

← Go to **ACL ARR 2026 January** homepage (</group?id=aclweb.org/ACL/ARR/2026/January>)

3 Versions ▾

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



Chuhan Wang (/profile?id=~Chuhan_Wang3),
Xintong Li (/profile?id=~Xintong_Li2),
Jennifer Yuntong Zhang (/profile?id=~Jennifer_Yuntong_Zhang1),
Junda Wu (/profile?id=~Junda_Wu1),
Chengkai Huang (/profile?id=~Chengkai_Huang1),
Lina Yao (/profile?id=~Lina_Yao2), *Julian McAuley* (/profile?id=~Julian_McAuley1),
Jingbo Shang (/profile?id=~Jingbo_Shang2)

05 Jan 2026 (modified: 17 Mar 2026) ACL ARR 2026 January Submission
 January, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers
 Revisions (</revisions?id=WV53Yt6Zzr>) CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/>)

Keywords: Multimodal Alignment, Scene Graph, Reasoning

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.

Paper Type: Long

Research Area: Multimodality and Language Grounding to Vision, Robotics and Beyond

Research Area Keywords: cross-modal application, multimodality

Contribution Types: NLP engineering experiment

Languages Studied: English

Previous URL: </forum?id=w694ryS46f> (</forum?id=w694ryS46f>)

Explanation Of Revisions PDF: pdf (/attachment?id=WV53Yt6Zzr&name=explanation_of_revisions_PDF)

Reassignment Request Area Chair: No, I want the same area chair from our previous submission (subject to their availability).

Reassignment Request Reviewers: No, I want the same set of reviewers from our previous submission (subject to their availability)

A1 Limitations Section: This paper has a limitations section.

A2 Potential Risks: N/A

B Use Or Create Scientific Artifacts: Yes

B4 Data Contains Personally Identifying Info Or Offensive Content: N/A

B6 Statistics For Data: Yes

B6 Elaboration: Section 4.1

C Computational Experiments: Yes

C2 Experimental Setup And Hyperparameters: Yes

C2 Elaboration: Section 4.1

C3 Descriptive Statistics: N/A

D Human Subjects Including Annotators: No

D1 Instructions Given To Participants: N/A

D2 Recruitment And Payment: N/A

D3 Data Consent: N/A

D4 Ethics Review Board Approval: N/A

E Ai Assistants In Research Or Writing: Yes

E1 Information About Use Of Ai Assistants: Yes

E1 Elaboration: Appendix A.5

Author Submission Checklist: yes

Preprint: no

Preprint Status: We plan to release a non-anonymous preprint in the next two months (i.e., during the reviewing process).

Preferred Venue: ACL

Consent To Share Data: yes

Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.

Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree

B1 Cite Creators Of Artifacts: Yes

B2 Discuss The License For Artifacts: N/A

B3 Artifact Use Consistent With Intended Use: N/A

B5 Documentation Of Artifacts: N/A

C1 Model Size And Budget: Yes

C1 Elaboration: Section 4.1

C4 Parameters For Packages: N/A

D5 Characteristics Of Annotators: N/A

B1 Elaboration: Section 4.1

Submission Number: 8046

Discussion (?id=WV53Yt6Zzr#discussion)

Filter by reply type...

Filter by author...

Search keywords...

Sort: Newest First



Everyone

Program Chairs

Submission8046...

Submission8046 Area...

11 / 11 replies shown

Submission8046 Authors

Submission8046...

Submission8046...

Submission8046...

Submission8046...

Submission8046...



Meta Review of Submission8046 by Area Chair uqfT

Meta Review by Area Chair uqfT 04 Mar 2026, 04:31 (modified: 17 Mar 2026, 09:12)

Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Revisions (/revisions?id=J1HSwMhPcb)

Metareview:

This paper tackles reasoning in complex visual scenes, where intricate entities and relations require precise visual grounding. The authors propose SceneAlign, a framework that leverages scene graphs to perform controllable structural interventions (following four predefined templates) on training samples; this way, they construct hard negative rationales that remain linguistically plausible but are grounded in inaccurate visual facts. The resulting preference pairs are filtered using an overlap constraint and a diversity sampling step, then used as preference data for Direct Preference Optimization (DPO) training. Across seven visual reasoning benchmarks and for multiple open MLLMs, SceneAlign consistently improves answer accuracy and reasoning faithfulness over the base models and an SFT baseline trained on positive reasoning samples only.

