MirrorCoT: Lightweight Multimodal Interleaved Chain-of-Thought

Anonymous ACL submission

Abstract

Recent advances in multimodal interleaved Chain-of-Thought(CoT) (Li et al., 2025) have exhibited great potential in boosting the reasoning performance of Multimodal Large Language Models (MLLMs). However, existing work rarely examines *how* and *when* visual information should be injected during multimodal reasoning. In this work, we systematically study **MirrorCoT**, a lightweight multimodal interleaved CoT with **query-triggered visual injection** mechanism. We conduct comprehensive comparisons against state-of-the-art baselines (e.g., VoCoT, VQD) on LLaVA-1.5 and InternVL-2 across seven benchmarks, including MMStar and HallusionBench. Our key finding is that a simple structural modification, which **forces the model to explicitly emit a sub-question that triggers a visual information injection**, consistently outperforms dense visual token insertion. This modification not only boosts task accuracy (e.g., delivering a 5.1% improvement on LLaVA-1.5) but also cuts the number of visual tokens required by 91.1%. Further analysis reveals that the gains stem from **Dynamic Inquiry** (deciding *when* to look) and **Targeted Feature Extraction** (deciding *what* to retrieve), which also mitigates hallucination in long-context generation.

1 Introduction

Chain-of-Thought (CoT) prompting has unlocked substantial reasoning gains in large language models (LLMs) by encouraging explicit intermediate steps (Kojima et al., 2022; Wei et al., 2022; Yao et al., 2023). This success naturally spurred the exploration of multimodal CoT, aiming to extend step-by-step reasoning to vision-language settings (Zhang et al., 2023b, 2024b; Xu et al., 2024; Xiang et al., 2024). However, long-horizon multimodal generation exposes a mechanistic flaw: a widening gap between *producing* a fluent rationale and *grounding* it in visual reality. Empirical studies

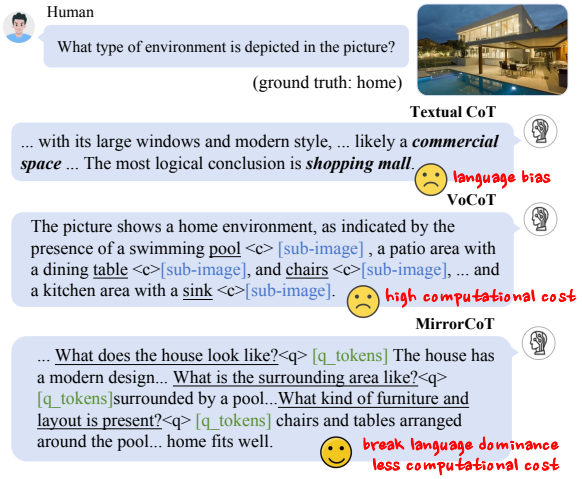


Figure 1: Comparison of multimodal reasoning paradigms. **Textual CoT** (Wei et al., 2022) solely relies on language priors, often leading to hallucinations. **Visual CoT** (e.g., VoCoT (Li et al., 2025)) inserts raw image tokens [sub-image], causing high computational redundancy. In contrast, our **MirrorCoT** actively generates sub-questions to extract query-specific visual features into learnable tokens [q_tokens].

reveal that MLLMs often suffer from "textual inertia": as the reasoning trace grows, the model becomes increasingly conditioned on its own generated text, causing attention to drift away from the image and leading to hallucinations that are logically plausible but visually ungrounded (Huang et al., 2024; Leng et al., 2024; Kang et al., 2025). This suggests that the effective influence of visual evidence systematically decays as auto-regressive decoding progresses.

A common belief is that grounding can be sustained by injecting more visual information into the reasoning stream. Recent systems thus attempt to strengthen the "vision pathway" via strategies such as delegating decomposition to external LLMs (Zheng et al., 2023; Surís et al., 2023) or reallocating inference-time attention from text to image (Liu et al., 2024a). More recently, VoCoT (Li et al.,

2025) proposed interleaving dense, object-centric image tokens directly into the output sequence. While effective, such approaches introduce substantial redundancy. As illustrated in Figure 1, VoCoT mechanically inserts raw image patches [sub-image] whenever an object is mentioned, flooding the context with visual tokens that may not be relevant to the *current* reasoning sub-goal. This passive injection increases compute overhead without ensuring attention to verification-critical features.

A motivating finding: Informativeness outweighs quantity in visual tokens. We posit that the core challenge of long-horizon CoT is not insufficient visual grounding capability, but the lack of a mechanism for extracting task-relevant compact features. We prompt the model to raise sub-questions to inquire about visual evidence that supports the reasoning. Our observations indicate that performance gains are still achievable even when the volume of injected visual tokens is drastically reduced, provided that these tokens are *contextually tailored* to the current reasoning step. This points to a first-principles diagnosis: effective grounding requires a *minimal-yet-efficient* visual evidence insertion mechanism that interrupts text dominance and amplifies visual influence exactly when needed.

MirrorCoT: target visual evidence acquisition under a token budget. Motivated by this view, we propose **MirrorCoT**, a paradigm that casts multimodal reasoning as *target visual evidence acquisition*. Instead of passively receiving a stream of visual tokens, the model learns to (i) explicitly emit a sub-question when textual priors are insufficient, and (ii) actively retrieve only the visual content needed to verify that step. Concretely, MirrorCoT interleaves reasoning with a lightweight vision operator. Conditioned on a generated query (e.g., `<IMAGE_QUERY>...`), this operator extracts *query-specific* features into a small set of learnable tokens [q_tokens], injecting them back into the stream (see Figure 1). This design establishes a bidirectional loop: (1) **Vision-to-Language Alignment**: Visual tokens are distilled strictly for the current sub-goal (e.g., checking the specific attributes of a "couch" rather than just seeing generic furniture features), filtering out irrelevant noise. (2) **Language-to-Vision Alignment**: The active injection of these evidence tokens breaks the language-dominant inertia, forcing the model to re-anchor its reasoning to the image without the cost of re-encoding full visual features.

Our main contributions are summarized as follows:

- We conceptualize multimodal CoT grounding as an **effective visual evidence acquisition** problem and refine the visual information injection manner in multimodal interleaved Chain-of-Thought.
- We propose **MirrorCoT**, a framework that employs a dynamic, query-triggered visual evidence operator to inject *step-specific* visual tokens, significantly reducing redundancy compared to dense insertion baselines while improving interpretability.
- We develop an automated data construction pipeline to distill strong teacher reasoning into MirrorCoT-style traces. Experiments show consistent improvements, boosting LLaVA-1.5 by **5.1%** on average across seven benchmarks and InternVL-2 by a 15.3% gain on the complex M^3CoT benchmark.

2 Related Works

Multimodal Chain-of-Thought. The huge success of CoT in textual modality inspires a lot of works to explore the potential of multimodal CoT in cross-modal scenarios.(Besta et al., 2024; Zhang et al., 2023a; Sprague et al., 2025) Visual Programming(VisProg)(Gupta and Kembhavi, 2023; Surís et al., 2023) makes use of the in-context learning ability of LLMs to generate stepwise solutions in the form of Python-like modular programs, and each line of the program invokes one of the VLMs or Python functions. Since LLMs are prone to generating programs that seem to match the questions but are irrelevant to the image content without "seeing" the images, the performance of VisProg is dependent on the quality of the programs. With the advancement of MLLMs and the emergence of various multimodal reasoning datasets, researchers attempt to guide the MLLMs to think step by step (Azzolini et al., 2025; Jaech et al., 2024). QVQ(Wang et al., 2024c), Kimi k-1.5(Team et al.) adopts long-CoT and exhibits powerful reasoning abilities in challenging math problems and Contest-Level Tasks. The generated CoT is expressed in text. VoCoT extends the modality of CoT by introducing coordinates and visual tokens of objects into the text tokens. However, this introduces an unnecessary computational burden when performing complex reasoning tasks.

