiring fine-grained perception: Fine-Grained Visual Classification (FGVC) and Visual Question Answering (VQA). For FGVC, we utilize FGVC-Aircraft [74], Oxford-IIIT Pet (Pet-37) [75], Stanford Cars (Car-196) [76], and the Google Landmarks Dataset [77]. For VQA tasks, which contain a more diverse query and answer format, we employ TallyQA [78] and GQA [79]. TallyQA comprises 287K complex counting questions across 165K images with numerical answers up to 15, and GQA is a dataset consisting of 22M challenging compositional questions over real-world images that evaluates spatial understanding, inherently demanding strong fine-grained visual perception capabilities from the models.

Baselines and metrics. We evaluate our method against two groups of baselines. First, we compare with leading open-source MLLMs of varying scales [80, 81], including the InternVL [82] series and Qwen2.5-VL [4] series. Second, to perform a rigorous internal ablation, we apply C2FT to foundational models (Qwen3-VL-4B and Gemma-4-E2B [83]) and compare the results against their Base, SFT, and GRPO counterparts. Following standard evaluation protocols, we report the Top-1 Accuracy for all individual datasets and calculate the overall average score.

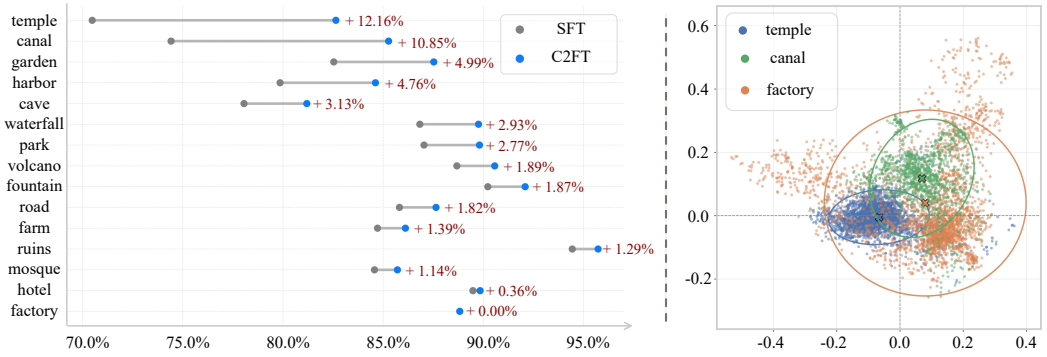
Implementation details. We utilize Qwen3-VL-4B and Gemma-4-E2B as our base models. For all of our experiments, we adopted the same learning rate as the baseline, utilized the AdamW optimizer, and trained the models until convergence. All models are trained using full-parameter fine-tuning, and the maximum sequence length is capped at 16384 tokens.

4.2 Main Results

Quantitative analysis. In this section, we train and evaluate our method across FGVC and VQA datasets, with results summarized in Tab. 1. C2FT achieves the best overall average score on

Table 1: **Quantitative comparisons on FGVC and VQA benchmarks.** We report the quantitative results on both FGVC and VQA datasets. The best results are **bolded** and the second best are underlined. Aircraft: FGVC-Aircraft dataset. GLD: Google Landmarks Dataset.

Models	FGVC				VQA		AVG
	Aircraft	Pet-37	Car-196	GLD	TallyQA	GQA	
Comparison with other MLLMs							
LLaVA-Onevision-7B	31.51	51.23	46.77	34.84	42.76	48.95	42.68
Idefics3-LLAMA3-8B	32.76	55.82	50.49	37.81	45.63	51.74	45.71
InternVL2.5-4B	29.45	43.16	51.04	33.42	39.70	46.93	40.62
InternVL2.5-8B	34.96	56.32	53.17	38.90	43.72	51.58	46.44
Qwen2.5-VL-3B	41.48	56.92	57.69	42.83	49.62	54.87	50.57
Qwen2.5-VL-7B	44.94	59.85	60.77	48.53	54.50	58.31	54.48
Comparison among Our Models							
DeepPerception-7B	44.24	40.90	54.14	49.35	-	-	47.15
No-Thinking-RFT	74.41	86.13	92.5	64.87	-	-	79.48
Qwen3-VL-4B	47.55	59.70	63.92	53.54	59.52	61.64	57.65
+ SFT	84.52	**92.23**	85.72	82.31	63.96	65.74	79.08
+ GRPO	85.13	89.64	86.96	81.60	63.42	64.83	78.60
+ C2FT	**89.43**	91.67	**93.27**	**84.72**	**68.45**	**70.16**	**82.95**
Gemma-4-E2B	21.18	33.92	32.75	28.64	36.81	38.73	32.01
+ SFT	37.39	62.68	59.86	66.72	45.25	44.81	52.79
+ GRPO	38.41	60.10	59.72	64.31	46.92	42.27	51.95
+ C2FT	**40.68**	**65.72**	**61.59**	**68.73**	**48.43**	**47.91**	**55.51**



