COUNTERGROUND3D: PROBING AND IMPROVING 3D LMM GROUNDING WITH A COUNTERFACTUAL INTERVENTION BENCHMARK

Anonymous authors
Paper under double-blind review

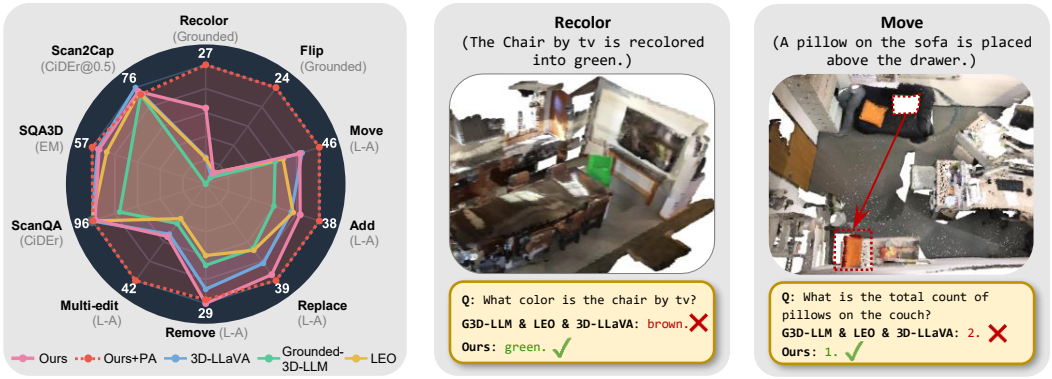


Figure 1: **CounterGround3D exposes language shortcut and dataset bias in 3D LMMs. Left**: Representative 3D LMMs perform decently on ScanQA, Scan2Cap, and SQA3D but fail frequently in various interventions from CounterGround3D. The proposed m^{2}DPO partially restores grounding capability in conventional setting, while additional pose input from the PoseAlign setup (Ours+PA) unleashes full performance (Liu et al., 2026). **Middle and Right**: 3D LMMs fail to notice obvious changes in the scene—the language prior of 'brown chairs' and dataset bias of 'two pillows on the sofa' induce incorrect answers without paying sufficient attention to the visual interventions. In contrast, m^{2}DPO successfully grounds the answer to the interventions.

ABSTRACT

Reliable 3D grounding requires large multimodal models (LMMs) to base their predictions on the visual evidence present in the point cloud. However, existing 3D LMMs tend to be dominated by language priors and dataset-level correlations, while frequently neglecting visual input. We introduce **CounterGround3D**, a paired counterfactual benchmark that directly evaluates whether model predictions track controlled interventions in the point cloud. CounterGround3D comprises six types of interventions, namely, *recolor*, *flip*, *add*, *remove*, *replace*, and *move*, covering a wide range of factors in grounding tasks. Each edited scene is paired with its original counterpart, enabling evaluation of both sensitivity to answer-critical interventions and invariance to irrelevant ones. Leveraging CounterGround3D, we propose **m^{2}DPO**, a mixed multimodal direct preference optimization framework that employs counterfactual pairs as direct preference supervision and boosts 3D grounding with little catastrophic forgetting. Experiments across multiple widely adopted 3D LMMs show that CounterGround3D reveals substantial grounding failures that remain hidden under conventional benchmarks, while m^{2}DPO significantly improves 3D grounding capabilities.

1 INTRODUCTION

Reliable 3D grounding requires a 3D large multimodal model (LMM) to attend to the visual evidence relevant to a query, which is essential for embodied agents operating in 3D environments (Mariani et al., 2025; Zeng et al., 2024). However, recent studies show that most existing 3D LMMs do not have clear performance change when point cloud inputs are removed or zeroed out, indicat-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ing their deep reliance on language shortcut and dataset-level correlations (Zhang et al., 2025; Ma et al., 2026). Specifically, 3D LMMs may confidently ignore changes in color, location, direction, or object relation, as illustrated in Figure 1. This limitation could affect a broad range of tasks severely, including question answering (QA) (Azuma et al., 2022; Yan et al., 2023), referring segmentation (Yan et al., 2023; Yuan et al., 2021), situated or dense captioning (Yuan et al., 2022; Chen et al., 2021; Achlioptas et al., 2020), and spatial reasoning (Ma et al., 2023a; Liu et al., 2026).