This paper is a resubmission. The previous version received an OA of 3, with suggested revisions mainly targeting clarity details, and the authors asked for the same set of reviewers and the same area chair for this resubmission. On top of the required clarifications, the authors performed a few extra experiments (more models and baselines, extra ablation study), as detailed in the revision notes; the improvement was acknowledged by one reviewer. The paper received 3 reviews in this round, with detailed responses from the authors, mostly clarifying misunderstandings or commenting on suggestions of extensions that are out of the paper's scope. Following this, one of the reviewers increased their score from 2.5 to 3, while the other two maintained their scores at 3.5 (and did not respond).

Summary Of Reasons To Publish:

- An innovative method that uses scene graphs to structurally align multimodal reasoning, offering a principled advance over text-based alignment.
- Consistent and significant improvements across seven challenging benchmarks and multiple model architectures (Qwen2.5/3-VL, InternVL3, LLaVA-Next), robustly validating the approach.
- The four interventions are mapped to concrete error types (role mis-binding, hallucination/substitution, omission, over-specification), making the supervision interpretable and extensible.

Summary Of Suggested Revisions:

- Given the misunderstanding by two reviewers on the use of the scene graph during training (instead of just generating the training samples), you should pay extra attention to clarify this point in the paper.
- Add the direct faithfulness evaluation done in the rebuttal.

Overall Assessment: 4 = Conference: I think this paper could be accepted to an *ACL conference.


Reported Issues:  No

Publication Ethics Policy Compliance: I did not use any generative AI tools for this review



Response to Meta Reviewer

Author-Editor Confidential Comment

by Authors ( Junda Wu (/profile?id=~Junda_Wu1), Julian McAuley (/profile?id=~Julian_McAuley1), Xintong Li (/profile?id=~Xintong_Li2), Jennifer Yuntong Zhang (/profile?id=~Jennifer_Yuntong_Zhang1), +4 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission8046/Authors))


 11 Mar 2026, 16:39 (modified: 11 Mar 2026, 16:55)

 Program Chairs, Senior Area Chairs, Area Chairs, Authors  Revisions (/revisions?id=WqUSTkvyER)

[Deleted]

Official Review of Submission8046 by Reviewer d9iW

Official Review by Reviewer d9iW  14 Feb 2026, 07:36 (modified: 17 Mar 2026, 09:12)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer d9iW, Commitment Readers

 Revisions (/revisions?id=fqcKKnLKDy)

Paper Summary:

SceneAlign targets reasoning unfaithfulness in multimodal LLMs on complex scenes by using scene graphs to create structure-aware, controllable counterfactuals rather than text-only perturbations. It identifies CoT-referenced subgraphs and applies four graph interventions to generate linguistically plausible but visually inconsistent negative

rationales, then uses DPO to prefer scene-faithful reasoning. Across multiple MLLMs and seven benchmarks, the method reports consistent improvements, suggesting that grounding-aware contrastive supervision is more effective than answer- or token-level negatives.

Summary Of Strengths:

- The paper precisely argues that token/answer perturbations let models exploit language priors and fail to localize grounding failures, motivating structure-aware supervision
- The four interventions are mapped to concrete error types (role mis-binding, hallucination/substitution, omission, over-specification), making the supervision interpretable and extensible.
- Using Jaccard-based filtering to avoid trivial/irrelevant negatives and an explicit diversity selection criterion is a sensible way to prevent weak preference pairs.
- The method is tested on multiple families (Qwen2.5/3-VL, InternVL3, LLaVA-Next) and shows consistent improvements beyond SFT, with especially large margins on hallucination/reasoning-heavy benchmarks.

Summary Of Weaknesses:

- Both scene-graph generation and positive CoT generation explicitly include “{ground-truth answer}”, which can contaminate the reasoning traces and weaken claims about learning grounding rather than learning to rationalize given the answer.
- The method assumes high-quality graph parsing from GPT-4o and even reports only a small human spot check; it remains unclear how performance degrades with weaker/cheaper parsers or out-of-distribution images.
- While testing spans many benchmarks, the preference training is constructed from a single dataset; more evidence is needed that the learned “structure-faithfulness” transfers when supervision is built from different distributions or tasks.