(a) Accuracy comparison on subcategories

(b) Feature distributions

Figure 3: **Per category analysis on the Google Landmarks dataset.** (a) Accuracy gains of C2FT over the SFT baseline across categories. (b) PCA visualization of image embeddings. The results demonstrate that classes with more densely clustered and visually entangled feature spaces can yield larger performance improvements.

both Qwen3-VL-4B and Gemma-4-E2B, raising the score from 57.65 to 82.95 and 32.01 to 55.51, respectively. It also surpasses recent fine-grained perception methods such as DeepPerception-7B [84] and No-Thinking-RFT [35]. Although No-Thinking-RFT shows competitive performance on several FGVC datasets, C2FT further improves the results on Aircraft, Car-196, and GLD, demonstrating its effectiveness in both fine-grained visual categorization and visual question answering.

Compared with different fine-tuning strategies, C2FT also shows clear advantages. On Qwen3-VL-4B, it surpasses standard SFT and GRPO by 3.87 and 4.35 average points, respectively. Similar trends are observed on Gemma-4-E2B, where C2FT outperforms both SFT and GRPO across all

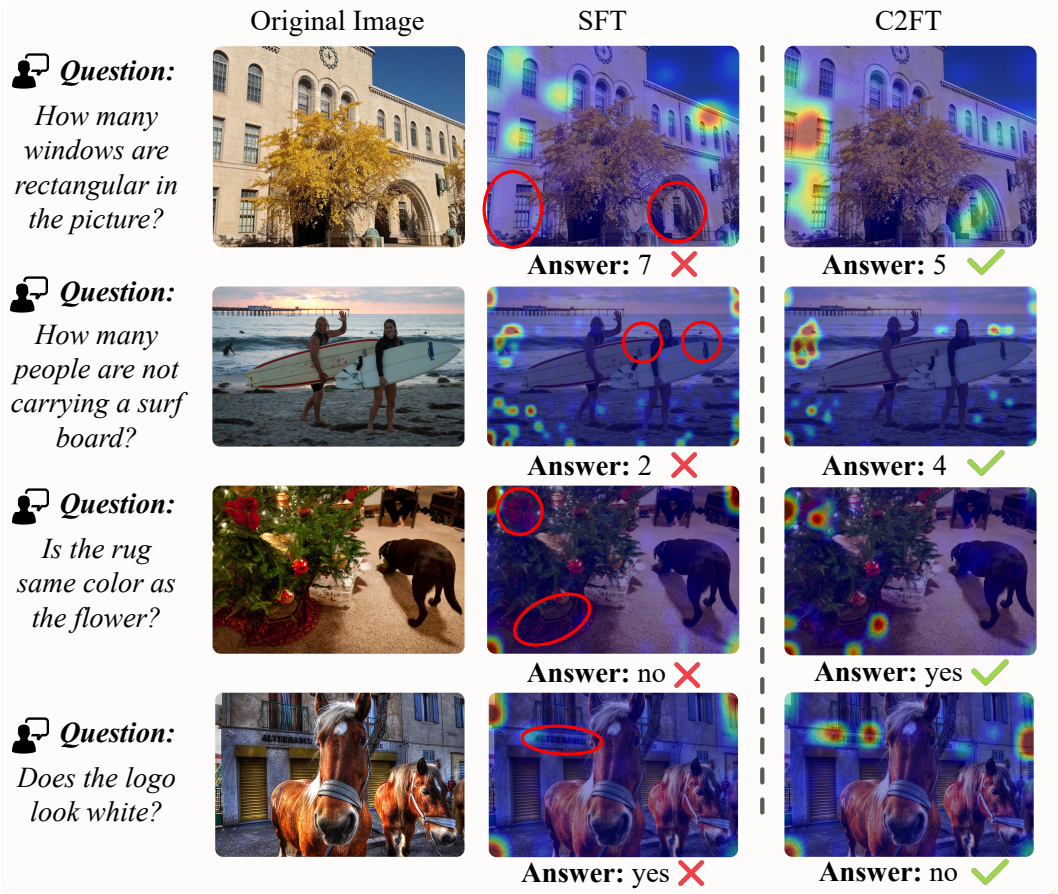


Figure 4: **Qualitative results on TallyQA and GQA.** Attention maps for the generated answer tokens are visualized, with red circles highlighting the critical fine-grained features. C2FT accurately localizes subtle discriminative regions, while the SFT baseline struggles with background distractions, resulting in incorrect predictions.

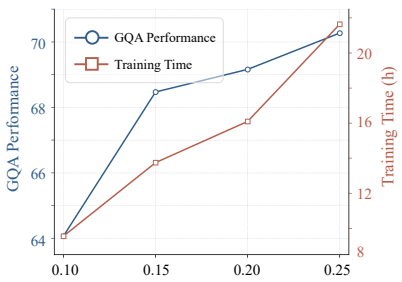
benchmarks. These consistent gains across different backbones suggest that introducing input-side comparison among confusing samples helps MLLMs learn more discriminative fine-grained visual cues than isolated single-sample optimization.

