SceneAlign: Aligning Multimodal Reasoning to Scene Graphs in Complex Visual Scenes

Chuhan Wang[1*], **Xintong Li**[1*], **Jennifer Yuntong Zhang**[2], **Junda Wu**[1], **Chengkai Huang**[3,4], **Lina Yao**[3], **Julian McAuley**[1], **Jingbo Shang**[1]
[1]University of California, San Diego [2]University of Toronto
[3]University of New South Wales [4]Macquarie University
{chw136,xil240,juw069,jmcauley,jshang}@ucsd.edu
jenniferyt.zhang@mail.utoronto.ca
{chengkai.huang1, lina.yao}@unsw.edu.au

Abstract

Multimodal large language models often struggle with faithful reasoning in complex visual scenes, where intricate entities and relations require precise visual grounding at each step. This reasoning unfaithfulness frequently manifests as hallucinated entities, mis-grounded relations, skipped steps, and over-specified reasoning. Existing preference-based approaches, typically relying on textual perturbations or answer-conditioned rationales, fail to address this challenge as they allow models to exploit language priors to bypass visual grounding. To address this, we propose SceneAlign, a framework that leverages scene graphs as structured visual information to perform controllable structural interventions. By identifying reasoning-critical nodes and perturbing them through four targeted strategies that mimic typical grounding failures, SceneAlign constructs hard negative rationales that remain linguistically plausible but are grounded in inaccurate visual facts. These contrastive pairs are used in Direct Preference Optimization to steer models toward fine-grained, structure-faithful reasoning. Across seven visual reasoning benchmarks, SceneAlign consistently improves answer accuracy and reasoning faithfulness, highlighting the effectiveness of grounding-aware alignment for multimodal reasoning.

1 Introduction

Multimodal large language models (MLLMs) have achieved remarkable progress across tasks such as visual question answering, captioning, and instruction following (Yin et al., 2024; Li et al., 2024c). However, they often struggle with visually grounded reasoning, frequently exhibiting reasoning unfaithfulness (Yu et al., 2025; Liu et al., 2025a) where the generated rationales are decoupled from the actual visual evidence, even when the final answer appears correct (Wu et al., 2025; Man et al., 2025). This issue becomes particularly severe in complex visual scenes (Akter et al., 2024; Yang et al., 2022; Du et al., 2023; Ding et al., 2021), where numerous objects and intricate relations require precise, step-wise grounding. Typical manifestations of such unfaithfulness include object hallucination, relation mis-grounding (e.g., subject–object swaps), incomplete grounding, and over-specification (e.g., irrelevant or tangential reasoning), all of which disrupt the alignment between textual reasoning and the underlying visual evidence (Chen et al., 2024).

Recent studies attempt to mitigate these issues through preference-based alignment of reasoning chains (Ouyang et al., 2022; Rafailov et al., 2023; Yu et al., 2024). Such methods generate positive and negative rationale pairs by editing tokens or final answers (Zhang et al., 2024a; Tan et al., 2025; Zheng et al., 2024; Zhou et al., 2024), encouraging coherent reasoning but leaving visual grounding untouched. As illustrated in Figure 1, token- or answer-level perturbations (e.g., replacing "skateboard" with "dog", or generating a negative rationale based on the wrong answer "yell") do not create hard, visually grounded negatives, so the model can still rely on language priors without engaging the image. Because these rationales are not tied to specific scene elements, grounding failures cannot be localized or controlled: when the model hallucinates relations, skips steps, or loosely associates objects, we cannot tell which node was misunderstood or where the reasoning chain broke. Thus, realistic grounding errors remain uncaptured.

To address these limitations, we propose SceneAlign, a scene graph–guided preference alignment framework for structure-faithful visual reasoning, where the scene graph represents a set of objects and relations that provide explicit structure to localize reasoning-critical entities. Given an image and question, SceneAlign first constructs a scene graph that encodes objects, attributes, and

[*]These authors contributed equally to this work.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17136–17150
July 2-7, 2026

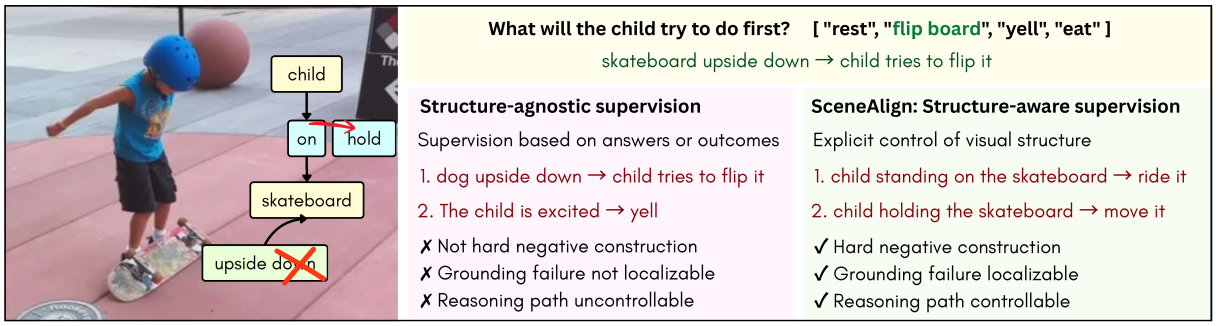

Figure 1: Motivation for structure-aware supervision. Under structure-agnostic supervision, reasoning errors that ignore, hallucinate, or loosely associate visual structure are indistinguishable at the answer level. Structure-aware supervision (SceneAlign) perturbs scene-graph elements to generate interpretable and controllable negative examples, making grounding failures observable and localizable.

relations, and generate a structure-consistent chain of thought (CoT) grounded in this graph. We then perturb the graph through controlled interventions, such as swapping, removing, or replacing nodes and edges, to simulate realistic grounding failures, and regenerate corresponding CoTs directly from these modified graphs.

For instance, as shown in Figure 1, removing a reasoning-critical node such as "upside down" or replacing the relation "on" with "hold" produces a reasoning chain that remains plausible but omits critical evidence from the scene. Guided by the scene graph, such CoTs form high-quality contrastive pairs: visually grounded CoTs as positives, and semantically coherent yet visually inconsistent CoTs as negatives, which expose a diverse set of reasoning failure modes. Finally, we filter the perturbed samples for plausibility and diversity, and apply Direct Preference Optimization (DPO) over these contrastive pairs to explicitly align the model's reasoning with the underlying scene structure, enhancing grounding fidelity in complex visual reasoning.

Our contributions are summarized as follows:

- We introduce SceneAlign, a scene-graph–guided preference alignment framework for grounding-faithful visual reasoning.
- We design controllable perturbation strategies that generate hard and semantically coherent negatives reflecting common grounding failures.
- We conduct extensive experiments across multiple visual reasoning benchmarks, showing that SceneAlign consistently improves grounding consistency, reasoning coherence, and answer accuracy over strong baselines.