Comments Suggestions And Typos:

NA

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims. Some extra experiments could be nice, but not essential.

Excitement: 3.5

Overall Assessment: 3.5 = Borderline Conference

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Publication Ethics Policy Compliance: I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits



Response to Reviewer d9iW

Official Comment

by Authors (👤 Junda Wu (/profile?id=~Junda_Wu1), Julian McAuley (/profile?id=~Julian_McAuley1), Xintong Li (/profile?id=~Xintong_Li2), Jennifer Yuntong Zhang (/profile?id=~Jennifer_Yuntong_Zhang1), +4 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission8046/Authors))

📅 22 Feb 2026, 02:17 (modified: 17 Mar 2026, 09:12)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer d9iW, Commitment Readers

📄 Revisions (/revisions?id=vTfeROcT3c)

Comment:

We thank the reviewer for the thorough and thoughtful evaluation. We address the raised concerns below and clarify how our work handles them.

Response to Weakness 1

We appreciate the reviewer's concern and would like to clarify that incorporating the ground-truth answer during scene-graph and positive CoT generation does not contaminate the learning objective nor introduce privileged information at inference time.

First, this setup follows standard supervised learning. We do not claim an unsupervised setting, and all baselines are trained using the same question-answer supervision signals. Our pipeline does not rely on any additional labels beyond what is commonly available in VQA-style benchmarks, and is therefore directly comparable to mainstream reasoning and data-augmentation approaches.

For scene-graph construction, the ground-truth answer is used only during data synthesis to guide the extraction of question-relevant entities and relations, ensuring that the structural representation captures semantically critical components rather than producing generic image descriptions. The graph does not encode answer tokens, and it is never exposed to the model during DPO training or inference. Similarly, conditioning positive CoT generation on the ground-truth answer improves trajectory quality and prevents semantic drift during data construction. Importantly, the DPO objective operates on pairwise trajectory preferences rather than direct answer-token supervision. The model ultimately learns to prefer structurally consistent reasoning grounded in the (image, question) pair, and at inference time it receives only the original inputs. Therefore, the improvements cannot be attributed to post-hoc rationalization, but instead reflect improved structure-aware alignment under the same supervision regime as prior work.

Response to Weakness 2

We appreciate the reviewer's concern regarding the reliance on GPT-4o for scene-graph parsing and the robustness of the method. In our framework, the scene graph serves as an intermediate scaffold during data construction rather than as a required component during training or inference. We employ GPT-4o at this stage primarily to enhance data quality and structural consistency, rather than because the method fundamentally depends on a specific parser.

To assess the reliability of the generated graphs, we conducted a human sanity check (Appendix A.1). Among 150 randomly sampled A-OKVQA instances, all sampled graphs captured the entities and relations necessary to answer their associated questions. When minor imperfections appeared, they consisted of redundant nodes rather than missing question-critical elements. This suggests that the structural signals are sufficiently reliable for data synthesis, even if not exhaustively complete.

More importantly, SceneAlign does not require perfect graph extraction. The DPO objective operates on relative preference comparisons between structurally consistent and perturbed reasoning trajectories. Since learning is driven by structural contrast rather than absolute graph fidelity, the approach remains effective as long as the structural signal is not systematically misleading. We agree that weaker parsers may introduce additional noise and potentially reduce performance, but the method relies on sufficient structural consistency rather than parser perfection.

Response to Weakness 3

We thank the reviewer for raising the concern about cross-distribution generalization. While the preference training in SceneAlign is constructed from a single dataset, we explicitly evaluate the trained model across multiple benchmarks spanning different tasks and distributions. The consistent improvements observed beyond the training dataset provide direct empirical evidence that the learned structural alignment generalizes rather than overfitting to dataset-specific patterns.

Importantly, the supervision signal in SceneAlign operates at the level of structural consistency between reasoning trajectories and image-grounded entities and relations, rather than dataset-specific lexical patterns. This form of structure-level preference alignment is inherently more transferable than answer-token supervision tied to a particular dataset distribution.