Qualitative analysis. Fig. 4 compares the attention maps of the standard SFT baseline and C2FT. The SFT model often attends to irrelevant regions or distractors, leading to incorrect answers in fine-grained counting and attribute comparison tasks. For example, it focuses on invalid windows, people carrying surfboards, or background areas unrelated to the queried objects. In contrast, C2FT concentrates on the discriminative regions required by the question and produces correct predictions. These results indicate that training with hard contrastive examples helps the model suppress spurious cues and improve fine-grained visual perception.

In-depth analysis on GLD. To further investigate how C2FT enhances fine-grained perception, we conduct a class-level analysis on selected subcategories from the GLD in Fig. 3. By correlating performance metrics with feature distributions, we observe a distinct pattern: C2FT yields the most substantial improvements on categories with densely clustered feature spaces. As shown in Fig. 3(b), the embeddings for *temple* and *canal* are heavily entangled, indicating high visual similarity that typically confuses standard MLLMs. Consequently, C2FT achieves its highest gains on these categories (+12.16% and +10.85%, respectively). Conversely, the *factory* category exhibits a widely dispersed feature distribution with less intra-class consistency, resulting in negligible improvement (+0.00%). This phenomenon validates our core motivation. The confuse-then-contrast mining strategy actively isolates visually ambiguous hard negatives within dense clusters, forcing the model to learn crucial discriminative details exactly where foundational models struggle the most.

(a) Ablation on core training strategies. "w/o C2M" denotes replacing C2M with randomized contrastive training; "w/o PAT" removes the PAT strategy directly.

Method	GLD	GQA
w/o C2M	81.26	65.07
w/o PAT	72.64	51.42
Ours	84.72	70.16



(b) Time and effect trade-off across τ.

(c) Ablation on threshold τ. $\tau = 0.20$ achieves the best trade-off.

τ	GLD	GQA
0.10	82.73	66.92
0.15	83.82	69.47
0.20	84.72	70.16
0.25	83.92	71.27

Figure 5: **Ablation results.** (a) Ablation on core training strategies demonstrates the necessity of the C2M and PAT modules. (b) The trade-off between performance and training time across different τ values. (c) Sensitivity analysis of the threshold τ.

4.3 Ablation Study

Effect of core training strategies. To validate the effectiveness of our proposed components, we ablate the C2M and PAT modules, as shown in Fig. 5(a). Replacing C2M with randomized contrastive training causes a notable performance drop of over 3 and 5 points on GLD and GQA, respectively. Meanwhile, omitting the PAT strategy leads to even more severe degradation across both metrics, underscoring its crucial role in maintaining the model's instruction-following behavior. These results confirm that both training strategies are essential and complementary for maximizing the model's overall performance.

Effect of threshold τ. We investigate the impact of the threshold τ, which controls the strictness of our hard negative mining. According to Eq. (2), increasing τ relaxes the margin constraint, thereby allowing a larger quantity of candidate samples to be retrieved into the hard negative set $\mathcal{H}_{neg}$.

As shown in Fig. 5(c), performance generally improves as τ increases, suggesting that a larger pool of hard negatives is beneficial. For example, GQA accuracy steadily increases from 66.92 to 71.27 as τ grows from 0.1 to 0.25. GLD shows a similar upward trend up to $\tau = 0.2$, improving from 82.73 to 84.72, although it slightly drops at $\tau = 0.25$; we discuss this anomaly in Sec. 5. However, Fig. 5(b) reveals a critical trade-off between performance and computational cost. Expanding the negative set drastically increases training time, with a particularly steep surge when moving from τ=0.20 to 0.25. Consequently, we adopt $\tau = 0.20$ as it achieves the optimal balance between effectiveness and efficiency. This scaling trend validates the effectiveness of C2FT, demonstrating that exposing the model to more visually ambiguous samples strengthens fine-grained perception.

5 Conclusion

In this paper, we introduce C2FT, a Confuse-then-Contrast Fine-Tuning framework designed to enhance the fine-grained visual perception of MLLMs. By dynamically assembling hard positive and negative samples into unified multi-image instructions, C2FT enforces explicit cross-image comparisons via joint-attention. Complemented by a progressive anti-shortcut training strategy, our approach guides models to capture subtle discriminative visual cues while preserving general instruction-following abilities. Extensive experiments demonstrate that C2FT achieves superior performance on both FGVC and VQA benchmarks, representing a practical approach for empowering MLLMs with fine-grained perception ability.

Limitation and future work. While C2FT effectively leverages contrastive samples, our ablation study indicates that incorporating an excessive number of hard negatives can occasionally introduce context noise, which may slightly degrade performance on certain datasets like GLD. Future work could explore combining C2FT with recent data synthesis frameworks [71], which aims to deliver a fully automated training and self-improving pipeline for fine-grained multimodal perception.

Acknowledgments and Disclosure of Funding

This work was supported by Alibaba Group Security and Alibaba Group through the Alibaba Research Intern Program. The authors declare no other competing interests.