2 Related Works

Recent efforts to improve multimodal reasoning span a broad range of approaches. We specifically build on negative CoT sampling in preference optimization and scene-graph–based multimodal reasoning. SceneAlign perturbs scene graphs to construct structurally inconsistent CoTs and employs preference optimization to align reasoning with the underlying visual structure.

Negative CoT Sampling and Preference Optimization. Modern alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) and DPO (Rafailov et al., 2023), have been adapted to multimodal settings to address issues such as hallucination (Yu et al., 2024). A central idea is to generate negative reasoning to provide counterfactual signals. Answer-oriented methods label rationales by final correctness (Zhang et al., 2024a; Tan et al., 2025). However, these approaches often miss structural reasoning errors since flawed reasoning can still produce correct answers. Other approaches perturb CoTs at the token level (e.g., SNSE-CoT (Zheng et al., 2024), QCRD (Wang et al., 2024), and NoRa (Zhou et al., 2024)) or treat hallucinated answers as negatives (Jiang et al., 2024; Sarkar et al., 2024). These approaches demonstrate the value of contrastive supervision, but they rarely capture structural reasoning errors. Recent work instead perturbs relational structures, such as scene graphs for multimodal reasoning (Chen et al., 2025). We extend prior work through a scene-graph–grounded framework that perturbs entities, attributes, and relations to create counterfactual but semantically plausible rationales, utilizing the structural reasoning that dominates the complex visual scenes.

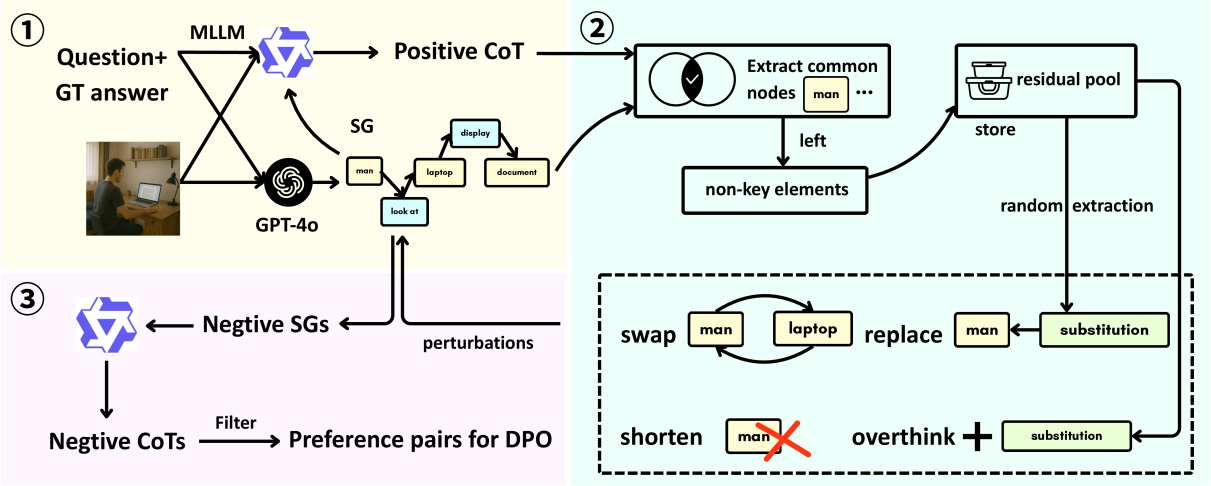

Figure 2: **Framework overview.** The SceneAlign pipeline first generates scene-graph–grounded positives (Sec. 3.1), applies four graph perturbations (*swap, replace, shorten, overthink*) to create negatives (Sec. 3.2), filters for diversity (Sec. 3.3), and finally performs DPO alignment (Sec. 3.4).

Scene Graphs for Multimodal Reasoning. Scene graph integration with large language models has emerged as a promising direction for structured visual understanding. Recent developments in graph-based reasoning reveal the potential for structured representations in complex visual tasks. (Mitra et al., 2024) utilizes generated scene graphs in zero-shot prompting to extract compositional knowledge without fine-tuning. (Lee et al., 2024) extends this approach through multimodal knowledge graphs with relation graph attention networks. Additional methods build on the scene graph CoT (Shao et al., 2024; Ji et al., 2025). Together, they show promise in helping MLLMs to maintain coherence and faithfulness when reasoning about complex scenes by injecting structured representations. Our work extends beyond prior works by integrating scene graphs into the learning process. The MLLM model is trained to prefer rationales that follow the true scene structure over those that deviate from it. Specifically, after generating positive graphs, we perturb them to obtain negative ones, and turn the pairs into reasoning trajectories to train the model for structural faithfulness. This design builds on earlier efforts in scene graph perturbation (Singh et al., 2023; Li et al., 2023; Huang et al., 2024), and leverages these perturbations to align CoT reasoning with preference supervision.

In summary, previous studies have not considered how to align the reasoning process itself with the visual scene structure. They often treat reasoning as a flat textual sequence, neglecting the graph-level dependencies among entities and relations. In contrast, our framework introduces structure-aware perturbations guided by scene graphs, allowing the model to learn from both grounded and structurally inconsistent rationales.

3 Methodology

To mitigate the inconsistency between answers and their visual grounding, we propose the SceneAlign framework that aligns multimodal reasoning with scene-level structure through preference optimization. Figure 2 provides an overview of our SceneAlign framework. To be specific, given an image-question pair, SceneAlign first constructs scene-graph–grounded positive reasoning, then perturbs scene graphs to produce negative CoTs, filters them for informativeness, and finally applies DPO (Rafailov et al., 2023) to align the model toward structure-consistent reasoning. Algorithm 1 summarizes its key steps.

3.1 Positive Reasoning Generation

To ensure that the reasoning process is grounded in the visual scene rather than driven by superficial correlations, we first construct *structure-consistent* reasoning traces that explicitly reference scene-graph elements as visual anchors. Formally, a scene graph is defined as,

$$SG = (E, A, R), \tag{1}$$

where E is the set of entities, A the set of attribute pairs, and R the set of relational triples (e_i, r, e_j).

Given a dataset D, for each image–question pair $(I, q) \in D$, we extract a positive scene graph SG_q^+

Algorithm 1: Workflow of SceneAlign

Preparation: Parser Φ (GPT-4o); perturbation operators $\mathcal{O}$ [Eq. (7)]; overlap thresholds $[\gamma_\ell, \gamma_u]$.
Input: Dataset $D = \{(I, q, a)\}$; base MLLM π_θ.
Output: Aligned parameters $\theta^\star$.