Moreover, training on one dataset and evaluating on a diverse suite of benchmarks is a common paradigm in reasoning research (e.g., math or logical reasoning models trained on a single corpus and tested across multiple datasets). Our experimental design follows this paradigm, and the observed cross-benchmark gains substantiate the generalizability of the learned structure-faithfulness objective.

Official Review of Submission8046 by Reviewer br5m

Official Review by Reviewer br5m 📅 03 Feb 2026, 05:48 (modified: 17 Mar 2026, 09:12)

👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer br5m, Commitment Readers

📄 Revisions (/revisions?id=0wsFuUscwr)

Paper Summary:

This paper proposes SceneAlign, a new data pipeline for LVLm DPO training. By first generating scene graphs of input images, the authors automatically generated DPO training examples by perturbing the generated scene graphs followed by maliciously guided CoT reasoning.-

Summary Of Strengths:

- The paper is well-written with vivid demonstrations and clarification.
- The method is easy to understand.

Summary Of Weaknesses:

- About scene graph extraction:
 - An evaluation of the GPT-4o extracted scene graphs might be necessary.
 - As far as I can tell, the scene graphs used in this paper are structurally represented data. I wonder how it compares with unstructured representations like image captions, since you can also alter the image caption and generate CoT reasoning rollout based on captions instead of scene graphs.
- About CoT generation:
 - If I am right, when you generate the negative CoT traces, you do not provide the images as inputs to the LVLms. Any ideas and experimental results behind this design?
- About DPO training:
 - Have you tried to include the scene graph as part of the training supervision in the final training objective?

Comments Suggestions And Typos:

Overall, this is an interesting paper which has improved a lot from the previous version. Check my comments above the further revision.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 4 = Strong: This study provides sufficient support for all of its claims. Some extra experiments could be nice, but not essential.

Excitement: 3 = Interesting: I might mention some points of this paper to others and/or attend its presentation in a conference if there's time.

Overall Assessment: 3.5 = Borderline Conference

Ethical Concerns:

There are no concerns with this submission

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.

Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Publication Ethics Policy Compliance: I did not use any generative AI tools for this review



Response to Reviewer br5m

Official Comment

by Authors (Junda Wu (/profile?id=~Junda_Wu1), Julian McAuley (/profile?id=~Julian_McAuley1), Xintong Li (/profile?id=~Xintong_Li2), Jennifer Yuntong Zhang (/profile?id=~Jennifer_Yuntong_Zhang1), +4 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission8046/Authors))

22 Feb 2026, 02:24 (modified: 17 Mar 2026, 09:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer br5m, Commitment Readers

Revisions (/revisions?id=PZfLCnOA5J)

Comment:

We thank the reviewer for the thoughtful assessment and feedback. We appreciate the recognition of our contributions and address the concerns raised below.

Response to Weakness 1

1. Regarding the evaluation of GPT-4o-extracted scene graphs, this is addressed in Appendix A.1. We conducted a human spot check on 150 randomly sampled A-OKVQA instances to verify that the generated graphs capture the entities and relations required to answer the associated questions. The sampled graphs consistently covered question-critical elements; minor imperfections involved redundant nodes rather than missing key relations. Since our method only requires sufficient structural fidelity to support controlled perturbations (rather than exhaustive completeness), these results indicate that the graphs are reliable for data construction.
2. Regarding comparisons with unstructured representations such as image captions, SceneAlign focuses on multimodal reasoning tasks where fine-grained entity-relation grounding is central. Scene graphs provide explicit structural decomposition that enables localized, controlled perturbations (e.g., swapping relations or modifying specific entities), which is critical for constructing minimally inconsistent negative reasoning trajectories. In contrast, caption editing shifts the problem toward holistic language generation rather than structured reasoning alignment and thus falls outside the primary scope of this work. Therefore, the use of scene graphs is motivated by controllability tailored to multimodal reasoning objectives, rather than by representational preference.

Response to Weakness 2

We thank the reviewer for the question. The exclusion of the image during negative CoT generation is a deliberate design choice. When generating negative CoT traces, we deliberately exclude the image to ensure that the reasoning strictly follows the perturbed scene graph. The goal is to enforce adherence to the controlled structural modification, thereby producing hard negatives that reflect the injected structural inconsistency.