References

[1] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[2] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[3] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025.

[4] 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.

[5] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[6] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[7] Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words. *arXiv preprint arXiv:2407.06581*, 2024.

[8] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? *Advances in Neural Information Processing Systems*, 37:51727–51753, 2024.

[9] Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

[10] Gregor Geigle, Radu Timofte, and Goran Glavaš. African or european swallow? benchmarking large vision-language models for fine-grained object classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2653–2669, 2024.

[11] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025.

[12] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024.

[13] Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–632, 2025.

[14] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19446–19455, 2025.

[15] 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.

[16] Yangyang Shu, Anton Van den Hengel, and Lingqiao Liu. Learning common rationale to improve self-supervised representation for fine-grained visual recognition problems. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11392–11401, 2023.

[17] Eran Hirsch, Aviv Slobodkin, David Wan, Elias Stengel-Eskin, Mohit Bansal, and Ido Dagan. Laquer: Localized attribution queries in content-grounded generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15355–15370, 2025.

[18] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. 2024 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13084–13094, 2023.

[19] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023.

[20] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9568–9578, 2024.

[21] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, and Tae-Hyun Oh. Beaf: Observing before-after changes to evaluate hallucination in vision-language models. In *European Conference on Computer Vision*, pages 232–248. Springer, 2024.

[22] Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuan-Jing Huang, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3769–3798, 2025.

[23] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, 2025.

[24] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[27] 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.

[28] Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. *arXiv preprint arXiv:2502.14044*, 2025.

[29] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Xinzhe Ni, Zicheng Lin, Songtao Jiang, Yiyao Yu, Chufan Shi, Lei Wang, Ruihang Chu, et al. Unlocking multimodal mathematical reasoning via process reward model. *arXiv preprint arXiv:2501.04686*, 2025.

[30] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025.

[31] Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24290–24315, 2025.

[32] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[33] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[34] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

[35] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. To think or not to think: A study of thinking in rule-based visual reinforcement fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[36] Benlei Cui, Ruize Wang, Junjie Li, Jinhao Chen, Longtao Huang, Yinghao Chen, Yuwen Zhai, Jingqun Tang, Ruijian Jia, Weiwei Wu, et al. Metavideoagent: Automated video-agent evolution for long-form video understanding. *arXiv preprint arXiv:2608.04587*, 2026.

[37] Yefei He, Yuanyu He, Shaoxuan He, Feng Chen, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. Neighboring autoregressive modeling for efficient visual generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19000–19010. IEEE, 2025.

[38] Yuanyu He, Zhuokun Chen, Yefei He, Zhiwei Tang, Jiasheng Tang, Yinghao Yu, Jianfei Cai, and Bohan Zhuang. Dynamics-aware sparse attention for efficient autoregressive video diffusion. *arXiv preprint arXiv:2610.26262*, 2026.

[39] Shaoxuan He, Benlei Cui, Bukun Huang, Zhizeng Ye, Yunyun Sun, Longtao Huang, Hui Xue, Yang Yang, Haiwen Hong, Jingqun Tang, et al. Tc-pade: Trajectory-consistent pade approximation for diffusion acceleration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 35768–35778, 2026.

[40] Shikai Qiu, Xiaowen Xu, Benlei Cui, Ting Ma, Xiufeng Huang, Wenjing Jiang, Shaoxuan He, Haolei Xu, Chunyang Chai, Yujian Li, et al. Yuvion vl: A multimodal foundation model for adversarial content and ai safety. *arXiv preprint arXiv:2606.25034*, 2026.

[41] Yuxin Peng, Zishuo Wang, Xiangtian Zheng, Geng Li, Sibo Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. 2025.

[42] Anwesha Mohanty, Venkatesh Balavadhani Parthasarathy, and Arsalan Shahid. The future of mllm prompting is adaptive: A comprehensive experimental evaluation of prompt engineering methods for robust multimodal performance. *arXiv preprint arXiv:2504.10179*, 2025.

[43] Ranran Huang, Yu Wang, and Huazhong Yang. Cross-layer attention network for fine-grained visual categorization. *arXiv preprint arXiv:2210.08784*, 2022.

[44] Huazhen Chen, Haimiao Zhang, Chang Liu, Jianpeng An, Zhongke Gao, and Jun Qiu. Fet-fgvc: Feature-enhanced transformer for fine-grained visual classification. *Pattern Recognition*, 149:110265, 2024.

[45] Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601, 2024.

[46] Gabriel Sarch, Lawrence Jang, Michael J Tarr, William W Cohen, Kenneth Marino, and Katerina Fragkiadaki. Vlm agents generate their own memories: Distilling experience into embodied programs of thought. *Advances in Neural Information Processing Systems*, 37:75942–75985, 2024.

[47] Renrong Shao, Wei Zhang, Jianhua Yin, and Jun Wang. Data-free knowledge distillation for fine-grained visual categorization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1515–1525, 2023.