Initialize preference set $\mathcal{P} \leftarrow \emptyset$.
// Stage I: Preference Data Construction
foreach $(I, q, a) \in D$ **do**
  // (1) Positive Reasoning Generation (Sec. 3.1)
  Extract $SG_q^+ \leftarrow \Phi(I, q)$; generate $\tau_q^+ \leftarrow \pi_\theta(I, q, SG_q^+)$.
  // (2) Subgraph Extraction (Sec. 3.2)
  Identify $SG_q^c \subseteq SG_q^+$ referenced in τ_q^+.
  // (3) Negative *SG* Construction (Eqs. (2)–(6))
  $\mathcal{S}_q^- = \{\mathcal{O}_i(SG_q^c) \oplus (SG_q^+ \setminus SG_q^c) \mid \mathcal{O}_i \in \mathcal{O}\}$.
  // (4) Negative CoT Generation
  Generate $\{\tau_{q,i}^-\}$ by prompting π_θ with $(q, SG_{q,i}^-)$.
  // (5) Scene-Graph Filtering (Eq. (9))
  Compute Jaccard overlap $J(SG_{q,i}^-, SG_q^+)$; retain indices $\mathcal{I}_{\text{mid}}$ where $J \in [\gamma_\ell, \gamma_u]$.
  // (6) Diversity-based Sampling (Eq. (10))
  Select $\tilde{\mathcal{I}} = \arg\max_{\mathcal{I} \subseteq \mathcal{I}_{\text{mid}}} \min_{i \neq j} \|z_i - z_j\|_2$; obtain diverse negatives $\{\tilde{\tau}_{q,i}^-\}$.
  // (7) Preference Pair Construction
  Define context $x_q = (I, q)$; add $\{(x_q, \tau_q^+, \tilde{\tau}_{q,i}^-)\}_{i \in \tilde{\mathcal{I}}}$ to $\mathcal{P}$.
// Stage II: Preference Optimization
// (8) Scene-Graph Preference Optimization (Eq. 11)
$\theta^\star \leftarrow$ DPO-TRAIN$(\pi_\theta, \pi_{\text{ref}}, \mathcal{P})$.
return $\theta^\star$

that captures all relevant objects, attributes, and relations using GPT-4o (Hurst et al., 2024). A grounded chain-of-thought $\tau_q^+ \in \mathcal{T}$ is then generated by a MLLM π_θ, where $\mathcal{T}$ denotes the space of possible reasoning traces. By anchoring each reasoning step to elements in SG_q^+, the model is encouraged to reference concrete visual evidence, improving both interpretability and grounding reliability. This multimodal conditioning ensures that the image remains the primary authoritative signal; any minor noise in the extracted scene graph can be rectified by the MLLM by directly referencing the visual evidence during reasoning generation.

Importantly, the scene graph serves exclusively as an intermediate structure for constructing preference data. During both DPO training and inference, the model receives only the image and question as input, without requiring any external parser or structured representation.

3.2 Construction of Negative Scene Graphs

To teach the model what incorrect grounding looks like, we construct structured counterfactuals that reflect common failure patterns, including object hallucination, relation mis-grounding, incomplete grounding, and over-specification, while keeping the overall semantics coherent. Given a grounded rationale τ_q^+, we first identify its CoT-grounded subgraph $SG_q^c \subseteq SG_q^+$ that contains only the elements explicitly referenced in τ_q^+. We then construct negative graphs by perturbing elements drawn from SG_q^c while keeping the remaining part $SG_q^+ \setminus SG_q^c$ unchanged, ensuring that all modifications remain semantically relevant yet alter the structural grounding.

Inspired by previous studies on graph-based multimodal alignment (Huang et al., 2025; Li et al., 2025; Liu et al., 2025b), we introduce four structured transformations: **swap**, **replace**, **shorten**, and **overthink**, each corresponding to a distinct type of grounding error. To support replacement-based perturbations, we maintain a *residual pool* $P_{\text{res}} = SG_q^+ \setminus SG_q^c$ consisting of non-key elements from the same scene graph. Formally, for a relation $(e_i, r, e_j) \in R$ and an element $x \in SG_q^c$,

- **Swap**: exchange the subject and object of a relation,
$$T_{swap}(e_i, r, e_j) = (e_j, r, e_i). \quad (2)$$
This reflects mis-grounding, where subject and object roles are confused, producing fluent rationales that point to wrong bindings.

- **Replace**: substitute an element using a candidate from the residual pool,
$$T_{rep}(x) = x', \quad x' \in P_{\text{res}}. \quad (3)$$
Such substitutions mimic subtle grounding errors or hallucination-like substitutions, where a non-key scene element replaces a critical one while the rationale remains plausible.

- **Shorten**: remove a key element to simulate omission,
$$T_{short}(SG_q^c) = SG_q^c \setminus \{x\}, \quad x \in SG_q^c. \quad (4)$$
Eliminating key nodes or relations creates traces that skip inference steps, highlighting the need for complete evidence in reasoning chains.

- **Overthink**: augment the graph with redundant or spurious content,
$$T_{over}(SG_q^c) = SG_q^c \cup \{x'\}, \quad x' \in P_{\text{res}}. \quad (5)$$
Adding redundant grounded details corresponds to over-specification, introducing superfluous steps that distract from the core causal path and reduce reasoning faithfulness.

Applying these operators yields a collection of counterfactual scene graphs, each constructed by perturbing the CoT-grounded subgraph and reattaching it to the unperturbed remainder,

$$SG^{-}_{q,i} = \mathcal{O}_i\big(SG^{c}_{q}\big) \oplus \big(SG^{+}_{q} \setminus SG^{c}_{q}\big), \quad (6)$$

$$\mathcal{O} = \{T_{swap}, T_{rep}, T_{short}, T_{over}\}. \quad (7)$$

which induces the set,

$$\mathcal{S}^{-}_{q} = \{SG^{-}_{q,i}\}_{i=1}^{k}, \quad (8)$$

covering the major error types in complex scene reasoning and providing structured negatives for preference optimization.

A *mix* mode randomly samples among perturbations to enhance diversity, while consistency constraints guarantee well-formed outputs. This perturbation design allows us to systematically expose recurring error patterns such as hallucination, mis-grounding, skipped inference, and over-specification, so that the model learns to distinguish visual-grounded reasoning from plausible but incorrect alternatives.

3.3 Selection of Negative Reasoning Chains

The purpose of this step is to construct informative and diverse negative reasoning chains that serve as effective contrastive signals for preference alignment. Without careful selection, negative samples that are either too similar or entirely irrelevant to the positive reasoning chain would provide weak or misleading supervision, undermining the goal of aligning the model's reasoning with the true scene structure. Thus, we first perform *scene-graph filtering*. Specifically, let $U(SG) = A \cup \{(e_s, e_o) : (e_s, r, e_o) \in R\}$ denote the set of attributes and subject–object pairs extracted from a scene graph SG, and compute the Jaccard overlap $J(SG^{-}_{q,i}, SG^{+}_{q})$ between each negative graph and the gold graph,

$$J\Big(SG^{-}_{q,i},\, SG^{+}_{q}\Big) = \frac{\big|U\big(SG^{-}_{q,i}\big) \cap U\big(SG^{+}_{q}\big)\big|}{\big|U\big(SG^{-}_{q,i}\big) \cup U\big(SG^{+}_{q}\big)\big|}. \quad (9)$$

We retain indices

$$\mathcal{I}_{\text{mid}} = \big\{i : \gamma_\ell \le J(SG^{-}_{q,i}, SG^{+}_{q}) \le \gamma_u\big\},$$

yielding a filtered set $\widehat{\mathcal{S}}^{-}_{q} = \{SG^{-}_{q,i}\}_{i \in \mathcal{I}_{\text{mid}}}$. This filtering yields hard negatives by excluding near-duplicates (γ_u) and irrelevant outliers (γ_ℓ). Combined with localized perturbations (Sec. 3.2), these samples remain semantically consistent with positives yet contain precise grounding errors, forcing DPO to prioritize subtle visual-logical distinctions over global semantic shifts.