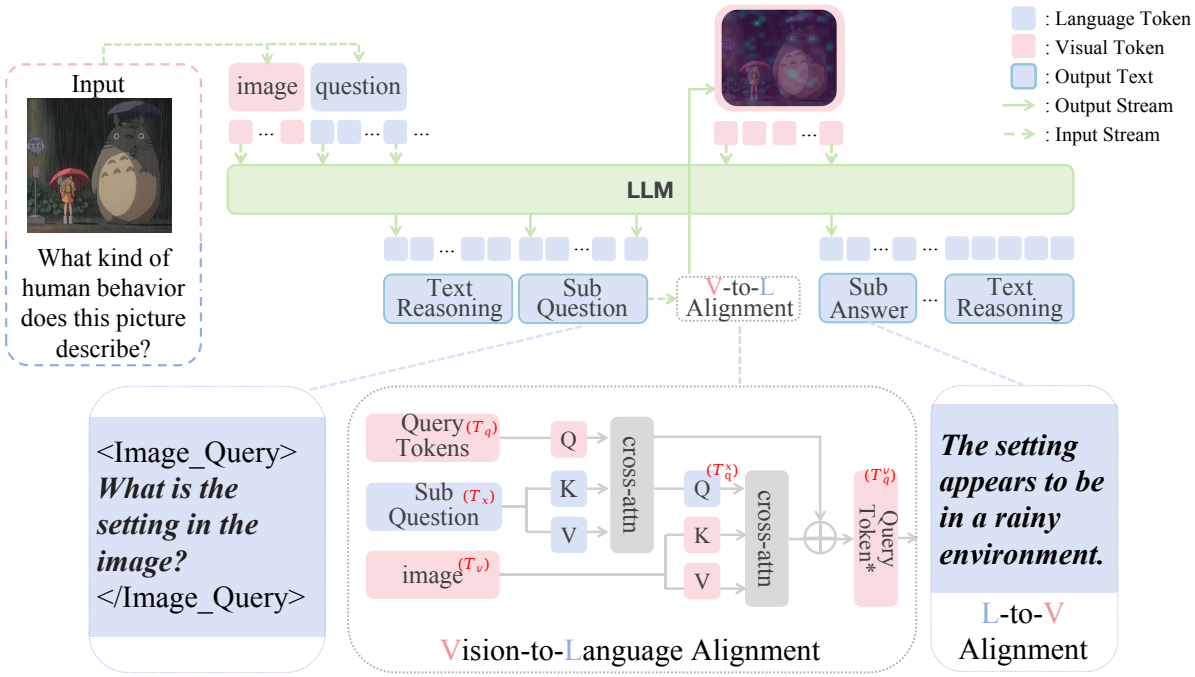


Figure 2: Pipeline of MirrorCoT. Given an image and a question, e.g., "What kind of human behavior does this picture describe", the MLLM begins standard next-token prediction to perform stepwise reasoning. During reasoning, the model is trained to explicitly emit sub-questions like "What is the girl doing" when visual information is required. A cross-attention module processes the image and current sub-question to extract question-relevant visual features into learnable query tokens. These tokens are inserted after the sub-question in the text sequence. Then, the text generation continued with this visually augmented context.

Vision Language Alignment in MLLMs. Hallucination refers to cases where the generated text contradicts or is not grounded in the input image. To address the hallucination caused by data bias, some researchers devise strict data-evaluation strategies (Yu et al., 2024a; Liu et al., 2023; Wang et al., 2024b) and spend a large amount of manpower on annotation data. Another cause of hallucination is the over-reliance on text and decreased attention to image context. (Huang et al., 2024; Liu et al., 2024a) intervene in the model's inference process and adjust the attention weights assigned to the image tokens. Although those methods are training-free, the effectiveness of those methods depends on the original ability of the MLLMs. Apart from processing during generation, post-processing (Zhou et al., 2024; Yin et al., 2024; Lee et al., 2024) is proposed to detect the hallucination and modify the generated text. (Yin et al., 2024) trains a model to inspect and verify the objects that appear in the text, and the result assists the model to re-generate answers. These methods greatly extend the inference time.

3 Methodology

When engaging in complex multimodal reasoning, humans typically decompose the task into a series of logically interconnected subtasks. As reasoning progresses, attention dynamically shifts across modalities to retrieve task-relevant information. Our proposed method aims to simulate this cognitive process to establish a more robust and interpretable chain-of-thought (CoT) paradigm. We name this approach **MirrorCoT**, as the two modalities mutually align with each other, like dual reflections in a mirror, ensuring coherent and step-aligned reasoning.

Preliminaries. We first recall some background on MLLMs in this section. Given an image I and an instruction T, the MLLMs processes multimodal inputs through three core components: (1) A vision encoder VE_ψ (e.g., CLIP-ViT(Radford et al., 2021)) extracts patch embeddings $E_v = \text{VE}_\psi(I) \in \mathbb{R}^{N_v \times d_v}$ from input image I; (2) A projector (typically an MLP) Proj_ϕ aligns visual features to the semantic space of LLM via $\hat{E}_v = \text{Proj}_\phi(E_v) \in \mathbb{R}^{N_v \times d_{\text{text}}}$; (3) An LLM (e.g., Vicuna(Chiang et al., 2023)) LLM_θ jointly attends to projected visual tokens $\hat{Z}_v$ and text tokens T through cross-modal

attention, generating coherent text outputs. This pipeline (Eq. 1) enables multimodal reasoning by unifying visual and linguistic representations in a shared latent space.

$$\text{Output} = \text{LLM}_\theta\left(\text{Proj}_\phi\left(\text{VE}_\psi(I)\right) \parallel T\right) \quad (1)$$

Multimodal Interleaved CoT. Different from textual CoT, MirrorCoT introduces the visual modality to simulate cross-modal interactions during reasoning. MirrorCoT uses the following interleaved format:"{text reasoning} <IMAGE_QUERY> {Visual Query} </IMAGE_QUERY> {query tokens} {text reasoning}..." We prompt the model to raise sub-questions(Visual Query) if visual evidence is needed and enclose each question with <IMAGE_QUERY> and </IMAGE_QUERY>. Once the </IMAGE_QUERY> token is detected, the query tokens carrying question-related image features will be inserted immediately after the </IMAGE_QUERY> tag. The process of text reasoning, question generation, and query token insertion can be iterated multiple times, allowing the model to perform multi-step reasoning while continuously incorporating visual information.

Vision-to-Language Alignment. Previous research in multimodal reasoning primarily focuses on improving the quality of pure text-based Chain-of-Thought (CoT), while some attempts have been made to incorporate image tokens that cover the mentioned objects. However, these approaches still lack advanced mechanisms for more fine-grained visual feature selection. To address this limitation, we propose leveraging query tokens (Li et al., 2023) to dynamically extract text-aligned visual features, which are then strategically inserted into the textual CoT sequence. Unlike heuristically selecting visual tokens (e.g., object crops), our method operates at a finer granularity, enabling more effective image feature selection. We build a cross-attention module to select image features. The activated cross-attention module identifies the question within the <IMAGE_QUERY> tags and subsequently extracts question-related image features into query tokens. The architecture of the module is followed (Tong et al., 2025) and is shown in Figure 2. Firstly, given query tokens T_q, image tokens T_v, and text tokens T_x. We first interact T_q and T_x via cross-attention:

$$\mathbf{T}_q^x = \text{CrossAttention}(\mathbf{T}_q, \mathbf{T}_x, \mathbf{T}_x) \quad (2)$$

Here, $\mathbf{T}_q^x \in \mathbb{R}^{n\times d}$ are the derived text queries. Then, we can use $\mathbf{T}_q^x$ to query visual information from the visual tokens $\mathbf{T}_v$, defined by

$$\mathbf{T}_q^v = \text{CrossAttention}(\mathbf{T}_q^x, \mathbf{T}_v, \mathbf{T}_v) \quad (3)$$

The query tokens output from the module of vision-to-language alignment will be appended to the sub-question. Both cross-attention operations share the same formulation as Eq. 4 but operate on different inputs:

$$\mathbf{CrossAttention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{Softmax}\left(\frac{(\mathbf{QW}_q)(\mathbf{KW}_k)^\top}{\sqrt{d_k}}\right)\cdot\mathbf{VW}_v. \quad (4)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are the projection weight matrices, with d_k denoting their dimension.