[48] Jinseong Jang, Chunfei Ma, and Byeongwon Lee. Vl2lite: Task-specific knowledge distillation from large vision-language models to lightweight networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 30073–30083, 2025.

[49] Jhe-Hao Lin, Yi Yao, Chan-Feng Hsu, Hong-Xia Xie, Hong-Han Shuai, and Wen-Huang Cheng. Perspective-aware teaching: Adapting knowledge for heterogeneous distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4178–4187, 2025.

[50] Jie Zhu, Yiyang Su, and Xiaoming Liu. Can textual reasoning improve the performance of mllms on fine-grained visual classification? *arXiv preprint arXiv:2601.06993*, 2026.

[51] Xuzheng Yang, Junzhuo Liu, Peng Wang, Guoqing Wang, Yang Yang, and Heng Tao Shen. New dataset and methods for fine-grained compositional referring expression comprehension via specialist-mllm collaboration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[52] Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin Peng. Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 14628–14637, 2024.

[53] Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-r1: Make multi-modal llms excel in fine-grained visual recognition by chain-of-thought reasoning. *arXiv preprint arXiv:2602.07605*, 2026.

[54] Hao Ding, Zhichuan Yang, Weijie Ge, Ziqin Gao, Chaoyi Lu, and Lei Zhao. Tikart: Stabilizing aperture-guided fine-grained visual reasoning with reinforcement learning. *arXiv preprint arXiv:2602.14482*, 2026.

[55] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044, 2025.

[56] Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

[57] Jiayi Fu, Xuandong Zhao, Chengyuan Yao, Heng Wang, Qi Han, and Yanghua Xiao. Reward shaping to mitigate reward hacking in rlhf. *arXiv preprint arXiv:2502.18770*, 2025.

[58] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[59] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

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

[61] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[62] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pages 776–794. Springer, 2020.

[63] Lei Ding, Jing Zhang, Haitao Guo, Kai Zhang, Bing Liu, and Lorenzo Bruzzone. Joint spatio-temporal modeling for semantic change detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[64] Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3108–3116, 2022.

[65] Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing differences in image sets with natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24199–24208, 2024.

[66] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15671–15680, 2022.

[67] 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.

[68] Wei Lin, Muhammad Jehanzeb Mirza, Sivan Doveh, Rogerio Feris, Raja Giryes, Sepp Hochreiter, and Leonid Karlinsky. Comparison visual instruction tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2973–2983, 2025.

[69] Zhiyu Pan, Yizheng Wu, Jiashen Hua, Junyi Feng, Shaotian Yan, Bing Deng, Zhiguo Cao, and Jieping Ye. Through the lens of contrast: Self-improving visual reasoning in vlms. *arXiv preprint arXiv:2603.02556*, 2026.

[70] Ji Ma, Wei Suo, Peng Wang, and Yanning Zhang. C3l: Content correlated vision-language instruction tuning data generation via contrastive learning. *arXiv preprint arXiv:2405.12752*, 2024.

[71] Qirui Jiao, Daoyuan Chen, Yilun Huang, Bolin Ding, Yaliang Li, and Ying Shen. Img-diff: Contrastive data synthesis for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9307, 2025.

[72] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[73] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, and Maxime Oquab. Cijo jose, vasil khalidov, marc szafraniec, seungeun yi, michaël ramamonjisoa, et al. dinov3. *arXiv preprint arXiv:2508.10104*, 2(4):5, 2025.

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

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

[76] 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*, pages 554–561, 2013.

[77] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.

[78] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084, 2019.

[79] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[80] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[81] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023.

[82] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[83] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[84] Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. Deepperception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*, 2025.

A More Implementation Details

C2M module. For the offline feature initialization, we employ a frozen, pre-trained vision-language encoder (e.g., Qwen3 Embedding [72], Dinov3 [73]) to extract unified visual and textual embeddings for all dataset instances and ground-truth categories. This offline bank is computed only once prior to training. During the dynamic semantic confusion estimation step, we use standard cosine similarity to measure the feature-level distance between the model's generated incorrect predictions and the ground-truth classes. Based on our ablation studies, we set the Teacher Forcing verification threshold $\tau = 0.20$, which optimally balances the mining of challenging hard negatives with computational overhead. For hard positive mining, we consistently retrieve the Top-$K(K = 2)$ instances that maximize the joint image-text similarity.

Algorithm 1 Confuse-then-Contrast Mining (C2M)

Require: Training dataset $\mathcal{D}$, current MLLM with weights θ, frozen pre-trained encoder E, threshold τ, Top-K value K.
Ensure: Multi-image contrastive dataset $\mathcal{B}_{multi}$.