Next, we apply *diversity-based sampling*. Let $z_i = f_\phi(\tau^{-}_{q,i}) \in \mathbb{R}^d$ denote the embedding of the negative CoT generated from $SG^{-}_{q,i}$. We then select m diverse negatives via a max–min criterion:

$$\tilde{\mathcal{I}} = \arg \max_{\substack{\mathcal{I} \subseteq \mathcal{I}_{\text{mid}} \\ |\mathcal{I}| = m}} \min_{\substack{i \neq j \\ i,j \in \mathcal{I}}} \|z_i - z_j\|_2. \quad (10)$$

This produces the final negative graph set $\tilde{\mathcal{S}}^{-}_{q} = \{SG^{-}_{q,i}\}_{i \in \tilde{\mathcal{I}}}$ and their associated CoTs $\tilde{\tau}^{-}_{q} = \{\tilde{\tau}^{-}_{q,i}\}_{i \in \tilde{\mathcal{I}}}$. By combining filtering with diversity control, we obtain a challenging yet balanced set of counterfactuals that provide rich training signals and prevent overfitting to narrow error modes.

3.4 Scene-Graph Preference Optimization

To ensure that the model internalizes a preference for scene-faithful reasoning, we optimize it with Direct Preference Optimization (DPO) (Rafailov et al., 2023) using positive and negative chains derived from scene graphs. This encourages the model to assign higher likelihoods to reasoning trajectories that are consistent with the underlying scene structure while penalizing inconsistent ones.

Formally, let $x_q = (I, q)$ denote the model input and let $\mathcal{P} = \{(x_q, \tau^{+}_{q}, \tilde{\tau}^{-}_{q,i})\}$ be the preference pairs with respect to a reference model π_{ref}. Note that although SG^{+}_{q} guides the construction of τ^{+}_{q} and $\tilde{\tau}^{-}_{q,i}$, it is not included in the model input during optimization. DPO optimizes

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{|\mathcal{P}|} \sum_{(x,\tau^{+},\tau^{-}) \in \mathcal{P}} \log \sigma\Big(\beta\big[\log \pi_\theta(\tau^{+} \mid x) - \log \pi_\theta(\tau^{-} \mid x) - \big(\log \pi_{\text{ref}}(\tau^{+} \mid x) - \log \pi_{\text{ref}}(\tau^{-} \mid x)\big)\big]\Big),$$

where σ is the logistic function and $\beta > 0$ controls preference sharpness. The final preference dataset comprises positives $\{\tau^{+}\}$ and selected negatives $\{\tilde{\tau}^{-}\}$. DPO trains the MLLM π_θ to prefer reasoning grounded in positive scene graphs, yielding the aligned parameters $\theta^{\star}$. This training objective encourages the model to produce correct answers while maintaining structural faithfulness, thereby improving reasoning across complex scenes.

4 Experiments

4.1 Experimental Settings

Models. We apply SceneAlign to five representative MLLMs: Qwen2.5-VL-3B (Bai et al., 2025b),

Model	Method	MME-RW	EMMA	ScienceQA	MMMU	Hallusion-Bench			GQA	SeedBench	
		Score (↑)	Score (↑)	Acc. (↑)	Acc. (↑)	aAcc (↑)	fAcc (↑)	qAcc (↑)	Exact (↑)	All (↑)	Img (↑)
Qwen2.5-VL-3B	Base	38.32	20.00	83.19	47.11	34.28	17.05	20.66	69.40	70.94	74.88
	SFT	39.81	22.75	83.26	48.09	53.28	28.15	28.36	**69.60**	70.30	74.83
	SceneAlign	**40.42**	**23.25**	**85.10**	**50.70**	**58.25**	**34.59**	**33.64**	**69.60**	**72.09**	**75.60**
Qwen3-VL-4B	Base	47.99	22.75	91.49	64.26	57.65	34.28	34.20	62.40	74.80	76.93
	SFT	47.36	24.00	91.53	66.33	62.10	39.45	39.20	**63.00**	74.55	77.12
	SceneAlign	**48.86**	**25.25**	**92.72**	**66.67**	**65.35**	**41.05**	**41.78**	**63.00**	**77.03**	**79.89**
Qwen2.5-VL-7B	Base	43.98	17.75	88.71	51.11	57.31	33.24	32.09	**71.80**	74.07	77.43
	SFT	44.45	21.00	88.80	51.53	60.34	36.72	35.84	71.40	75.17	78.46
	SceneAlign	**45.02**	**22.75**	**89.12**	**52.83**	**62.88**	**37.57**	**38.02**	71.40	**76.54**	**79.12**
InternVL3-8B	Base	49.60	21.00	**90.77**	60.51	49.58	26.03	24.45	60.20	70.72	73.67
	SFT	50.33	22.25	90.02	**60.60**	57.84	32.96	31.42	60.40	72.10	75.05
	SceneAlign	**51.94**	**23.25**	90.15	60.37	**61.72**	**36.88**	**35.01**	**60.80**	**73.45**	**76.22**
LLaVA-Next-8B	Base	37.26	15.75	80.12	40.67	26.81	18.79	12.75	73.00	57.58	72.73
	SFT	38.95	17.00	81.99	41.20	39.35	20.02	14.74	73.00	57.71	72.80
	LLaVA-Reasoner	37.68	14.00	81.10	42.44	41.85	20.81	12.53	72.80	58.12	73.52
	AoT	37.68	16.75	81.25	40.89	41.64	20.23	14.95	73.00	57.58	72.74
	SceneAlign	**39.42**	**19.50**	**83.25**	**42.85**	**46.89**	**24.57**	**18.02**	**73.40**	**59.80**	**74.60**

Table 1: Performance on **reasoning-oriented** benchmarks. Higher is better (↑). Best results are in bold. Comparison of different methods for constructing CoT-based preference pairs. All values are in percentage form.

Qwen3-VL-4B(Bai et al., 2025a), Qwen2.5-VL-7B (Bai et al., 2025b), InternVL3-8B(Zhu et al., 2025), and LLaVA-Next-8B (Li et al., 2024b). The three Qwen models primarily differ in scale and vision–language coupling. InternVL3-8B integrates a strong vision encoder with tight cross-modal fusion, while LLaVA-Next-8B combines CLIP with an LLaMA3-8B backbone (Dubey et al., 2024) and serves as a standard baseline.

Training Data. We construct training data from A-OKVQA (Schwenk et al., 2022) by pairing each image–question instance with a scene-graph–consistent positive CoT and multiple perturbed negatives. A-OKVQA is selected for its challenging open-ended questions requiring both visual grounding and external knowledge, making it a widely adopted benchmark for multimodal reasoning. Negatives are generated through four structured perturbations, then filtered and diversified to form DPO preference pairs. In all experiments, we use 3 negatives per instance as the default setting.