Language-to-Vision Alignment. As described in the above paragraph, once the cross-attention module is triggered, it can integrate relevant visual features into the query tokens, which are subsequently inserted into the output text sequence. Then the model continues to perform next-token prediction. The following generated textual tokens are based on both textual and visual context. Let T_x^i denote the text token, T_v^i denote the image token, and T_q^i denote the query token. The LLM generates the next token by jointly attending to previous text tokens and the sequence of query tokens, ensuring that each step of reasoning is tightly anchored to the visual input. The predicted text tokens follow the formulation in Eq.5:

$$\hat{T_x^i} \sim \mathcal{LLM}_\theta(T_x^i \mid T_v, T_x^0, \hat{T_x^1}, \hat{T_q^1}, \cdots, \hat{T_x^{i-1}}, \hat{T_q^i}) \quad (5)$$

This is consistent with the existing training paradigm of causal language modeling (Radford et al., 2019). Therefore, we can employ the standard auto-regressive loss to optimize in the same fashion.

Training Strategy. Our training involves two stages: 1) **Pre-training Stage.** We observe that random initialization of the cross-attention module and query tokens increases fine-tuning difficulty. Thus, pre-training for these components is essential. The objective of pre-training is to enable the cross-attention module to extract question-relevant image features. For this stage, we collect single-step solvable image-question-answer pairs as training data. The query tokens are appended to the input questions and output sub-questions. 2) **CoT Fine-tuning Stage.** Following pre-training, the second stage fine-tunes the model to align with

the MirrorCoT reasoning paradigm. The learnable query tokens are appended not only to the original question but also to each sub-question in the output. The embeddings at these query token positions are then replaced with the outputs of the cross-attention module, which processes the input image, question, and the corresponding answer at every step. For this stage, we train with our constructed dataset in the format of MirrorCoT.

4 Experiment

4.1 Experiments Setup

Training Data. We construct our training dataset based on LVIS-INSTRUCT4V (Wang et al., 2023) and LLaVA-CoT-100k (Xu et al., 2024). In the first training phase, only query tokens are learnable to learn to extract image features relevant to individual sub-questions, using single-step question-answer pairs sampled from LVIS-INSTRUCT4V and the multi-step reasoning chains during the subsequent CoT fine-tuning stage. For the CoT fine-tuning stage, we sample image-question pairs from the LLaVA-CoT-100k dataset and employ the QwenVL-2.5-72B (Bai et al., 2025; Wang et al., 2024c) model to generate MirrorCoT reasoning chains, explicitly prompting sub-question generation when visual evidence is required.

Benchmarks and Models. To evaluate the effectiveness of MirrorCoT, we conduct comprehensive experiments by integrating it with LLaVA-1.5-7B across six datasets spanning multimodal reasoning (MMStar (Chen et al., 2024a), MMBench (Liu et al., 2024b), MM-Vet (Yu et al., 2024b), ScienceQA (Lu et al., 2022)) and hallucination tasks (HallusionBench (Guan et al., 2024), HaloQuest (Wang et al., 2024d)). We also evaluate our method on the highly challenging M^3CoT (Chen et al., 2024b) dataset, which demands an average of 10.9 reasoning steps per question. To demonstrate the superiority of our MirrorCoT, we compare it with state-of-the-art multimodal reasoning methods, i.e., VAR(adjust attention assignment when inference) (Kang et al., 2025), VoCoT (Li et al., 2025) (interleave raw sub-image tokens with textual tokens), and textual CoT, including classic CoT (Wei et al., 2022) and VQD (Zhang et al., 2024a)(decompose the complex question into simple sub-questions), under a fair comparison.

Implementation Details. All experiments are conducted on NVIDIA A100 GPUs. Our training involves two stages, i.e., the Pre-training and CoT Fine-tuning stages with our constructed data. For the implementation of the comparison methods, we directly reproduce the results of these SOTA methods by using their official open-source code with default hyperparameters. Please refer to Section A.1 in Appendix for more implementation details about our training data and experimental settings.

4.2 Experimental Results and Analyses

MirrorCoT strengthens the textual CoT. As shown in Table 1, *our MirrorCoT consistently outperforms the direct answer baseline (Direct) on all the benchmarks* while offering enhanced interpretability. Furthermore, we evaluate MirrorCoT against the textual CoT fine-tuning(CoT) methods to assess its effectiveness. Given the same amount of data for fine-tuning LLaVA-1.5, *MirrorCoT achieves comparable or superior performance over four reasoning tasks over CoT.* Moreover, MirrorCoT exhibits reduced hallucination compared to textual CoT on the two hallucination benchmarks. MirrorCoT also surpasses CoT on M^3CoT by 1.1% as shown in Table 2. (The results in the Theory split exhibit significant discrepancy due to its minimal size, i.e., 11 questions.) Since our MirrorCoT introduces sub-questions, we further contrast MirrorCoT with VQD (Zhang et al., 2024a). (See Section A.4 in Appendix for detailed case studies comparing MirrorCoT with VQD.) As evidenced in Table 1, MirrorCoT outperforms VQD by 2.7% on MMStar, 1.8% on MMBench, and 2.0% on HallusionBench with fewer parameters. These findings indicate that MirrorCoT not only preserves the strengths of traditional CoT but also demonstrates significant improvements in both reasoning accuracy and anti-hallucination capabilities of MLLMs. This justifies the effectiveness of the proposed mirror-like bidirectional alignment between vision and language modalities.

MirrorCoT surpasses multimodal interleaved CoT. As demonstrated in Table 1, *our MirrorCoT outperforms VoCoT across 5 benchmarks* except MMStar. We speculate that this is because VoCoT is optimized specifically for handling object-centric tasks, which make up a large proportion (62.4%) of MMStar. Furthermore, we compare the average visual tokens introduced by VoCoT and MirrorCoT. The results (see Section A.3 in Appendix) indicate that MirrorCoT achieves comparable performance while requiring 91.1% fewer visual tokens (23.9 per question) than VoCoT (268.7 per question). Meanwhile, MirrorCoT also exhibits better task versatil-

Model	Param	MMStar	MMB	MMVet	SQA	HalluBench	Haloq
Random	-	24.9	25.3	-	25.2	47.3	19.1
Closed-source LVLMs							
GPT-4V(Yang et al., 2023)	-	**46.1**	**69.6**	**67.7**	**81.4**	**65.3**	-
GeminiPro-Vision(Team et al., 2023)	-	42.6	68.1	63.1	80.6	36.9	-
Open-source LVLMs							
VAR(Kang et al., 2025)	7B	29.9	65.4	29.5	64.3	27.8	23.8
VoCoT(Li et al., 2025)	7B	36.2	68.1	29.0	62.8	48.2	18.0
VQD(Zhang et al., 2024a)	13B	33.4	68.0	27.8	**70.8**	48.4	20.6
LLaVA-1.5	7B	31.9	64.3	29.0	64.7	47.4	20.3
+CoT	7B	34.7	70.8	22.2	63.2	28.9	20.1
+MirrorCoT	7B	**37.3**	68.7	23.5	65.3	29.4	23.1
+CoT‡	7B	33.7	**72.6**	29.5	64.2	**50.5**	25.4
+MirrorCoT‡	7B	35.9	69.8	**29.5**	**65.8**	50.4	**33.2**
InternVL2	8B	**56.8**	**84.8**	48.1	**81.1**	**54.1**	8.2
+Zero-shot CoT	8B	51.8	80.7	47.5	51.1	33.1	27.2
+CoT	8B	27.9	35.7	15.0	**78.6**	48.1	39.4
+MirrorCoT	8B	45.4	68.1	**48.9**	77.9	48.9	**50.0**

Table 1: Benchmark results on general vision-language task. Performance is evaluated across two categories of datasets: general cross-modal benchmarks (MMStar, MMBench, MM-VET, ScienceQA) and hallucination-specific benchmarks (HallusionBench, HaloQuest). The experiments marked with ‡ are conducted using the maximum amount of data available.