Offline Initialization: Extract class text embeddings $\mathbf{e}_c^{gt}$ for all classes $c \in \mathcal{C}$.
Extract image/text embeddings $(\mathbf{e}_x^i, \mathbf{e}_x^t)$ for all $x \in \mathcal{D}$ using E.
Initialize multi-image contrastive dataset $\mathcal{B}_{multi} \leftarrow \emptyset$.
for each anchor sample $x_a \in \mathcal{D}$ with ground-truth label y_a **do**
  Generate prediction $pred$ using the current MLLM.
  if $pred \neq y_a$ **then**
    % Model-Centric Confusion Estimation
    Extract text embedding $\mathbf{e}_{pred}$ for the incorrect prediction.
    Compute confusion score $S_{confuse}(c \mid x_a) = \mathrm{Sim}(\mathbf{e}_{pred}, \mathbf{e}_c^{gt})$.
    Obtain semantic confusion set $\mathcal{C}_{confuse}$ by sorting $S_{confuse}$ in descending order.
    % Fine-Grained Visual Contrastive Retrieval
    Initialize hard negative set $H_{neg} \leftarrow \emptyset$.
    for each candidate $x_{neg} \in \mathcal{D}_c$ where $c \in \mathcal{C}_{confuse}$ **do**
      if $|P_{TF}(y_a \mid x_{neg}) - P_{TF}(c \mid x_{neg})| < \tau$ **then**
        $H_{neg} \leftarrow H_{neg} \cup \{x_{neg}\}$.
      end if
    end for
    % Mine Hard Positives via multi-modal feature similarity
    Select $H_{pos} \leftarrow \text{Top-}K_{x_{pos} \in \mathcal{D}_{y_a} \setminus \{x_a\}} \left(\frac{\mathrm{Sim}(\mathbf{e}^i_{x_{pos}}, \mathbf{e}^i_{x_a}) + \mathrm{Sim}(\mathbf{e}^t_{x_{pos}}, \mathbf{e}^t_{x_a})}{2} \right)$.
    % Assemble Context
    $\mathcal{X}_{context} \leftarrow \mathrm{Shuffle}(H_{pos} \cup H_{neg})$.
    $\mathcal{B}_{multi} \leftarrow \mathcal{B}_{multi} \cup \{[x_a, \mathcal{X}_{context}]\}$.
  end if
end for
return $\mathcal{B}_{multi}$.

PAT module. To mitigate positional biases and prevent the model from learning unintended sequence shortcuts, the retrieved hard positives and negatives are randomly shuffled before being concatenated into the multimodal prompt. To preserve the base model's general instruction-following capabilities and mitigate catastrophic forgetting, we employ a mixed-format batching strategy. Each training batch is constructed to contain a fixed proportion of single-image samples and multi-image contrastive samples. Empirically, we set the single-image sampling ratio to $\rho = 0.3$. This configuration effectively forces the model to engage in explicit cross-image discriminative reasoning while retaining standard conversational robustness.

B Limitation and Societal Impacts

Limitation analysis. As discussed in Sec. 5 and our ablation study in Sec. 4.3, while C2FT effectively leverages hard contrastive samples to enhance fine-grained perception, scaling this multi-image

Algorithm 2 Progressive Anti-Shortcut Training (PAT)

Require: Training dataset $\mathcal{D}$, multi-image contrastive dataset $\mathcal{B}_{multi}$, MLLM with weights θ, mixing ratio ρ.
Ensure: Updated MLLM weights θ.
for each training epoch **do**
  for each training step **do**
    Sample a mini-batch containing a fraction ρ of standard single-image instances from $\mathcal{D}$.
    Sample a fraction $(1-\rho)$ of multi-image instances from $\mathcal{B}_{multi}$.
    Compute the total training loss $\mathcal{L}_{total} = \rho\,\mathbb{E}[\mathcal{L}_{CE_{single}}] + (1-\rho)\,\mathbb{E}[\mathcal{L}_{contrast}]$.
    Update model weights θ by minimizing $\mathcal{L}_{total}$.
  end for
end for
return θ.

context presents two primary limitations. First, an excessive number of hard negatives can introduce overwhelming context noise. As observed in our sensitivity analysis of the threshold ω, relaxing the margin constraint ($\omega = 0.25$) expands the negative set H_{neg} but causes a slight performance degradation on certain datasets, such as GLD. This anomaly indicates that beyond a certain point, cramming too many visually ambiguous distractors into a single prompt may confound the model's cross-image attention mechanism rather than refine it. Second, there is a strict trade-off between perception performance and computational efficiency. Expanding the contrastive pool drastically increases the multimodal input sequence length, leading to a steep surge in training time and memory footprint (as demonstrated in Fig. 5(b)). Consequently, boundless scaling of the joint-supervision context is currently computationally impractical. Future work will explore integrating C2FT with advanced data synthesis and filtering frameworks [70, 71] to dynamically assess and control the quality of the contrastive context, thereby minimizing noise and optimizing computational overhead in a fully automated, self-improving pipeline.

Potential positive societal impacts. Beyond advancing academic benchmarks, the fine-grained perception capabilities introduced by C2FT hold significant potential for high-stakes, real-world applications. By empowering models to accurately discriminate subtle visual differences and reduce hallucinations, our framework can be broadly applied to domains requiring precise comparative reasoning. For instance, in security, it can assist in detecting deepfakes and subtle anomalies; in law, it can aid in analyzing nuanced forensic evidence or trademark disputes; and in politics, it can support rigorous fact-checking against manipulated visual media. Ultimately, C2FT contributes to the development of more accountable, precise, and trustworthy AI systems for detail-sensitive tasks.