Scene Graph Generation. To construct scene graphs for each image-question pair, we use GPT-4o to generate structured scene representations conditioned on both the visual input and the corresponding question. The model is prompted to output entities, attribute pairs, and relational triples in a consistent JSON format, ensuring fine-grained grounding of objects and relations relevant to reasoning. The detailed generation prompt and examples are provided in the Appendix A.2.

Baselines. We compare SceneAlign against the pretrained model and an SFT baseline (fine-tuned on SceneAlign's positive CoTs) to isolate the gains from preference alignment. We also compare SceneAlign with two prior methods, **AoT** (Tan et al., 2025) and **LLaVA-Reasoner** (Zhang et al., 2024a), for constructing CoT-based preference pairs on A-OKVQA, using **LLaVA-Next-8B** as the backbone since prior work conducted their evaluations on this model. AoT reformulates each question–answer pair into an answer-conditioned reasoning task to generate contrastive rationales that logically support or contradict the answer. LLaVA-Reasoner adopts a two-stage pipeline that first augments short-answer data with GPT-generated reasoning chains and then constructs outcome-based preference pairs optimized via DPO. We compare SceneAlign against these approaches as representative *negative CoT construction methods*: AoT and LLaVA-Reasoner generate textual negatives through answer or outcome perturbations, while SceneAlign produces **structurally perturbed** negatives guided by scene graphs, enabling contrastive alignment at the level of visual grounding.

Evaluation Benchmarks. We evaluate SceneAlign on a suite of benchmarks that jointly assess its visual grounding and complex scene reasoning capabilities. For visual grounding, MME-RealWorld (Zhang et al., 2024b) tests perception and grounding under real-world scenes, while GQA (Hudson and Manning, 2019) examines compositional question answering requiring precise grounding of objects and relations. For com-

Method	MME-RW	EMMA	ScienceQA	MMMU	Hallusion-Bench			GQA	SeedBench	
	Score (↑)	Score (↑)	Acc. (↑)	Acc. (↑)	aAcc (↑)	fAcc (↑)	qAcc (↑)	Exact (↑)	All (↑)	Img (↑)
Base	38.32	20.00	83.19	47.11	34.28	17.05	20.66	69.40	70.94	74.88
w/o swap	40.13	19.25	83.55	47.60	57.52	33.82	32.09	69.40	71.30	75.35
w/o replace	39.60	19.00	83.70	49.65	57.20	33.24	32.53	69.60	71.65	75.50
w/o shorten	39.86	**23.25**	84.00	48.85	58.04	34.39	33.41	69.60	71.80	75.65
w/o overthink	39.55	20.75	83.45	47.90	57.75	34.02	33.23	69.20	72.05	**75.75**
entity	39.76	22.75	84.31	47.57	55.84	30.90	30.13	69.40	71.48	75.42
relations	39.76	21.25	83.19	49.00	56.62	32.35	29.91	**69.60**	**72.16**	**75.75**
attributes	39.92	19.50	83.35	46.92	55.42	32.35	29.69	69.40	71.72	75.58
SceneAlign	**40.42**	**23.25**	**85.10**	**50.70**	**58.25**	**34.59**	**33.64**	**69.60**	72.09	75.60

Table 2: Ablation study of structured perturbations on reasoning-oriented benchmarks using **Qwen2.5-VL-3B**. The relative performance order of each variant across benchmarks is preserved, with consistent overall growth. All values are in percentage form. Higher is better (↑). Best results are in bold.

plex scene reasoning, EMMA-mini (Hao et al., 2025), SEEDBench (Li et al., 2024a), Hallusion-Bench (Guan et al., 2024), ScienceQA (Saikh et al., 2022), and MMMU-Reasoning (Yue et al., 2024) evaluate multi-step reasoning, hallucination robustness, and structure-dependent inference across multimodal inputs. All evaluations are performed using the LMMs-Eval framework (Zhang et al., 2025) for consistency and reproducibility. At test time, all models operate as standard MLLMs, receiving only the image and question as input without any scene graph or external parser.

Implementation Details. To ensure reproducibility, we generated all positive scene graphs using the gpt-4o-2024-11-20 snapshot with a temperature of 0 and fixed random seeds. Training was conducted on 2 NVIDIA A100 GPUs. We set the preference optimization coefficient $\beta = 0.1$ and learning rate 5×10^{-6}, training for 1 epoch. We employed a per-device batch size of 6 with 20 gradient accumulation steps, resulting in an effective batch size of 120. The maximum input length is 4096 tokens, with prompts truncated to 2048 tokens. All hyperparameters were kept consistent across all models to ensure a fair comparison. Further details on training efficiency are provided in Appendix A.4.

4.2 Main Result

Table 1 reports results across reasoning-oriented and vision-centric benchmarks for different MLLMs of varying scales and architectures. Across all settings, SceneAlign consistently outperforms both the pretrained and SFT baselines, showing that scene graph–guided preference alignment provides benefits beyond positive-CoT supervision, where structure-aware negative samples supply a contrastive grounding signal that discourages incorrect or poorly grounded reasoning paths. Gains are most pronounced on reasoning-intensive benchmarks such as EMMA, HallusionBench, and MMMU-Reasoning, where SceneAlign yields average improvements of 3%–5%. These tasks require step-wise compositional reasoning and multi-entity grounding, highlighting that our perturbation strategy exposes fine-grained, structure-level grounding errors that standard textual perturbations overlook. Notably, even lightweight models such as Qwen2.5-VL-3B and Qwen3-VL-4B achieve clear gains (e.g., +23.97% and +7.70% respectively on HallusionBench), demonstrating that structural grounding is effective even without large capacity. Larger models (e.g., Qwen2.5-VL-7B, InternVL3-8B) show similar benefits, implying that SceneAlign provides orthogonal improvements to model capacity. On vision-centric benchmarks like GQA, SeedBench, and MME-RealWorld, SceneAlign also delivers consistent gains. We further compare SceneAlign with two prior CoT-based preference construction approaches, LLaVA-Reasoner and AoT, using the LLaVA-Next-8B backbone. Results illustrate that SceneAlign surpasses both across nearly all metrics, with particularly large margins on HallusionBench and EMMA, confirming that grounding-aware contrastive pairs provide stronger supervision than text-only perturbations. Overall, these results demonstrate that SceneAlign is a general and scalable framework for improving multimodal reasoning faithfulness through structure-aware preference alignment.

4.3 Ablation Studies

Effect of Perturbation Operators. Table 2 ablates the four perturbation operators: *swap*, *replace*, *shorten*, and *overthink*. Each operator corresponds to a distinct failure mode (role mis-binding, hallucination, skipped inference, and over-specification). Removing any operator reduces performance on the benchmarks most sensitive to that error type. Full SceneAlign, which integrates all four, achieves the greatest and most balanced improvements.