If the image were provided during negative generation, the LLM could access the original visual evidence and partially compensate for the perturbation. In such cases, the resulting reasoning would tend to drift back toward visually correct explanations, weakening the structural contrast between positive and negative

trajectories. Our objective is to construct negatives that are structurally plausible yet grounded in perturbed graph information, so that the primary controlled difference between positive and negative samples lies in their grounding consistency.

During preliminary ablations in development, we observed that including the image in negative generation frequently led to softened or partially corrected rationales, reducing the clarity of the preference signal. In contrast, removing the image produced reasoning traces that adhered more consistently to the perturbed structure, resulting in stronger trajectory-level contrast. This contrast is particularly important for DPO, which learns from relative preference comparisons rather than from absolute correctness labels. Therefore, this design is a deliberate mechanism to produce structurally controllable hard negatives and to maintain a clean and interpretable preference signal during DPO training.

Response to Weakness 3

We thank the reviewer for the insightful question regarding whether the scene graph could be incorporated as part of the training supervision in the final DPO objective.

We intentionally do not include the scene graph in the final training objective. In our framework, the scene graph serves only as an intermediate structural scaffold for constructing controlled positive and negative reasoning trajectories. It is not designed to be a supervision target or an additional model input.

Incorporating the scene graph directly into the training objective (e.g., as an additional supervision signal or conditioning input) would fundamentally alter the problem setting. It would convert the method into a graph-conditioned reasoning model, thereby introducing an additional structured input at inference time and changing the deployment contract. Our goal, however, is to improve a standalone multimodal LLM that operates solely on image-question pairs, without requiring external parsers or structured representations during inference.

Directly supervising on scene graphs could encourage the model to rely on graph-specific artifacts rather than improving intrinsic visual grounding. In contrast, our DPO objective operates at the trajectory level, where structural information influences learning indirectly through preference construction. For this reason, we restrict the scene graph to the data construction stage and do not incorporate it into the final training objective.

Official Review of Submission8046 by Reviewer SbHj

Official Review by Reviewer SbHj  02 Feb 2026, 01:56 (modified: 17 Mar 2026, 09:12)

 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer SbHj, Commitment Readers

 Revisions (/revisions?id=74t1Hyrdj4)

Paper Summary:

This paper targets a practical failure mode in multimodal large language models: they can produce fluent chains of thought that look plausible yet are not grounded in the image, especially in busy scenes with many entities and relations. The authors propose SceneAlign, a preference alignment pipeline that uses scene graphs to make grounding errors structured, localizable, and controllable rather than purely textual.

For each image and question, a GPT-4o based parser produces a scene graph of entities, attributes, and relations. A base MLLM then generates a positive chain of thought conditioned on the image, question, and graph. The method extracts the subgraph referenced by that reasoning and applies four targeted perturbations: swap for role confusion, replace for substituting key elements, shorten for removing key evidence, and overthink for injecting unsupported detail. These perturbed graphs drive the generation of negative chains of thought. The resulting preference pairs are filtered using an overlap constraint and a diversity sampling step, then used to train the model with DPO.

SceneAlign is applied to multiple open MLLMs and evaluated on seven benchmarks spanning grounding and reasoning. The reported results show consistent gains over the base models and an SFT baseline trained on positive reasoning, with additional comparisons against prior preference construction baselines on LLaVA-Next-8B.

Summary Of Strengths:

1. The supervision targets the right failure mode. The negative examples correspond to recognizable grounding errors at the level of entities and relations, not cosmetic token edits. This makes the training signal more interpretable and better aligned with how these systems fail in practice.
2. The perturbation design is clean and diagnostic. The four operators map to distinct error types and are easy to sanity check. This is valuable both for debugging and for communicating what the method is actually teaching the model.
3. The pipeline includes important “unsexy” details that matter for DPO. Overlap filtering and diversity sampling reduce trivial negatives and collapse, and the ablations suggest these components contribute meaningfully.
4. Empirical coverage is strong. The method is tested across multiple model families and a broad benchmark suite, and gains are not confined to a single dataset or a single model.