Model	Param	Science			Commonsense			Mathematics			Total
		Lang	Natural	Social	Physical	Social	Temporal	Algebra	Geometry	Theory	
Random	-	32.7	30.6	26.7	33.0	22.2	20.3	35.7	27.5	23.8	28.6
VAR(Kang et al., 2025)	7B	73.2	73.3	18.3	41.4	23.5	15.0	15.4	19.4	9.1	28.6
VoCoT(Li et al., 2025)	7B	26.4	36.3	25.2	41.4	31.0	45.0	35.4	25.0	27.3	32.1
VQD(Zhang et al., 2024a)	13B	48.3	29.2	26.6	63.4	64.7	36.7	32.3	22.2	27.3	35.3
LLaVA-1.5	7B	43.7	29.7	17.6	75.6	68.9	16.7	18.4	16.7	18.1	31.4
+CoT	7B	40.2	34.3	27.2	80.5	62.1	36.7	26.2	22.2	72.7	37.1(+5.7)
+MirrorCoT	7B	34.5	34.3	27.9	68.3	62.1	48.3	53.8	22.2	63.6	**38.2**(+6.8)
InternVL-2	8B	26.4	29.7	24.9	53.7	56.9	30.0	35.4	19.4	10.0	31.6
+CoT	8B	24.1	26.3	33.6	39.0	54.3	36.7	29.2	25.0	18.2	32.1(+0.5)
+MirrorCoT	8B	70.1	55.0	23.9	78.0	60.3	63.3	23.1	33.3	45.5	**46.9**(+15.3)

Table 2: Results on M^{3}CoT. The baseline methods encompass direct answer and CoT fine-tuning.

ity in non-object-centric tasks. This validates the superiority of our MirrorCoT, thanks to the proposed dynamic visual token insertion strategy.

MirrorCoT mitigates performance degradation of CoT fine-tuning. For InternVL2-8B, direct-answer accuracy on datasets like MMStar and MMBench is already near the ceiling. Both standard CoT fine-tuning and our MirrorCoT hurt performance on these high-baseline sets; as reported in (Wang et al., 2025), the drop is caused by the distribution shift inherent to SFT-teacher forcing. MirrorCoT mitigates this degradation by injecting visual-modality tokens during reasoning: even if the model drifts slightly in the auto-regressive phase, the injected query tokens recurrently realign the attention distribution with the evidence present in the image, mitigating error accumulation. Moreover, on low-baseline datasets such as M^{3}CoT and HaloQuest, MirrorCoT delivers larger improvements than purely textual CoT.

5 Ablation Studies

Ablation on CoT formats. As depicted in Figure 2, our MirrorCoT has two key components: sub-questions and learnable query tokens. We ablate the two key components of our method to analyze their contributions. Adding sub-questions alone improves the CoT baseline on MMStar and MM-Vet, while further incorporating learnable query tokens brings an additional gain of +1.0% (from 47.7% to 48.7%) on average, showing the combination of sub-question and query token is necessary.

Effects of training strategy. As delineated in Section 4.1, our proposed framework adopts a two-phase training paradigm: the pretraining stage(Stage 1) followed by the CoT fine-tuning stage(Stage 2). To evaluate the contribution of each training stage, we conduct ablation experiments based on LLaVA-1.5-7B with the maximum amount of data. The baseline model without training achieves 47.5% task accuracy on average.

Model	Sub Question (L → V)	Query Tokens (V → L)	MMStar	MMB	MMVet	ScienceQA	Average
Direct			31.9	64.3	29.0	64.7	47.5
MirrorCoT	✓		35.1	68.4	**25.2**	61.9	47.7
MirrorCoT	✓	✓	**37.3**	**68.7**	23.5	**65.3**	**48.7**

Table 3: Ablation on components of MirrorCoT. Sub-questions activate the language-to-vision alignment, and query tokens serve as the carrier of visual information in vision-to-language alignment.

When introducing only the CoT fine-tuning stage, it demonstrates a 1.3% improvement in average task performance compared to the baseline, but shows a 2.6% reduction compared to the model trained with both stages.

Effects of training data volume and number of query tokens. To investigate the impact of training data volume on model performance, we conduct a series of ablation studies based on LLaVA-1.5-7B by progressively varying the size of the training set. As the amount of training data increases, the performance of LLaVA-1.5-MirrorCoT on the four datasets shows an overall upward trend. While MirrorCoT exhibits a performance decline before improving at low data volumes(6k,30k). We speculate that training with limited data amplifies data bias-induced hallucinations, which manifest most prominently on hallucination benchmarks. We also evaluate different amounts of query tokens, ranging from 6 to 18. The results indicate that the model's performance generally improves as the number of query tokens increases. For example, increasing the token count from 9 to 18 led to an improvement in MMBench from 69.8% to 70.6%. This trend suggests that more query tokens may provide the model with richer representations to capture complex patterns of data. However, the marginal benefits decrease as the length increases.

Effects of the quality of sub-questions. To investigate how the quality of sub-questions influences model performance, we conduct an experiment by employing three prompt variants with distinct preferences for sub-question types to fine-tune LLaVA-1.5 via our MirrorCoT. Concretely, these three levels of prompt variants, listed as follows, can significantly influence the generated sub-questions with different levels of requirements: 1) Fine-grained Level: This variant discourages reasoning-based sub-questions, favors specific and clear sub-questions.(e.g., "What is the color of the strips?"). 2) Default Level: Currently used by our MirrorCoT. 3) Coarse-grained Level: Tends to yield global and abstract questions requiring reasoning, linking language to visual regions, or exploring implicit concepts. (e.g., "What visible elements suggest an artistic activity?") The results in Table 5 demonstrate that these three prompt variants achieve comparable performance. Most notably, the Default Level averagedly delivers the best performance on three of four benchmarks. This suggests that excessive prompt engineering may not always yield superior results, and less constrained prompts might be superior by allowing for more natural model reasoning.

6 Discussions

We visualize the bidirectional loop of MirrorCoT in Figure 4, allowing us to explore the effectiveness of its active visual evidence extraction mechanism.

Vision-to-Language Alignment. As the model progressively generates each reasoning step, it simultaneously raises sub-questions to inquire about visual evidence. For example, when identifying the human activity in the image, MirrorCoT generates the sub-question "What item is the man carrying?". Rather than simply selecting image tokens that cover the man's location, our cross-modal attention mechanism simultaneously focuses on both the subject and his surrounding environment and embeds these features into tokens after the sub-question, which can be observed in the attention map of <Query Token 2> in the case of Figure 4. This enables a visual finding that "the man is carrying a suitcase". From an information entropy perspective, the vision-to-language alignment increases the effective information, which reduces uncertainty and narrows down possible answers, making the reasoning process more efficient and accurate.

Language-to-Vision Alignment. To mitigate the over-reliance on language priors, we employ a language-to-vision alignment mechanism. When query tokens are inserted into the textual CoT, the model's reasoning is more closely tied to the visual context. Figure 8 in the appendix demonstrates this improvement through a comparison with textual CoT. Textual CoT may generate visually inconsistent answers (e.g., describing the furniture

Model	Stage1	Stage2	MMStar	MMB	MMVet	ScienceQA	Average
Direct			31.9	64.3	29.0	64.7	47.5
MirrorCoT		✓	35.7	68.2	29.0	62.1	48.8
MirrorCoT	✓	✓	**35.9**	**69.8**	**29.5**	**65.8**	**50.3**

Table 4: Ablation on training stages. The experiment is conducted with the maximum amount of data.