Effect of Scene Graph Elements. Table 2 further ablates the contribution of different scene-graph elements by selectively perturbing entities, relations, and attributes. Relation-level grounding brings the largest performance gains, showing that correctly binding subjects and objects is most critical for grounded reasoning. Entity information also plays a major role, confirming the importance of identifying the right participants. Attribute grounding has a smaller but still consistent effect, suggesting that fine-grained properties provide complementary cues beyond structural role assignment.

Effect of Negative Sampling Strategies. Table 4 in the appendix compares three sampling variants: random negatives, diversity-only sampling, and our full method (scene-graph filtering followed by diversity sampling). Random negatives yield weak supervision due to trivial or irrelevant samples, while removing graph filtering introduces overly close or overly distant negatives. Our full method consistently achieves the highest performance, highlighting the value of carefully selecting medium-difficulty negatives.

Effect of Overlap Thresholds. To study how scene-graph overlap thresholds affect preference construction, we vary the lower bound τ_ℓ and upper bound τ_u. Overly narrow ranges filter out many candidate pairs, reducing data coverage and causing moderate accuracy drops, while overly loose ranges introduce trivial or overly distant negatives that weaken supervision. We find that moderate thresholds strike the best balance; the default SceneAlign setting ($\tau_\ell = 0.3$, $\tau_u = 0.7$) consistently yields strong and well-balanced performance on MMMU-Reasoning and SEED-Bench (Figure 3).

Effect of the Number of Negatives. Table 4 studies the impact of using different numbers of negatives per instance. One negative provides limited coverage of reasoning errors, while three negatives

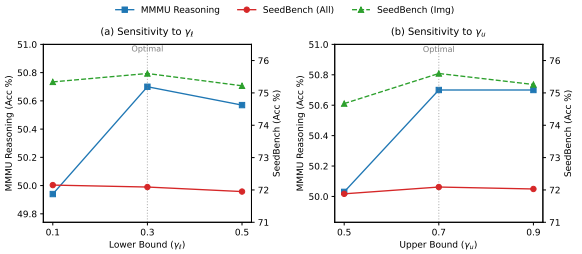

Figure 3: Sensitivity to Jaccard bounds γ_ℓ and γ_u, Performance shows an inverted U-shape with optimal performance at $\gamma_\ell = 0.3$ and $\gamma_u = 0.7$.

significantly improve structural supervision with modest additional cost. Larger numbers offer diminishing returns. We adopt three negatives per instance by default as the best trade-off between efficiency and effectiveness.

4.4 Case Study

To better illustrate how scene-graph grounding improves reasoning faithfulness, we present a case study in Figure 4. The example asks: *"What kind of activity with respect to the motorcycle is the man on the floor most likely engaging in?"* The scene contains several interacting entities, including the man, the motorcycle, and a piece of paper, which calls for structured visual reasoning that goes beyond surface-level cues.

Positive CoT. Our model correctly grounds each reasoning step in the scene graph: the man is on the ground next to the motorcycle, others are gathered around, and one holds a paper while looking toward the motorcycle, which jointly supports *inspecting or diagnosing* the motorcycle.

Prior Methods. Answer-based approaches fix an incorrect final answer (e.g., *riding*) and fabricate reasoning to justify it, leading to fluent but semantically inconsistent CoTs. Token-level perturbations merely alter surface text and often yield logically broken scenes (e.g., "arguing with the tree"), which is grammatically correct yet structurally invalid.

SceneAlign (Ours). In contrast, SceneAlign perturb the **scene graph structure** through four controlled operations: *swap*, *replace*, *shorten*, and *overthink*. Each produces a natural yet structurally inconsistent CoT, revealing distinct failure modes such as role misalignment, hallucination, missing evidence, and over-specification. These structured negatives expose reasoning drift that purely text-based perturbations fail to capture. Full CoT details

Method	Grounding Score (↑)
Pretrained	1.997
SFT	2.215
SceneAlign	**2.832**

Table 3: Step-level grounding faithfulness on A-OKVQA using GPT-4o evaluation. Each reasoning step is rated on a 0–5 scale for visual evidence support.

are provided in Appendix A.3.

4.5 Direct Faithfulness Evaluation

Beyond answer accuracy, we directly evaluate whether SceneAlign improves step-level reasoning faithfulness. Following recent work on automated faithfulness assessment (Balasubramanian et al., 2025; Moll et al., 2025; Lv et al., 2026), we use GPT-4o as an external evaluator under a fixed rubric. For each reasoning trajectory generated by Qwen2.5-VL-3B-Instruct on the A-OKVQA test set, every numbered reasoning step is rated on a 0–5 scale based on whether it is supported by visible evidence in the image. We compute per-sample averages and report the overall mean grounding score. As shown in Table 3, SceneAlign achieves a grounding score of 2.832, improving over the SFT baseline by +0.617 and over the pretrained model by +0.835. This result confirms that SceneAlign not only improves final answer accuracy but also enhances the visual grounding of intermediate reasoning steps, reducing unsupported claims in the chain of thought.

5 Conclusion

We presented SceneAlign, a scene-graph–guided preference alignment framework that improves grounding-faithful reasoning in multimodal large language models. By generating semantically coherent yet structurally inconsistent rationales through controlled scene-graph perturbations, SceneAlign addresses reasoning-grounding inconsistency and explicitly aligns chain-of-thought reasoning with visual structure. Experiments across seven benchmarks show consistent 3–5% gains on reasoning-intensive tasks, confirming that structure-aware supervision yields significantly more faithful reasoning than traditional text-based perturbations.

Limitations

Our study focuses on single-image reasoning and does not extend to multi-image or video-based inputs, where temporal and cross-view consistency introduce additional grounding challenges. Moreover, we rely on GPT-based models for generating chains of thought and for automatic evaluation of reasoning coherence and hallucination rates, which may not fully capture fine-grained human judgments. Future work could explore dynamic or temporal scene graphs and adopt human-centered evaluation protocols to provide a more comprehensive assessment of grounding fidelity in open-ended multimodal reasoning.

References

Syeda Nahida Akter, Sangwu Lee, Yingshan Chang, Yonatan Bisk, and Eric Nyberg. 2024. Visreas: Complex visual reasoning with unanswerable questions. *arXiv preprint arXiv:2403.10534*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

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

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. 2025. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. *arXiv preprint arXiv:2505.23945*.

Guoqing Chen, Fu Zhang, Jinghao Lin, Chenglong Lu, and Jingwei Cheng. 2025. Rrhf-v: Ranking responses to mitigate hallucinations in multimodal large language models with human feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6798–6815.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. 2024. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418.

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matt Botvinick. 2021. Attention over learned object embeddings enables complex visual reasoning. *Advances in neural information processing systems*, 34:9112–9124.

Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. 2023. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *arXiv preprint arXiv:2311.01487*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

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.

Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. 2025. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*.

Xin Huang, Ruibin Li, Tong Jia, Wei Zheng, and Ya Wang. 2025. Visual perturbation and adaptive hard negative contrastive learning for compositional reasoning in vision-language models. *arXiv preprint arXiv:2505.15576*.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, and 1 others. 2024. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 2417–2425.