Summary Of Weaknesses:

1. Major methodological confound: the scene graph generator appears to use the ground truth answer. The appendix prompt includes the question and the ground truth answer when generating the scene graph. Even if the output is “just structure,” answer access can bias which entities and relations are included, effectively distilling label information into the graph. If the model is trained while conditioned on that graph, the reported gains may reflect privileged hints rather than improved grounding.
2. The inference time contract is unclear. The method is described as conditioning on a scene graph during training and during positive reasoning generation. It is not stated clearly whether evaluation also supplies a scene graph at test time, and if so how it is produced without labels. This ambiguity changes the contribution from improving a standalone MLLM to improving a pipeline that depends on an external parser, with different cost and reproducibility implications.
3. Faithfulness is a central claim, but the evidence is mostly indirect. Most results are accuracy style benchmark scores. Those are necessary, but they do not directly validate step level grounding. A small number of qualitative examples is not enough to support a claim about faithful reasoning.
4. The data reliability claim is overstated. The appendix reports a spot check where GPT-4o graphs capture all question relevant entities and relations in all cases. For open world images this is an unusually strong statement. It also becomes less informative if the answer is provided to the graph generator, since “relevance” is then answer guided.

Comments Suggestions And Typos:

1. State the test time setup explicitly for every main table. For each benchmark, specify whether the model receives a scene graph at inference. If it does, specify the generator, prompt, decoding settings, and approximate cost. If it does not, include a train with graph, test without graph ablation to address train test mismatch.
2. Add at least one direct faithfulness evaluation. Even a modest human study can be high value if it is focused: annotate whether each reasoning step is supported by visible evidence, or whether the final answer depends on unsupported intermediate claims. If human evaluation is infeasible, consider an evidence based protocol that checks consistency between reasoning steps and graph facts generated without labels.
3. Rule out shortcut learning from style cues. The overthink and shorten operators may correlate with length and verbosity. DPO can learn to prefer shorter or more cautious rationales without becoming more grounded. Please report statistics on reasoning length and provide a length controlled comparison, or enforce similar lengths for positive and negative chains of thought.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 3.5

Excitement: 3 = Interesting: I might mention some points of this paper to others and/or attend its presentation in a conference if there's time.

Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.

Datasets: 1 = No usable datasets submitted.

Software: 1 = No usable software released.

Knowledge Of Or Educated Guess At Author Identity: No

Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources

Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Publication Ethics Policy Compliance: I used a privacy-preserving tool exclusively for the use case(s) approved by PEC policy, such as language edits



Official Comment by Authors

Official Comment

by Authors (Junda Wu (/profile?id=~Junda_Wu1), Julian McAuley (/profile?id=~Julian_McAuley1), Xintong Li (/profile?id=~Xintong_Li2), Jennifer Yuntong Zhang (/profile?id=~Jennifer_Yuntong_Zhang1), +4 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission8046/Authors))

22 Feb 2026, 02:45 (modified: 22 Feb 2026, 02:47)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer SbHj

Revisions (/revisions?id=TbjZXw39D6)

[Deleted]



Response to Reviewer SbHj [1/3]

Official Comment

by Authors (Junda Wu (/profile?id=~Junda_Wu1), Julian McAuley (/profile?id=~Julian_McAuley1), Xintong Li (/profile?id=~Xintong_Li2), Jennifer Yuntong Zhang (/profile?id=~Jennifer_Yuntong_Zhang1), +4 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission8046/Authors))

22 Feb 2026, 02:41 (modified: 17 Mar 2026, 09:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer SbHj, Commitment Readers

Revisions (/revisions?id=NKZNg5rIK3)

Comment:

We thank the reviewer for the detailed and thoughtful comments. We are encouraged by the recognition of our motivation, empirical results, and the relevance of our work. Below, we provide responses to the key concerns and suggestions, and will revise the paper accordingly to improve clarity and rigor.

Response to Weakness 1

We would like to clarify that including the ground-truth answer during scene graph generation does not introduce a methodological confound. The scene graph is used only during data construction and is never provided during DPO training or inference. At test time, the model operates solely on the (image, question) pair. If answer-conditioned graphs introduced privileged shortcuts, one would expect degraded generalization once such signals are removed at inference. Instead, we observe consistent improvements over both the pretrained and SFT baselines.