We argue that a strong diagnostic is *counterfactual grounding*. Given a scene-question-answer triplet $(\mathcal{P}, q, a)$, consider a controlled intervention τ that produces an edited scene $\mathcal{P}^\tau$. If an answer-critical intervention τ changes the answer $a^\tau \neq a$, a grounded model should follow accordingly. On the contrary, if τ makes an irrelevant intervention, the prediction should remain unchanged. Such formulation tests both *grounding* capability and *invariance* to irrelevant changes, providing a strong indicator for visual grounding capabilities.

While counterfactual grounding has been extensively explored in 2D vision-language models (Lee et al., 2025; Sar et al., 2026; Jin et al., 2026), relevant studies in 3D domain remain largely limited. Several preliminary 3D benchmarks attempt to address this challenge recently. For example, Yang et al. (2025a) propose 3D-GRAND and 3D-POPE benchmarks to investigate object-language binding, but only focus on existence grounding. Hypo3D studies hypothetical reasoning on an obsolete point cloud and a language description of scene modification, its counterfactual reasoning is therefore focused on language space rather than 3D input grounding (Mao et al., 2025). DEER-3D introduces error-driven counterfactual scene editing for training and improves grounding in referring tasks (Zhang et al., 2026), but does not provide a broad paired benchmark spanning multiple intervention types and task families. Additionally, existing 3D datasets evaluate examples independently and rarely provide matched scene pairs with answer-critical interventions. This leaves a central question unresolved: can 3D grounding be evaluated through controlled scene-level interventions that isolate visual causality, and can 3D LMMs be trained with such pairs to effectively enhance their grounding capability?

To bridge this gap, we introduce **CounterGround3D**, a benchmark of physically realized counterfactual pairs for diagnosing, evaluating, and improving grounding in 3D LMMs. CounterGround3D spans complementary grounding factors through six types of controlled intervention: *recolor*, *flip*, *add*, *remove*, *replace*, and *move*. These interventions modify object color, directional relations, object existence and identity, spatial configuration, and instance-level binding. Specifically, we construct clean recoloring and global flip tracks from existing 3D QA and captioning annotations, and materialize additional object-level interventions from Hypo3D descriptions through structured edit extraction, instance grounding, physically reliable point cloud editing, and manual labeling. The resulting benchmark contains paired original and edited examples, and further includes a multi-intervention subset that combines multiple interventions into harder counterfactual cases. Unlike existing 3D benchmarks, CounterGround3D enables direct measurement of grounding capability and invariance to irrelevant interventions, providing a causal evaluation of grounding reliability.

Leveraging CounterGround3D, we propose **m^{2}DPO**, a mixed multimodal direct preference optimization framework built upon mDPO (Wang et al., 2024). For vanilla mDPO on counterfactual samples, the counterfactual answer forms the visually consistent response, while the original answer becomes a hard negative, yet is unbalanced and causes catastrophic forgetting on original data. In contrast, our m^{2}DPO leverages the pair structure of CounterGround3D by symmetric loss designs and guards the performance with supervised counterfactual training, encouraging models to track visual interventions while preserving their original capabilities.

Extensive experiments show that most existing 3D LMMs exhibit substantial failures under controlled counterfactual interventions despite competitive performance on their used benchmarks. CounterGround3D exposes these failures through paired evaluation and separates grounding capability from exploitation of language shortcuts. The proposed m^{2}DPO substantially improves counterfactual grounding, and further integration with the PoseAlign setup (Liu et al., 2026) elevates overall performance by a large margin. Hence, CounterGround3D is not only useful for diagnosis but also competitive on improving grounding capability.

Our contributions are threefold: (1) We introduce **CounterGround3D**, a paired 3D counterfactual grounding benchmark spanning VQA and captioning tasks, with physically realized interventions over color, direction, object existence, identity, and spatial configuration. (2) We propose **m^{2}DPO**, a

mixed multimodal preference-training framework that leverages counterfactual pairs while incorporating supervised training to improve grounding capabilities. (3) Extensive experiments demonstrate that CounterGround3D reveals language shortcut behavior that is obscured by conventional benchmark accuracy, and the proposed **m^{2}DPO** improves grounding performance significantly by training over CounterGround3D..

2 Related Work

Visual grounding and hallucination in 2D VLMs. A large body of work studies whether multimodal models rely on visual evidence or exploit language priors and dataset correlations. Early hallucination benchmarks such as CHAIR and POPE focus primarily on object existence, measuring whether models mention or affirm objects absent from the image (Rohrbach et al., 2018; Li et al., 2023). Later benchmarks extend evaluation to attributes and relations, e.g., AMBER and hierarchical variants of POPE (Wang et al., 2023; Pham & Schott, 2024). In parallel, compositionality benchmarks such as Winoground, ARO, CREPE, SugarCrepe, and ARPGrounding construct hard negatives by swapping attributes, relations, or subject object roles while largely preserving lexical content (Thrush et al., 2022; Yuksekgonul et al., 2022; Ma et al., 2023b; Hsieh et al., 2023; Zeng et al., 2024). These datasets reveal that correct category recognition or image-text matching does not necessarily imply correct attribute binding or relational grounding.