Drew A Hudson and Christopher D Manning. 2019. 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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Binbin Ji, Siddharth Agrawal, Qiance Tang, and Yvonne Wu. 2025. Enhancing spatial reasoning in vision-language models via chain-of-thought prompting and reinforcement learning. *arXiv preprint arXiv:2507.13362*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. 2024. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seedbench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, and 1 others. 2024c. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, and Long Chen. 2023. Compositional feature augmentation for unbiased scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21685–21695.

Yuting Li, Lai Wei, Kaipeng Zheng, Jingyuan Huang, Linghe Kong, Lichao Sun, and Weiran Huang. 2025. Vision matters: Simple visual perturbations can boost multimodal math reasoning. *arXiv preprint arXiv:2506.09736*.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025a. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Preprint*, arXiv:2505.21523.

Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Ding Wang, and Botian Shi. 2025b. Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 981–992.

Weijiang Lv, Yaoxuan Feng, Xiaobo Xia, Jiayu Wang, Yan Jing, Wenchao Chen, and Bo Chen. 2026. Spdfaith bench: Diagnosing and improving faithfulness in chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2602.07833*.

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

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Johannes Moll, Markus Graf, Tristan Lemke, Nicolas Lenhart, Daniel Truhn, Jean-Benoit Delbrouck,

Jiazhen Pan, Daniel Rueckert, Lisa C Adams, and Keno K Bressem. 2025. Evaluating reasoning faithfulness in medical vision-language models using multimodal perturbations. *arXiv preprint arXiv:2510.11196*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. *arXiv preprint arXiv:2405.18654*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. *arXiv preprint arXiv:2305.13812*.

Wentao Tan, Qiong Cao, Yibing Zhan, Chao Xue, and Changxing Ding. 2025. From answers to rationales: Self-aligning multimodal reasoning with answer-oriented chain-of-thought. *arXiv preprint arXiv:2507.02984*.

Wei Wang, Zhaowei Li, Qi Xu, Yiqing Cai, Hang Song, Qi Qi, Ran Zhou, Zhida Huang, Tao Wang, and Li Xiao. 2024. Qcrd: Quality-guided contrastive rationale distillation for large language models. *arXiv preprint arXiv:2405.13014*.

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

Zuopeng Yang, Daqing Liu, Chaoyue Wang, Jie Yang, and Dacheng Tao. 2022. Modeling image composition for complex scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7764–7773.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Jiachen Yu, Yufei Zhan, Ziheng Wu, Yousong Zhu, Jinqiao Wang, and Minghui Qiu. 2025. Vfaith: Do large multimodal models really reason on seen images rather than previous memories? *Preprint*, arXiv:2506.11571.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. 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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and 1 others. 2025. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 881–916.

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

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, and 1 others. 2024b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

Guangmin Zheng, Jin Wang, Xiaobing Zhou, and Xuejie Zhang. 2024. Enhancing semantics in multimodal chain of thought via soft negative sampling. *arXiv preprint arXiv:2405.09848*.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. 2024. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? *Advances in Neural Information Processing Systems*, 37:123846–123910.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Appendix

A.1 Analysis of Data Reliability

To validate the quality of our structural interventions and positive rationales, we conducted a human spot check on 150 randomly sampled A-OKVQA instances. Our evaluation confirms that the gpt-4o-generated scene graphs correctly capture all question-relevant entities and relations in all of the cases. While occasional redundant nodes (e.g., background objects) appear, they do not degrade the quality of the positive CoTs, as the MLLM conditions primarily on the *(image, question)* pair and treats the scene graph as auxiliary guidance. This effectively bypasses extraction noise and underscores SceneAlign 's high-fidelity data synthesis.

A.2 Prompt Design

We use structured prompts (Figure 5, 6, 7) to decouple scene grounding from reasoning. A scene-graph prompt first extracts entities, attributes, and relations in strict JSON format to anchor reasoning to visual structure. We then generate positive CoTs using both the image and graph, and negative CoTs using only the graph, enabling controlled contrast in structure consistency.

A.3 Case Study Explanation

We present a case study (Table 5) showing that SceneAlign yields structure-consistent positive CoTs, while perturbations expose distinct failure modes.

A.4 Training Curves

We visualize four reward-related training curves (Figure 8) for five MLLMs to validate our method.

A.5 LLM Usage

Large language models were used exclusively to improve grammar and make minor wording edits in parts of this paper.

Figure 4: Example Image from Case Study.

Method	MME-RW	EMMA	ScienceQA	MMMU	Hallusion-Bench			GQA	SeedBench	
	Score (↑)	Score (↑)	Acc. (↑)	Acc. (↑)	aAcc (↑)	fAcc (↑)	qAcc (↑)	Exact (↑)	All (↑)	Img (↑)
Base	43.98	17.75	88.71	51.11	57.31	33.24	32.09	**71.80**	74.07	77.43
Random	44.22	19.25	88.79	51.48	58.15	32.08	33.19	71.40	74.25	77.60
w/o scene-graph filter	44.68	22.00	88.83	51.05	62.15	36.99	38.02	71.60	75.02	78.10
num=1	44.29	20.00	88.80	51.52	58.15	32.65	32.97	**71.80**	74.45	77.72
num=2	44.97	21.50	88.85	51.86	61.83	36.13	37.36	71.00	75.38	78.26
num=4	44.89	22.25	88.91	52.24	**62.88**	36.94	36.76	71.60	75.95	78.64
SceneAlign (num=3)	**45.02**	**22.75**	**89.12**	**52.83**	**62.88**	**37.57**	**38.02**	71.40	**76.54**	**79.12**

Table 4: Ablation on negative CoT sampling strategies using **Qwen2.5-VL-7B**. All values are in percentage form. Higher is better (↑). Best results are in bold.

Structured Scene Graph Generation Prompt

You are given an image and its associated question. Your task is to generate a scene graph in *strict JSON format* that includes the following three fields:

1. `"entity"`: a list of all objects and concepts relevant to answering the question.
2. `"attribute pairs"`: a list of `[object, attribute]` pairs describing each entity's key features (e.g., color, size, state).
3. `"relationships"`: a list of `[subject, relation, object]` triples describing spatial or semantic relationships.

Format Example:

```
{
  "entity": ["man", "motorcycle", "paper", "ground"],
  "attribute pairs": [
    ["motorcycle", "silver"],
    ["paper", "white"],
    ["ground", "paved"]
  ],
  "relationships": [
    ["man", "look at", "motorcycle"],
    ["man", "crouch on", "ground"],
    ["man", "hold", "paper"],
    ["motorcycle", "stand on", "ground"]
  ]
}
```

Attention:

1. Only return a valid JSON object with the three required fields.
2. Do **not** include any explanations or natural language text.
3. Ensure the format strictly matches the example above.

Question: `{question}`, `{ground-truth answer}`

Scene Graph:

Figure 5: Prompt for structured scene graph generation.

Positive CoT Generation Prompt

You are given a scene graph and its associated question and image. Your task is to provide step-by-step reasoning to answer the question based on the image and scene graph. Do not mention the data source. Treat the scene graph elements as the visual scene itself.