The role of the ground-truth answer in graph generation is to ensure that the extracted structure captures question-relevant and semantically critical entities and relations, rather than producing a generic or exhaustive image description. This improves the reliability of structural perturbations used to construct positive and negative reasoning trajectories. Importantly, the DPO objective operates on pairwise preference comparisons between complete reasoning trajectories, not on answer tokens or graph tokens. The model never observes the scene graph during optimization; it only learns to prefer reasoning chains that are structurally consistent with the image over those that are inconsistent due to controlled perturbations.

Furthermore, negative trajectories are generated by explicitly modifying semantically critical components of the scene graph to create structurally inconsistent alternatives. This design ensures that the supervision signal arises from structural grounding differences rather than from answer exposure. Since neither the scene graph nor the ground-truth answer is available at inference time, the observed gains cannot be attributed to privileged label distillation. Instead, they reflect improved alignment toward semantically grounded reasoning grounded in the original image-question pair.

Response to Weakness 2 and Suggestion 1

Thank you for the thoughtful question and suggestion. We would like to make it explicit that no scene graph is used during inference. At test time, the model receives only the image and the question and produces a standard “chain-of-thought + answer” output. There is no external parser, no scene-graph generator, and no additional pipeline component involved at deployment. Therefore, our contribution reflects an improvement to a standalone MLLM, without introducing additional cost or reproducibility concerns at inference time.

We further clarify that scene graphs are never part of the model input during training. They are used solely as an intermediate structure during data construction to induce structured positive and negative reasoning traces. The model is trained on preference pairs over outputs formatted as “chain-of-thought + answer,” where the answer is generated at the end of the reasoning trace rather than directly copied from the ground truth. In this sense, the scene graph functions only as a scaffolding mechanism for constructing higher-quality preference data.

Importantly, the training and inference paradigms are aligned. In both cases, the model consumes the same multimodal inputs (image + question) and produces outputs in the same format (“chain-of-thought + answer”). Since scene graphs are never included in the model input during either training or inference, there is no train-test mismatch, and the concern regarding reliance on an external parser does not apply to our setting.



Response to Reviewer SbHj [2/3]

Official Comment

by Authors (Junda Wu (/profile?id=~Junda_Wu1), Julian McAuley (/profile?id=~Julian_McAuley1), Xintong Li (/profile?id=~Xintong_Li2), Jennifer Yuntong Zhang (/profile?id=~Jennifer_Yuntong_Zhang1), +4 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission8046/Authors))

22 Feb 2026, 02:45 (modified: 17 Mar 2026, 09:12)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer SbHj, Commitment Readers

Revisions (/revisions?id=nr0fHEO7Ub)

Comment:

Response to Weakness 3 and Suggestion 2

We thank the reviewer for the suggestion to include a direct faithfulness evaluation. Using a stronger external model as an automated evaluator is standard practice in recent faithfulness research [1-3]. Therefore, we conduct a step-level grounding assessment on Qwen2.5-VL-3B-Instruct model with A-OKVQA test set using GPT-4o as an external evaluator. In our setup, GPT-4o is used solely for evaluation under a fixed rubric, consistently across all models. For each reasoning trajectory, every numbered step is rated on a 0-5 scale based on whether it is supported by visible evidence in the image. We then compute per-sample averages and report the overall mean grounding score.

The results show a clear improvement in step-level grounding:

Pretrained model: 1.997; SFT baseline: 2.215; SceneAlign: 2.832.

Compared to the SFT baseline, our method improves the average grounding score by +0.617, and by +0.835 over the pretrained model. This evaluation directly measures whether intermediate reasoning steps are supported by image evidence, rather than relying solely on

final answer accuracy. The consistent margin indicates that SceneAlign not only improves outcome accuracy but also significantly enhances step-level visual grounding and reduces unsupported intermediate claims.

[1] Balasubramanian, Sriram, et al. "A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models." In EMNLP 2025.

[2] Moll, Johannes, et al. "Evaluating Reasoning Faithfulness in Medical Vision-Language Models using Multimodal Perturbations." In ML4H 2025.

[3] Lv, Weijian, et al. "SPD-Faith Bench: Diagnosing and Improving Faithfulness in Chain-of-Thought for Multimodal Large Language Models." In arXiv:2602.07833v1.

Response to Weakness 4

Generating structured scene representations with MLLMs for open-world imagery is a common intermediate step in multimodal reasoning pipelines [1–7]. In Appendix A.1, the statement that GPT-4o-generated graphs captured all question-relevant entities and relations refers specifically to the evaluated spot-check sample, rather than claiming universal structural completeness in arbitrary open-world settings.