Counterfactual evaluation and language-prior diagnosis. More recent work evaluates grounding through controlled interventions. VLind-Bench separates failures caused by language priors from failures in basic perception by introducing prerequisite visual and commonsense checks before counterfactual evaluation (Lee et al., 2025). HallusionBench uses logically related control questions to diagnose visual hallucination and inconsistency (Guan et al., 2024). VLM Reality Check constructs minimally edited images through interventions such as object removal, color changes, mirroring, and background replacement, allowing robustness to specific visual perturbations to be measured directly (Sar et al., 2026). Chakraborty et al. (2025) also discover that strong LLM-driven agents fail under scene-task inconsistencies due to language priors. Related work also constructs counterfactual negatives by modifying entities, relations, or temporal factors while keeping other content fixed (Goyal et al., 2017; Jin et al., 2026; Ma et al., 2025). These benchmarks motivate a stronger notion of grounding: predictions should change under interventions on task-relevant evidence while remaining invariant to irrelevant changes. This principle is particularly important for distinguishing visual dependence from shortcut-based prediction.

Methods for increasing visual dependence. This line of work increase visual dependence by comparing disturbances during training or inference. In the inference stage, VCD suppresses tokens that remain likely even when visual evidence is degraded (Leng et al., 2024). M3ID similarly emphasizes the growth of language-only prior with sequence length (Favero et al., 2024). 3D-VCD applies contrastive decoding on a language-only LLM, parsing the disturbed point cloud as edited scene graph. These methods require basic model competence and therefore rarely work on the smaller 3D-native models with heavy language shortcuts. Training-time approaches include hallucination-aware data cleaning and rebalancing (Yu et al., 2024), hallucination-augmented contrastive learning (Jiang et al., 2024), and multimodal preference optimization (Wang et al., 2024; Huang et al., 2025). Recent DPO variants construct on-policy, localized, or multi-negative preference pairs to expose models to diverse visually inconsistent alternatives (Yang et al., 2025b; Yoon et al., 2025; Li et al., 2026). A related line introduces causal signals into policy optimization; COPO estimates the causal contribution of intermediate reasoning tokens and incorporates causal completeness into GRPO-style training (Guo et al., 2026). These papers suggest a path of combining structured counterfactual negatives with targeted training objectives that explicitly reward visual sensitivity.

Grounding in 3D large multimodal models. Most existing 3D grounding benchmarks evaluate reference resolution or object localization. ScanRefer, ReferIt3D, and Multi3DRefer measure whether a referring expression is mapped to the correct 3D instance or region (Chen et al., 2020; Achlioptas et al., 2020; Zhang et al., 2023). However, these datasets lack a causal decomposition of grounding capability versus reliance of language shortcuts. To bridge this gap, Mariani et al. (2025) propose an isolated object-level attribute grounding benchmark. 3D-POPE and 3D-GRAND evaluate object existence grounding and visual-language alignment (Yang et al., 2025a). Hypo3D

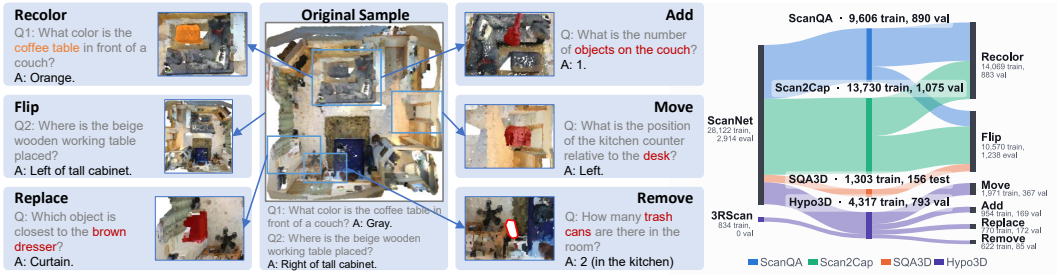


Figure 2: **The proposed counterfactual intervention dataset CounterGround3D and data sources.** The dataset consists of six types of interventions: recolor, flip, replace, add, move, and remove, covering a wide range of attribute, location, direction, or relation hallucinations. The benchmark composes of 31.9k counterfactual intervention data.

requires the 3D LMM to implicitly update its visual input with a counterfactual language description before VQA (Mao et al., 2025). Direction-aware 3D LMMs show that directional reasoning requires an explicitly defined and aligned reference frame (Liu et al., 2026). DEER-3D diagnoses grounding errors and generates targeted edits such as recoloring, cloning, and rotation (Zhang et al., 2026), but it does not provide a benchmark and has limited edit variety and task diversity.