Potential negative societal impacts. C2FT significantly enhances the capacity of multimodal models to distinguish subtle visual nuances and reduce hallucinations. While this precise fine-grained perception positions C2FT as a powerful tool for complex visual analysis, it is important to acknowledge that its reliability can still be compromised by highly crafted adversarial examples or excessive visual noise. Consequently, in high-stakes domains such as medical diagnosis or automated biometric identification, over-reliance on the model's predictions could lead to severe misjudgments. Therefore, the deployment of C2FT in these critical areas necessitates human-in-the-loop verification and robust fail-safe mechanisms to mitigate potential societal risks.

C More Qualitative Results

In this section, we provide more qualitative results to further illustrate the effectiveness of C2FT. Through various visual examples, we showcase its superior ability to captures fine-grained details in complex visual contexts.

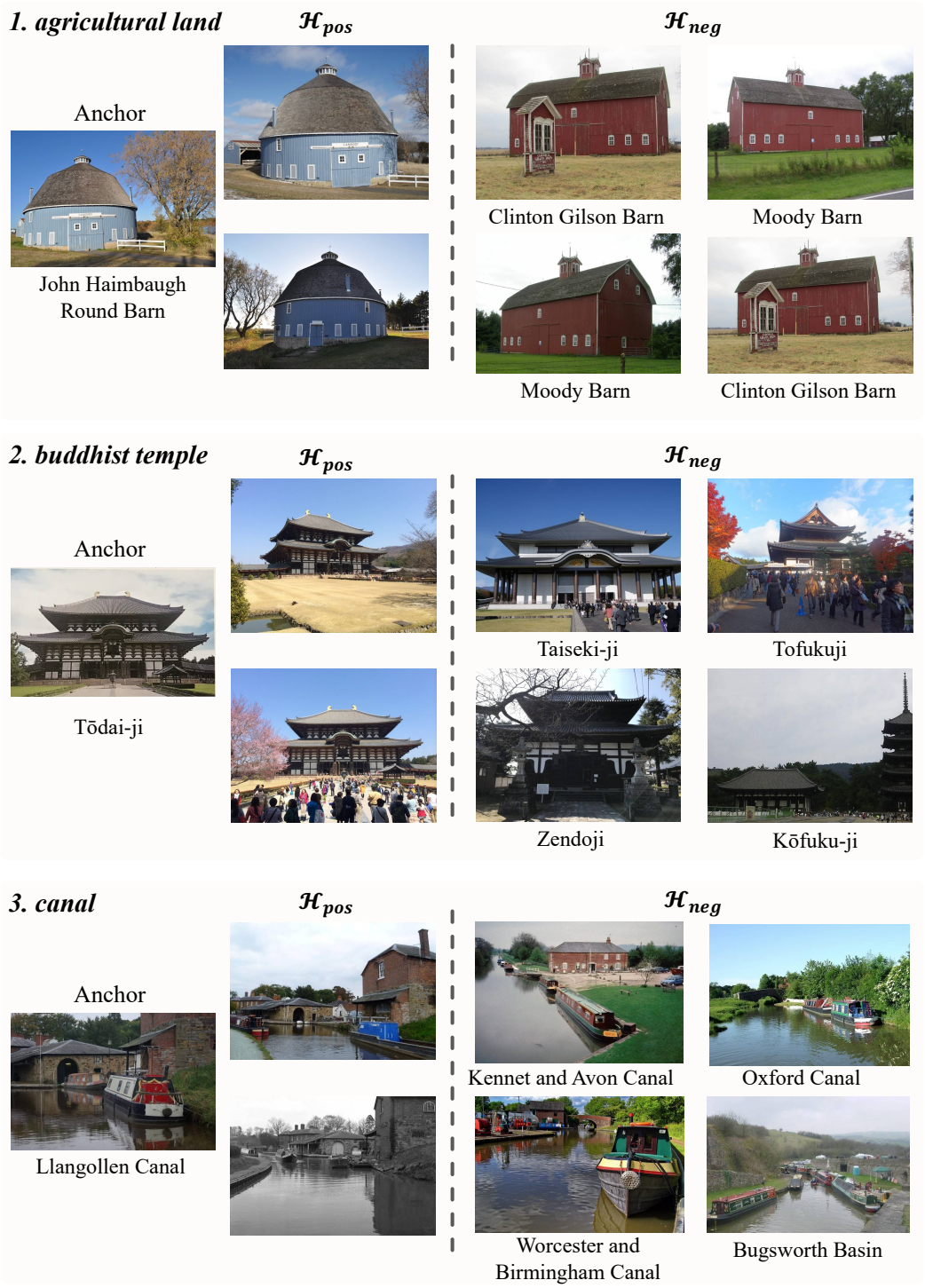


Figure A: **Hard positives and negatives sample on GLD [77]**. Here we showcase 1 anchor sample with 2 postives and 4 negatives.