Importantly, our method does not rely on scene graphs being perfectly exhaustive. As described in Sec. 3.1, the positive rationale is always a natural-language chain-of-thought grounded directly in the (image, question) pair, and the final MLLM conditions only on these inputs at inference time. The scene graph serves solely as an auxiliary scaffold during data construction to induce structured positive and perturbed rationales. Minor omissions or redundancies in the graph therefore do not propagate to deployment.

Regarding answer-guided relevance, providing the ground-truth answer to the graph generator is intended to focus structural abstraction on question-critical components, rather than to encode the answer itself. The graph is never used as a supervisory label or model input; it functions only as an intermediate structure for constructing controlled perturbations. Thus, the appendix analysis should be interpreted as verifying that the structural signal is sufficiently reliable for preference-pair construction, rather than as a claim of exhaustive open-world representation.

[1] Chen, Guoqing, et al. "RRHF-V: Ranking Responses to Mitigate Hallucinations in Multimodal Large Language Models with Human Feedback." in COLING 2025.

[2] Mitra, Chancharik, et al. "Compositional chain-of-thought prompting for large multimodal models." In CVPR 2024.

[3] Zheng, Changmeng, et al. "A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning." In ACMMM 2024.

[4] Long, Lin, et al. "On llms-driven synthetic data generation, curation, and evaluation: A survey." In ACL Findings 2024.

[5] Zhang, Wenqi, et al. "Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model." In EMNLP 2024.

[6] Li, Zhuochun, et al. "Learning from committee: Reasoning distillation from a mixture of teachers with peer-review." ACL Finding 2025.

[7] Li, Jiazheng, Hanqi Yan, and Yulan He. "Drift: Enhancing LLM Faithfulness in Rationale Generation via Dual-Reward Probabilistic Inference." In ACL 2025.



Response to Reviewer SbHj [3/3]

Official Comment

by Authors (👁️ Junda Wu (/profile?id=~Junda_Wu1), Julian McAuley (/profile?id=~Julian_McAuley1), Xintong Li (/profile?id=~Xintong_Li2), Jennifer Yuntong Zhang (/profile?id=~Jennifer_Yuntong_Zhang1), +4 more (/group/edit?id=aclweb.org/ACL/ARR/2026/January/Submission8046/Authors))

📅 22 Feb 2026, 02:49 (modified: 17 Mar 2026, 09:12)

👁️ Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer SbHj, Commitment Readers

📄 Revisions (/revisions?id=lvUa4MsSc1)

Comment:

Response to Suggestion 3

We thank the reviewer for this suggestion and report explicit reasoning length statistics to rule out shortcut learning from verbosity cues. We measured character-level CoT length for all perturbation methods (80 samples each).

Under Qwen2.5-VL-3B-Instruct, the mean lengths (± std) are: swap 981.8 (± 403.9), replace 996.3 (± 397.1), shorten 951.2 (± 393.4), and overthink 972.8 (± 347.8) characters;

Under Qwen2.5-VL-7B-Instruct, the corresponding values are: swap 1007.7 (± 367.9), replace 964.0 (± 363.1), shorten 948.8 (± 363.2), and overthink 975.3 (± 367.8) characters.

All perturbation types fall within a narrow range of approximately 950–1000 characters with highly overlapping standard deviations. There is no systematic inflation or reduction in length associated with any operator, and the distributions are comparable across settings. This is

expected because the perturbation operators are applied to the scene graph structure, not

directly to the chain of thought itself; the CoT is generated based on frequently asked questions

scene graph under the same prompting template. As a result, reasoning length is governed primarily by the shared prompt format rather than by the specific perturbation type.

All Venues (/venues)

Therefore, reasoning length is unlikely to serve as a shortcut signal for DPO; the improvements instead stem from learning to prefer structurally grounded reasoning rather than from

News (/groups?id=OpenReview.net/News&referrer=

Privacy Policy (/legal/privacy)

[Homepage](/)

[OpenReview \(/about\)](#) is a long-term project to advance science through improved peer review with legal nonprofit status. We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](#). © 2026 OpenReview