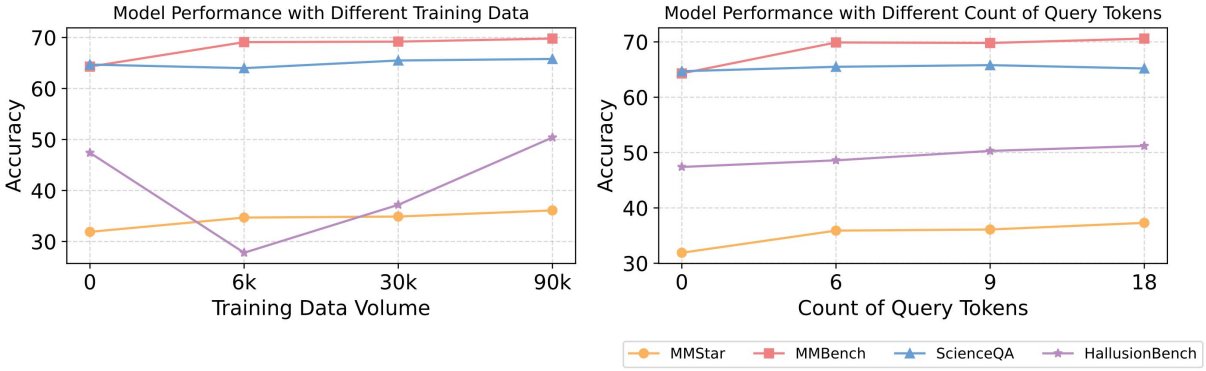


Figure 3: Ablations on the volume of training data and the count of query tokens.

Prompt Variant	MMStar	MMB	MMVet	SQA	Avg.
Fine-grained	34.7	68.9	23.9	65.1	48.2
Default	37.3	68.7	23.5	65.3	**48.7**
Coarse-grained	35.6	63.2	22.0	64.8	46.4

Table 5: The evaluation results on the MMStar, MMBench, MMVet, and SQA benchmarks.

as "a table with a flat surface" despite the image clearly showing a sofa), MirrorCoT identifies discriminative visual features like backrest structures to determine the answer as "couch" through dynamic visual grounding. By inserting query tokens into textual sequences, our method disrupts the language dominance through attention redistribution. This intervention increases the overall attention assigned to the visual modality, effectively counteracting language bias.

7 Conclusion

This paper focused on the core challenge in MLLMs when performing long-term multimodal interleaved CoT generation: effective visual information extraction. We propose MirrorCoT, a lightweight multimodal CoT with query-triggered injection: (1) a vision operator to extract text-relevant visual features instead of raw image patches for vision-to-language grounding, and (2) a token-level conditioning mechanism that dynamically integrates these query tokens into decoder inputs during text generation, ensuring language-to-vision alignment by forcing explicit reference to visual context. This design alleviates the excessive visual-token overhead inherent in prior multimodal interleaved approaches.

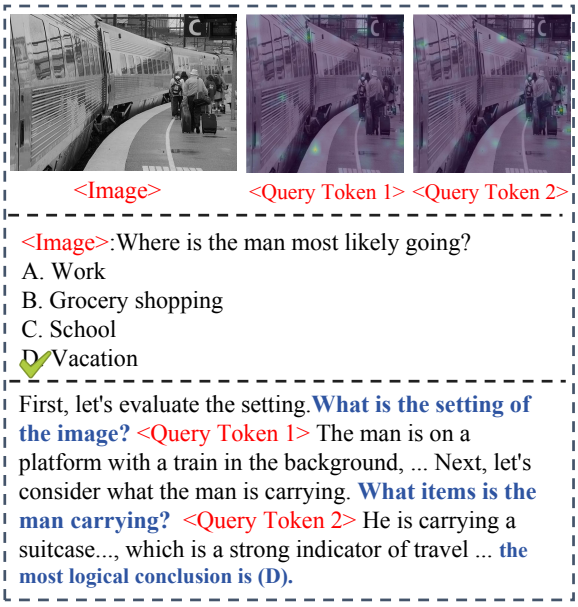


Figure 4: Visualization of the attention weights of the last cross-attention layer formulated by Eq.3. The attention maps highlight the regions of interest corresponding to the specific questions posed.

Limitations

The bidirectional alignment of MirrorCoT requires additional cross-modal attention operations, which may increase inference latency compared to textual CoT methods. Future work could explore more lightweight modality interaction method.

References

Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, and 1 others. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024a. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*, volume 37, pages 27056–27087. Curran Associates, Inc.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024b. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8199–8221, Bangkok, Thailand. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024c. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2024. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 391–404, Mexico City, Mexico. Association for Computational Linguistics.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

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

Zejun Li, Ruipu Luo, Jiwen Zhang, Minghui Qiu, Xuanjing Huang, and Zhongyu Wei. 2025. 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, Albuquerque, New Mexico. Association for Computational Linguistics.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Aligning large

multi-modal model with robust instruction tuning. *CoRR*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11888–11898.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, A Du, B Gao, B Xing, C Jiang, C Chen, C Li, C Xiao, C Du, C Liao, and 1 others. Kimi k1. 5: Scaling reinforcement learning with llms, 2025. *URL https://arxiv. org/abs/2501.12599*.

Bo Tong, Bokai Lai, Yiyi Zhou, Gen Luo, Yunhang Shen, Ke Li, Xiaoshuai Sun, and Rongrong Ji. 2025. Flashsloth: Lightning multimodal large language models via embedded visual compression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. 2023. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2025. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *Preprint*, arXiv:2411.10442.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. 2024d. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *European Conference on Computer Vision*, pages 288–304. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kun Xiang, Zhili Liu, Zihao Jiang, Yunshuang Nie, Runhui Huang, Haoxiang Fan, Hanhui Li, Weiran Huang, Yihan Zeng, Jianhua Han, and 1 others. 2024. Atomthink: A slow thinking framework for multimodal mathematical reasoning. *arXiv preprint arXiv:2411.11930*.

Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024b. Mm-vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Haowei Zhang, Jianzhe Liu, Zhen Han, Shuo Chen, Bailan He, Volker Tresp, Zhiqiang Xu, and Jindong Gu. 2024a. Visual question decomposition on multimodal large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1926–1949, Miami, Florida, USA. Association for Computational Linguistics.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2024b. Improve vision language model chain-of-thought reasoning. *arXiv preprint arXiv:2410.16198*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023a. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

A Technical Appendices and Supplementary Material

A.1 More Implementation Details

Training Settings. Our training involves two stages, i.e., the Pre-training and CoT Fine-tuning stages. The objective of pre-training is to enable the cross-attention module to extract question-relevant image features. Following pre-training, the second stage fine-tunes the model to align with the MirrorCoT reasoning paradigm. All experiments are conducted on NVIDIA A100 GPUs. For LLaVA-1.5-7B, we pre-trained the model for 5 epochs (batch size 8, learning rate 2×10^{-5}), followed by 1 CoT fine-tuning epoch (batch size 10, learning rate 2×10^{-5}). The pre-training requires 6 hours, and the CoT fine-tuning requires 16 hours with the maximum amount of data(90k). For InternVL-Chat-V1.5-4B, we adjust the hyperparameters to 3 pre-training epochs (batch size 8, learning rate 4×10^{-5}) and 4 fine-tuning epochs (batch size 10, learning rate 4×10^{-6}). The pre-training requires 7 hours, and the CoT fine-tuning requires 9 hours. In the first stage, only the parameters of the cross-attention module and query tokens are updated. In the second stage, besides query tokens and the cross-attention module, the last 6 layers of the LLaVA-1.5 language backbone and the last 4 layers of the InternVL-Chat-V1.5 language backbone are unfrozen.

Training data construction. We construct our training dataset based on LVIS-INSTRUCT4V (Wang et al., 2023) and LLaVA-CoT-100k (Xu et al., 2024). For the pre-training stage, we curate a dataset consisting of 12,000 question-answer pairs sampled from LVIS-INSTRUCT4V, and 8,000 question-answer pairs extracted from the multi-step reasoning chains in the CoT fine-tuning stage. For the CoT fine-tuning stage, we sample 95,000 image-question pairs from LLaVA-CoT-100K and use QwenVL-2.5-72B (Bai et al., 2025) to generate the reasoning chains in the MirrorCoT paradigm automatically. This transformation introduces explicitly dynamic sub-question generation when visual evidence is required and structured reasoning steps that alternate between visual verification and logical inference. The examples of our data are demonstrated in Figure 5.