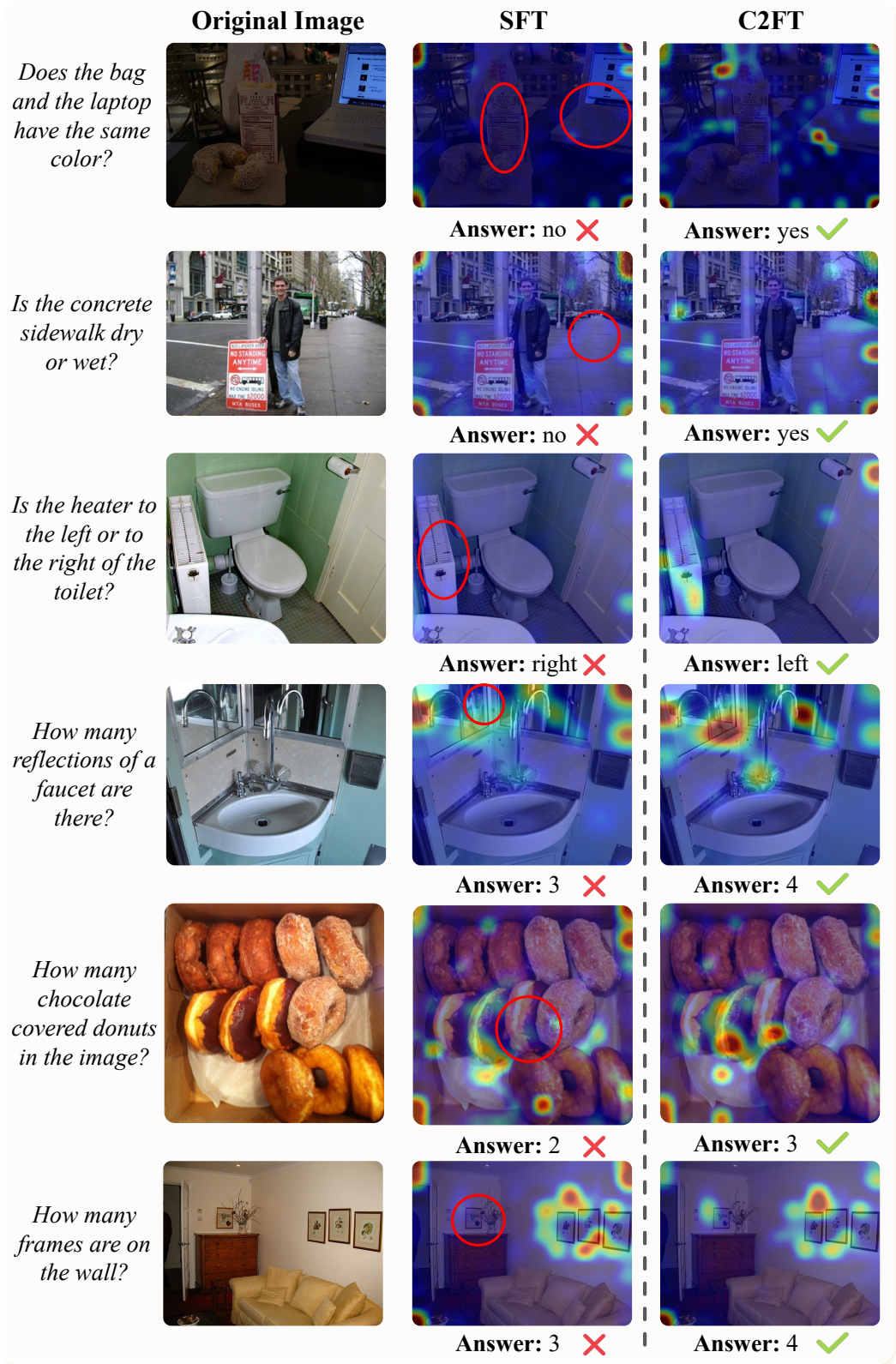


Figure B: **More Qualitative results.** Examples are randomly sampleed from TallyQA and GQA datasets.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly summarize our contribution in the abstract and introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We claim our limitation in conclusion.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Our method is detailed introduced in the paper.

   Guidelines:

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

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

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

Answer: [No]

Justification: Currently, we can not provide the code, but we will provide it after acceptance.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Our method is detailed introduced in the paper.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: We do not provide the error bar, but all experiments are done in fair settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We include the GPU resources we used in paper.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This work fully meets the demand of NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We claim the broader impact in appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite all paper related to the code base, baseline, compared method, and dataset.

    Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset or benchmark asset. If code or model checkpoints are released, they should be documented with training details, licenses, intended use, and limitations.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: Our work does not involve any human subjects, user studies, or personally identifiable information, and therefore does not require IRB approval.

    Guidelines:

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

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

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