Format Example:

```
1. ...
2. ...
3. ...
4. ...
Conclusion: ...
```

Scene Graph: `{scene_graph}`

Question: `{question}`, `{ground-truth answer}`

Step-by-step reasoning:

Figure 6: Prompt for generating positive chain-of-thought reasoning from image and its corresponding scene graph.

Negative CoT Generation Prompt

You are given a scene graph and its associated question. Your task is to provide step-by-step reasoning to answer the question based on the scene graph. Do not mention the data source. Treat the scene graph elements as the visual scene itself.

Format Example:

```
1. ...
2. ...
3. ...
4. ...
Conclusion: ...
```

Scene Graph: `{scene_graph}`

Question: `{question}`

Step-by-step reasoning:

Figure 7: Prompt for generating negative chain-of-thought reasoning from a structured scene graph.

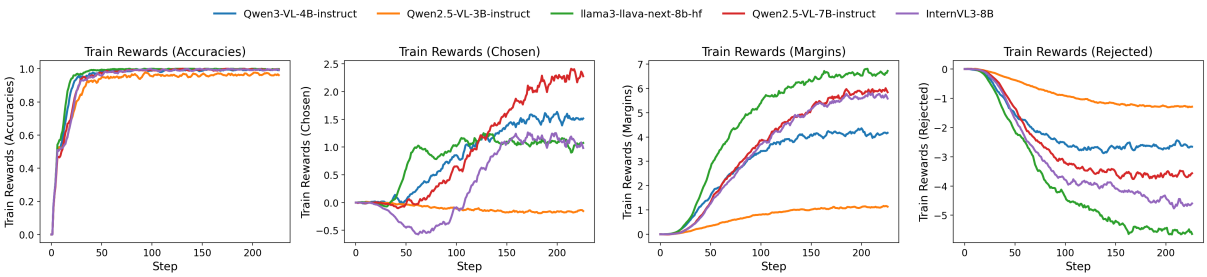

Figure 8: Training curves of five multimodal models (Qwen2.5-VL-3B-Instruct, Qwen3-VL-4B-Instruct, Qwen2.5-VL-7B-Instruct, InternVL3-8B, and Llama3-LLaVA-Next-8B-hf) across four reward-related metrics: accuracies, chosen, margins, and rejected, evaluated over 1 epoch.

Case Study: Scene-Graph–Grounded Reasoning	**Question:** What kind of activity with respect to the motorcycle is the man on the floor most likely engaging in?
Image	Figure 4
Ground Truth / MLLM Answer	**Ground Truth:** Inspecting / Diagnosing **MLLM Answer:** Inspecting. (correct)
Scene Graph	{ "entity": ["man", "motorcycle", "ground", "paper", "building", "window", "car"], "attribute pairs": [["motorcycle", "silver"], ["motorcycle", "parked"], ["ground", "paved"], ["paper", "white"], ["building", "white"], ["window", "glass"], ["car", "parked"]], "relationships": [["man", "look at", "motorcycle"], ["man", "crouch on", "ground"], ["man", "hold", "paper"], ["motorcycle", "stand on", "ground"], ["building", "behind", "motorcycle"], ["car", "behind", "motorcycle"]]}
CoT-grounded Subgraph	{ "entity": ["man","motorcycle","paper","ground"], "attribute pairs": [["motorcycle","silver"],["motorcycle", "parked"],["paper","white"],["ground","paved"]], "triples": [["man","look at","motorcycle"],["man","crouch on","ground"], ["man","hold","paper"],["motorcycle","stand on","ground"]] }
Positive CoT	1. The man is on the floor next to the motorcycle, not sitting on it. 2. Several people are gathered around the motorcycle, suggesting a repair or inspection setting. 3. One person holds a piece of paper, likely a manual or document. 4. The man's position implies he is interacting with the motorcycle. **Conclusion:** He is most likely inspecting or performing maintenance on the motorcycle.
Prior: Answer-Driven	1. The question already frames the situation around a motorcycle, which naturally suggests a riding scenario. 2. In everyday situations, people who interact with motorcycles are most often using them rather than performing technical diagnostics. 3. Without explicit evidence of a repair task, it is reasonable to treat the activity as related to normal use. **Conclusion:** So the man is more likely preparing for or engaging in a riding activity, instead of diagnosing the motorcycle.
Prior: Token-level perturb	1. The man is on the floor next to the banana, not sitting or riding on it. 2. Several people are gathered around the sky, suggesting a repair or inspection of clouds. 3. One person holds a dog, likely a manual or document. 4. The man's position and the scene imply he is arguing with the tree. **Conclusion:** He is most likely cooking dinner, not maintaining the motorcycle.
SceneAlign swap	*Change:* ["man","look at","motorcycle"] → ["motorcycle","look at","man"] 1. A motorcycle looking at the man, instead of the man looking at it. 2. Since the motorcycle is described as the active observer, the man is no longer directly interacting with it. 3. The man is sitting on the paved ground, holding a piece of white paper. 4. Without a clear connection between the man and the motorcycle, his activity is interpreted as passive or unrelated to the bike. **Conclusion:** The man is most likely just sitting and reading the paper, not doing anything with the motorcycle.
SceneAlign replace	*Change:* Replace the entity "paper" with a randomly selected entity from the residual pool ("window"). 1. The man is crouching on the paved ground near a silver motorcycle. 2. He is described as holding a white window in his hands. 3. The man is looking toward the motorcycle while holding the window, but there is no clear indication that he is using the window for repair or maintenance. 4. Instead, the window appears to be an unrelated object that he is simply handling while being near the motorcycle. **Conclusion:** The man is most likely handling or examining the window, rather than engaging in an activity related to the motorcycle.
SceneAlign shorten	*Change:* delete a main entity "man" and all attributes and relationships the node involved with. 1. The scene shows a silver motorcycle standing on a paved ground. 2. Paved ground is typically suitable for riding. 3. Since bikes are usually used for riding or recreation, the setting suggests a riding activity. **Conclusion:** The man on the floor is most likely engaged in riding the motorcycle.
SceneAlign overthink	*Change:* Randomly select and add a relationship (["building", "behind", "motorcycle"]) from the residual pool. 1. The scene shows a silver motorcycle standing on a paved ground, with a building located behind it. 2. A man is crouching on the ground, holding a white piece of paper and looking toward the motorcycle. 3. The explicit mention of the surrounding building introduces additional contextual structure, suggesting the motorcycle may be in a service, maintenance, or official facility area rather than a casual setting. 4. This additional environmental cue may bias the interpretation toward more formal or professional use of the motorcycle, rather than simple recreation. **Conclusion:** The man is most likely engaged in an inspection or diagnostic activity related to the motorcycle within a structured environment.

Table 5: **Case Study: Scene-Graph–Grounded Reasoning.** Structured comparison of grounded reasoning, showing how each perturbation (*swap, replace, shorten, overthink*) leads to distinct failure modes while the positive CoT remains consistent with the scene graph.