A.2 Benchmarks and Evaluation

We conduct extensive evaluations on seven benchmarks, which can be categorized into two distinct classes. The first class comprises cross-modal reasoning datasets, designed to assess the models' abilities to integrate and reason across different modalities. Benchmarks requiring LLM-based evaluation use Qwen-Plus (Yang et al., 2024) as the judge model.

MMStar MMStar (Chen et al., 2024c) is an elite vision-indispensable multimodal benchmark comprising 1,500 challenge samples meticulously selected by humans. The samples are chosen to ensure visual dependency, minimal data leakage, and the need for advanced multimodal capabilities for solutions.

MMBench MMBench (Liu et al., 2024b) is a comprehensive benchmark designed to evaluate the multimodal understanding capability of large vision-language models.

MM-Vet MM-Vet (Yu et al., 2024b) assesses the performance of models in six fundamental vision-language skills: numeracy, recognition, knowledge, spatial awareness, language creation, and optical character recognition (OCR).

M^{3}CoT M^{3}CoT (Chen et al., 2024b) aims to advance multi-domain, multi-step, and multi-modal reasoning by integrating visual and textual modalities. The dataset requires models to perform complex, multi-step reasoning. On average, each question involves 10.9 steps to reach a conclusion, significantly higher than other benchmarks like ScienceQA (which averages 2.5 steps).

ScienceQA ScienceQA (Lu et al., 2022) is a benchmark focused on evaluating models' ability to answer science-related questions using both visual and textual information. The second class encompasses hallucination question-answering datasets that evaluate the models' performance in addressing hallucination-related issues.

HallusionBench HallusionBench (Guan et al., 2024) is an advanced diagnostic suite designed to evaluate the visual illusions and knowledge hallucinations of large vision-language models. Given the cross-modal nature of our research, we evaluate model performance on its image split.

HaloQuest HaloQuest (Wang et al., 2024d) is a novel Visual Question Answering (VQA) dataset focusing on multimodal hallucinations in vision-language models (VLMs). It combines real images and synthetic images, with questions and answers designed to trigger and evaluate hallucinations.

Pre-training Stage

Source: LVIS-INSTRUCT4V

Question: <Fig1>What is the position of the man relative to the plants?
Answer: The man is in the foreground, and the plants are behind him, separated by a glass partition.

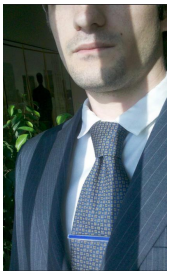
Source: Sub-questions extracted from MirrorCoT reasoning chains

Question: <Fig2>What letters are visible on the train?
Answer: The letters "OBB" are clearly visible on the side of the train.

CoT Finetuning Stage

Source: LLaVA-CoT-100k

Question: <Fig2>What letter appears twice in a row on the train?
A. g B. d C. w D. b
Answer: <SUMMARY> The core objective of the question is to identify which letter appears twice in a row on the train from the given options: A. g, B. d, C. w, D. b. </SUMMARY>
<CAPTION> The image shows a train at a station platform during nighttime. The train is red and has the letters "OBB" prominently displayed on its side. The platform is well-lit, and there are some people in the background near the entrance of the station. </CAPTION>
<REASONING> To answer this question, we need to carefully examine the letters visible on the train.<IMAGE_QUERY> What letters are visible on the train? </IMAGE_QUERY> The letters "OBB" are clearly visible on the side of the train.<IMAGE_QUERY> Are any of these letters repeated consecutively? </IMAGE_QUERY> Yes, the letter "B" appears twice in a row in the sequence "OBB".By comparing the visible letters with the given options, we can determine that the letter "B" matches option D.</REASONING>
<CONCLUSION> D. b </CONCLUSION>



<Fig1>

<Fig2>

Figure 5: Examples of our training data in the two training stages.

Table 6: Benchmarks and Evaluation strategy.

	MMStar	MMBench	MM-Vet	ScienceQA	M^3CoT	HallusionBench	Haloquest
Split	-	dev	test	test	dev	image	val
Metric	Acc	Acc	Score	Acc	Acc	Acc	Acc
LLM-Eval			✓				✓

A.3 Comparison with VoCoT

To better illustrate the advantage of MirrorCoT over VoCoT, we perform a quantitative comparison between MirrorCoT and VoCoT.

Task Versatility. While VoCoT specializes in object-centric tasks to enhance visual grounding, we propose MirrorCoT as a more versatile framework for general multimodal reasoning. To systematically evaluate this advantage, we categorize each dataset into object-centric and non-object-centric subsets through LLM-assisted annotation. As depicted in Figure 6(a), MirrorCoT surpasses VoCoT over the non-object-centric split of MMBench, MM-Vet, ScienceQA, and M^3CoT.

Less Computational Cost. VoCoT inserts raw image tokens whenever object references are detected in the text, regardless of their relevance to the reasoning task. As demonstrated in Section 6, this leads to excessive insertion of non-critical visual tokens. We quantify this inefficiency by measuring the average number of visual tokens per sample. Meanwhile, we compare the per-sample average visual tokens required between MirrorCoT and VoCoT across three multimodal benchmarks: MMStar, MMBench, and MM-Vet. The results demonstrate that MirrorCoT achieves comparable performance while requiring 91.1% (23.9) fewer visual tokens than VoCoT (268.7).

A.4 Case Study

Comparison with VQD. We further contrast MirrorCoT with VQD(Zhang et al., 2024a) (Visual Question Decomposition), a representative approach that decomposes main questions into sub-questions and draws the conclusion based on those sub-questions and corresponding answers. Instead of proposing all sub-questions at first, MirrorCoT introduces them progressively as the reasoning unfolds. For example, as illustrated in Figure 9, VQD decomposes "What can be inferred about the location shown in the image?" into a sub-question ("Are the animals in the picture already fenced in?"). The information gleaned from the sub-question is insufficient to address the original query. However, MirrorCoT progressively asks "What types of animals are visible, and how are they being treated", "Are there any signs of structured activity or infrastructure?". Each sub-question builds on the insights gained from the previous ones, creating

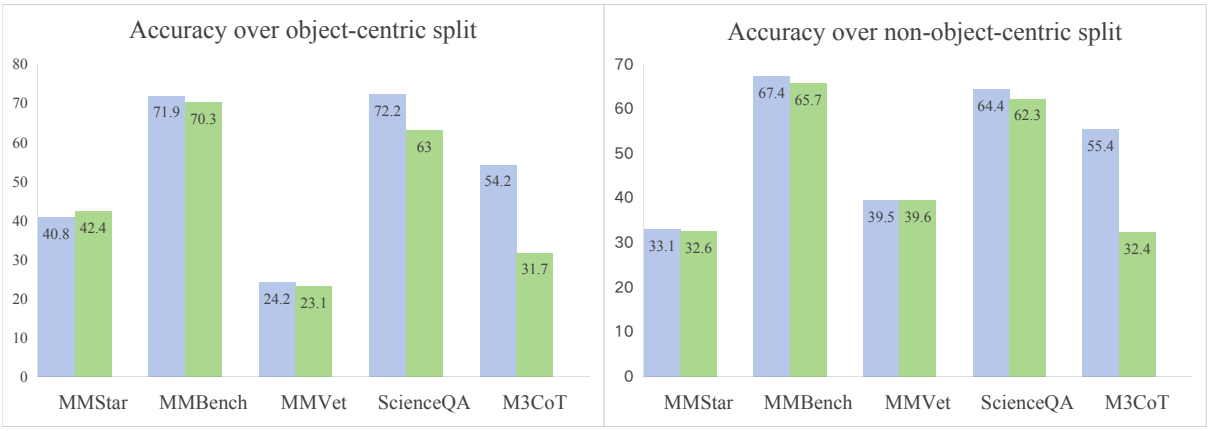


Figure 6: The results of MirrorCoT and VoCoT on benchmarks that are divided into object-centric and non-object-centric splits.

Method	CHAIR↓	Cover↑	Hal↓	Cog
VAR	8.3	52.5	38.8	4.3
VQD	6.6	42.4	18.8	1.2
Direct	7.5	51.7	35.1	4.2
CoT	7.5	51.7	34.7	4.2
MirrorCoT	**5.0**	**53.6**	**29.6**	2.7

Table 7: Results on AMBER generative task.

a chain of reasoning that progressively collects visual evidence and drives the model toward a more accurate final answer. More cases please refer to Figure 9 and Figure 10.

Comparison with Textual CoT and VoCoT. To highlight the comparative advantages of MirrorCoT over conventional CoT and VoCoT(Li et al., 2025), Figure 11 and Figure 12 present representative case studies. For example, in the second case of Figure 11, MirrorCoT accurately identifies the distance between Pair 1 and Pair 2, whereas both CoT and VoCoT produce erroneous measurements, demonstrating the superior modality alignment capability of MirrorCoT.

A.5 Hallucination Alleviation

Contradictions between the image input and the textual output are called hallucinations. The experimental results in Section 3.2 of the main text demonstrate that our MirrorCoT is also effective in alleviating the hallucination of MLLMs. Since our paper aims to improve reasoning ability through the promotion of modality alignment, we did not extensively address the capability of our method to mitigate hallucination in the main text. In the supplementary material, we demonstrate more experimental results of our MirrorCoT on the hallucination benchmark.

Benchmark and Baselines. AMBER(Wang et al., 2024a) is an LLM-free multi-dimensional benchmark designed for evaluating multi-modal hallucinations in large language models (LLMs). It is capable of assessing both generative and discriminative tasks. We evaluate MirrorCoT and CoT fine-tuning methods on the generative task of AMBER. AMBER introduces several metrics to evaluate hallucinations in the generative task: **CHAIR**(Rohrbach et al., 2019) measures the frequency of hallucinatory objects appearing in the responses. **Cover** measures the extent to which the response covers the image description. **Hal** represents the proportion of responses with hallucinations. **Cog** assesses whether the hallucinations in MLLMs are similar to those in human cognition.

Qualitative Results. As presented in Table 7, MirrorCoT demonstrates significant advantages over textual CoT(VQD(Zhang et al., 2024a) and CoT fine-tuning) and VAR(Kang et al., 2025)(identify and reallocate excess attention weights from visual trap tokens to more informative visual tokens within an image) across all four evaluation metrics in the AMBER generative task. Notably, our method achieves the lowest CHAIR (5.0) and Hal (29.6) scores among all methods. The superior performance is further evidenced by MirrorCoT attaining the highest Cover score (53.6), indicating that the responses of MirrorCoT are not only more accurate but also more detailed. These quantitative improvements demonstrate MirrorCoT's effectiveness in producing more reliable and semantically comprehensive outputs while mitigating hallucination risks.

Your Task: Classify the given question into object-centric or non-object-centric based on the definitions and examples below.
Definitions:
Object-Centric Questions:
Explicitly refer to specific objects/entities (e.g., "car", "person", "book").
Focus on physical attributes (color, shape, position), counts, or spatial relationships (e.g., "left of", "holding").
Examples: "What color is the vase on the table?", "How many dogs are in the image?", "Is the laptop open or closed?"
Non-Object-Centric Questions:
Concern abstract concepts, scene-level understanding, or subjective interpretation.
May involve emotions, intentions, temporal dynamics, or non-visual context.
Examples: "Why is this scene unusual?", "Describe the mood of the photograph.", "What might happen next in this situation?"
If it is object-centric, please answer with "1". Otherwise, answer with "0".
question: {question}
your response:

Figure 7: Prompt for dividing datasets into object-centric and non-object-centric splits.

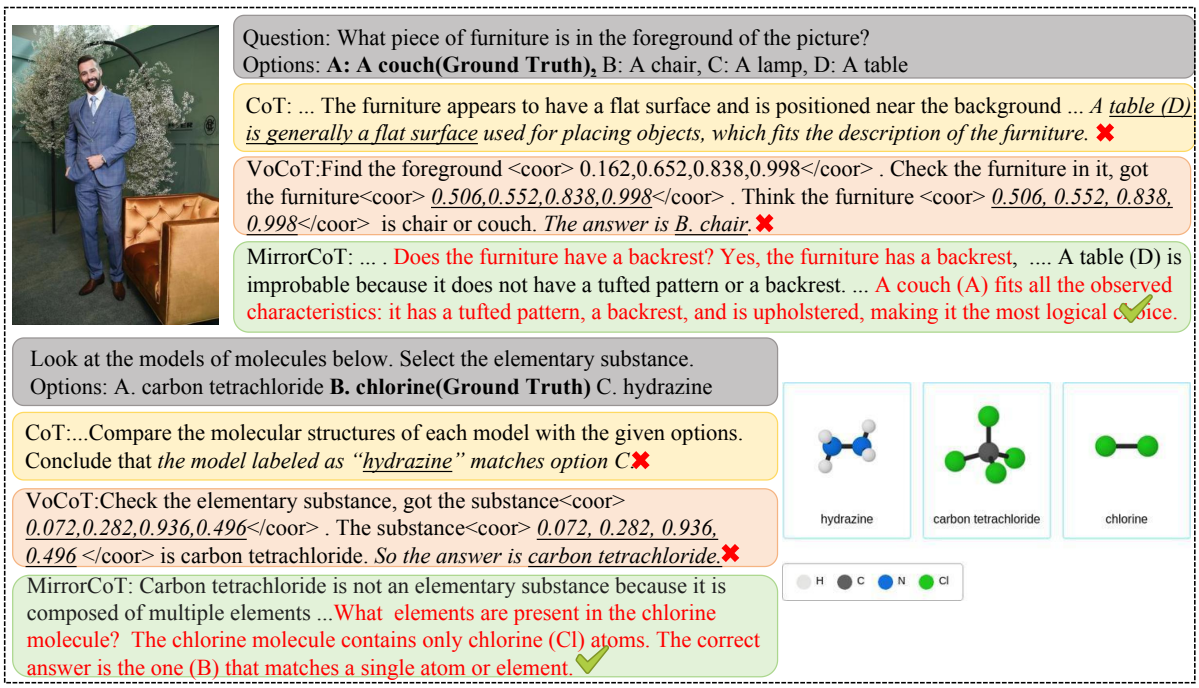


Figure 8: Case study. Comparison of CoT, VoCoT, and our MirrorCoT.

Question: What can be inferred about the location shown in the image?
Options: A. It is a petting zoo B. It is a wild animal sanctuary
C. It is a hobby farm(Ground Truth) D. It is a commercial farm

VQD: Sub-Question: 1. Are the animals in the picture already fenced in?
Sub-Answer: 1. Yes, the animals are already located in the fence.
Final Answer: petting zoo ✗

MirrorCoT: ...What types of animals are visible, and how are they being treated? ... the animals appear domesticated and are being cared for casually ...Are there any signs of structured activity or infrastructure? The absence of fences or barriers implies no strict control over movement, which aligns more with a hobby farm ...the location is a hobby farm.✓

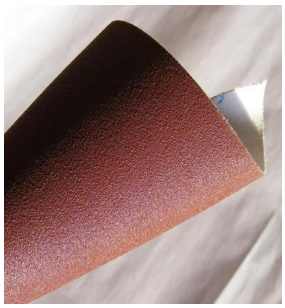
Select the mammal below.
A. Figure A B. Figure B C. Figure C **D. Figure D(Ground Truth)**

VQD: Sub-Question: 1. Which figure is a mammal?
2. A, B, C or D is the mammal among the four figures?
Sub-Answer: 1. Figure C is a mammal. 2. C) figure C
Final Answer: Figure C ✗

MirrorCoT: ...What does Figure A show? Figures A appears to be an aquatic animal resembling an alligator ... Figure C depicts fish swimming underwater... Figure C can be eliminated. Does Figure D fit the description of a mammal? ... Figure D portrays a human child sitting outdoors. Figure D (human child) fits the criteria for being a mammal.✓

Now, let's delve into the context: Mammals have hair or fur and feed their young milk. Mammals are warm-blooded. Warm-blooded animals can control their body temperature. A koala is an example of a mammal.

Figure 9: Comparison of our MirrorCoT and VQD.



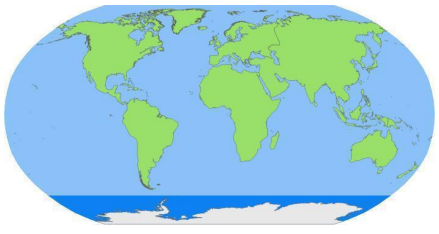
Question: What's the function of the demonstrated object?
A. Measure the temperature **B. burnish(Ground Truth)** C. Brushing D. Cut the grass

VQD: Sub-Question: 1. Does the bumpy surface of the object suggest that it is designed for brushing? 2. Does the object's shape and its function of cutting grass indicate its purpose?
Sub-Answer: Yes, ... *2. Yes, the object is quite like a lawn mower.*
Final Answer: D) cut the grass ✗

MirrorCoT: ...What is the shape and texture of the object? The object has a curved shape and a textured surface, which suggests it might be used for brushing or polishing. ... the most logical conclusion is that the object is used for Burnish (B).✓

Question: Which ocean is highlighted?
Options: **A: the Southern Ocean(Ground Truth)** B: the Pacific Ocean
C: the Arctic Ocean D: the Indian Ocean

VQD: Sub-Question: 1. Does the map show the outline of the continents and the shape of the oceans? 2. Which ocean has its shape and the outlines of the continents that border it clearly visible?
Sub-Answer: 1. Yes, the map shows the outline of the continents and the shape of the oceans clearly. 2. *The Pacific Ocean*
Final Answer: B: the Pacific Ocean ✗



MirrorCoT: ...What is the size and shape of the highlighted ocean? The ocean appears to be relatively large and has a distinctive shape that aligns with the Southern Ocean ... the Southern Ocean as the correct answer.✓

Figure 10: Comparison of our MirrorCoT and VQD.

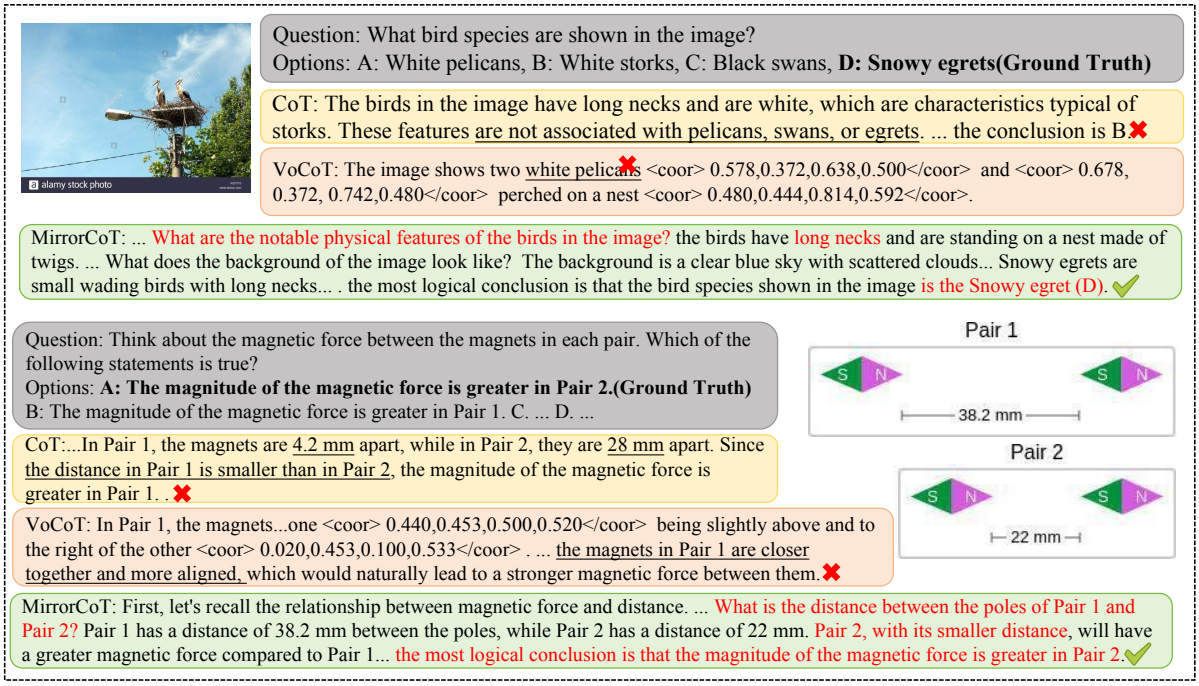


Figure 11: Comparison of our MirrorCoT, CoT, and VoCoT.

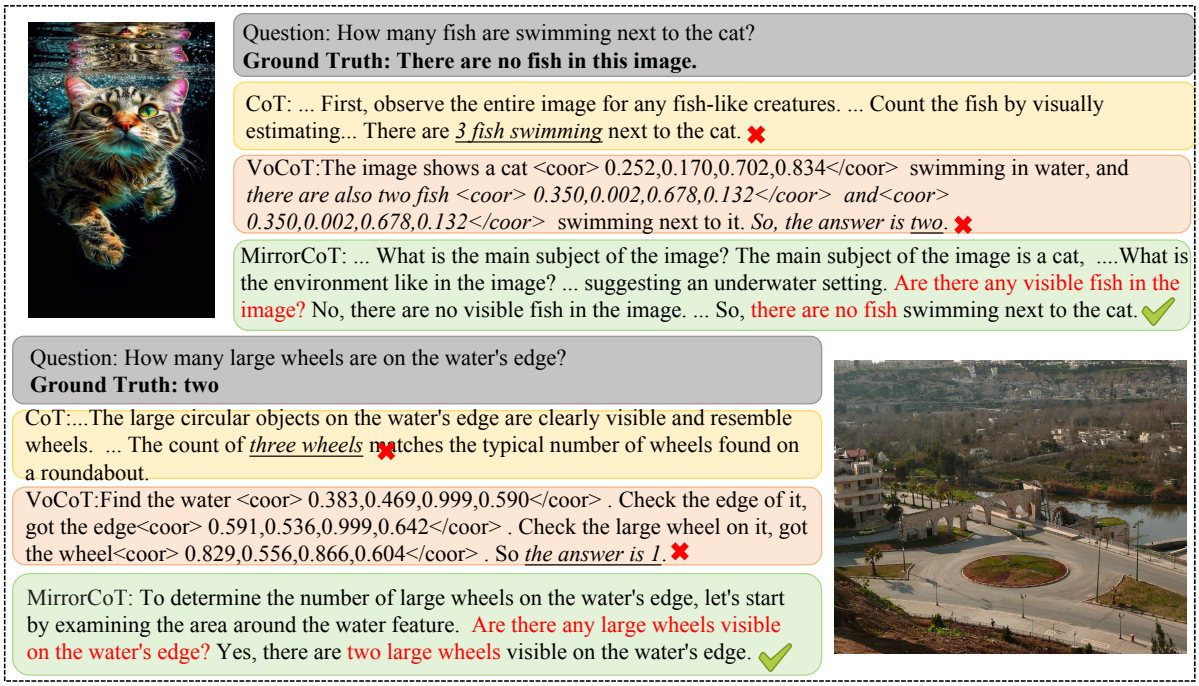


Figure 12: Comparison of our MirrorCoT, CoT, and VoCoT.