3 Dataset Generation

CounterGround3D evaluates whether a 3D LMM responds to the *visual state of the scene*, rather than exploiting correlations between questions and answers. As displayed in Figure 2, we construct paired examples in which a controlled intervention changes a specific scene property while preserving unrelated content. We first formalize the counterfactual construction, then introduce two shortcut-free interventions, *recolor* and *flip*, and finally describe a specific pipeline that converts hypothetical changes from Hypo3D into physically realized point cloud edits, i.e., to *move, replace, add, and remove* the object instances. Scenes are drawn from ScanNet and 3RScan, with object instances represented by segmented point sets and their associated semantic labels (Dai et al., 2017; Wald et al., 2019).

3.1 Counterfactual Pair Definition

Let a source example be $e = (\mathcal{P}, q, a)$, where $\mathcal{P} = \{(p_i, c_i, s_i)\}_{i=1}^{N}$ denotes a colored point cloud with 3D coordinates $p_i \in \mathbb{R}^3$, color $c_i \in \mathbb{R}^3$, and instance segmentation mask s_i; q is the language query and a its ground-truth answer. For an intervention τ, we construct $e^\tau = (\mathcal{P}^\tau, q^\tau, a^\tau)$ where P^τ physically realizes the intervention in 3D, q^τ and a^τ are the modified the language sections according to the intervention. We therefore obtain a counterfactual pair: $\{e, e_\tau\} = \{(\mathcal{P}, q, a), (\mathcal{P}^\tau, q^\tau, a^\tau)\}$. Crucially, whether the expected response changes is determined by the *post-intervention scene*. For an intervention τ that is relevant to the queried predicate, the target fact changes and hence

$$f^\star(\mathcal{P}^\tau, q^\tau) = a^\tau \neq a = f^\star(\mathcal{P}, q), \tag{1}$$

where $f^\star$ denotes the oracle grounded answer function. Similarly, an irrelevant intervention τ' that operates on a query-irrelevant factor should satisfy

$$f^\star(\mathcal{P}^{\tau'}, q^{\tau'}) = a^{\tau'} = a = f^\star(\mathcal{P}, q). \tag{2}$$

These two cases respectively test *counterfactual grounding* and *invariance to irrelevant intervention*. Their generation is unified into the same procedure but given either a task-relevant intervention or an irrelevant one.

3.2 Counterfactual Recoloring

We first construct a bias-free recolor intervention as displayed in Figure 3. We synthesize color change samples from ScanQA and Scan2Cap, where the queried or described target instance is explicitly annotated. We identify examples involving object color in the answers as the basis for

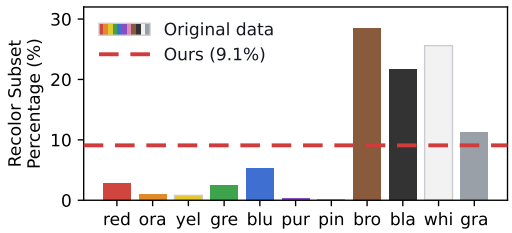


Figure 3: Distribution of color in answers before and after recolor intervention.

Table 1: Number of original scenes about which all GT question-answer pairs ask only one direction (either left or right), denoted as #Biased. Such correlation creates exploitable prior, which is eliminated by our dataset.

SOURCE	POINT CLOUDS	#BIASED	%
ScanQA	588	144	24.5
SQA3D	570	78	13.7
Scan2Cap	703	14	2.0
Ours	1861	**0**	**0.0**

recolor examples. For an annotated target instance o with point-index set $M_o = \{i | s_i = o\}$, recoloring modifies only its RGB channels:

$$c_i^\tau = \begin{cases} \text{color_adapt}(\tilde{c}, c_i), & i \in M_o, \\ c_i, & i \notin M_o, \end{cases} \qquad p_i^\tau = p_i, \tag{3}$$

where $\tilde{c}$ is uniformly sampled from a fixed basic color vocabulary, and $\text{color_adapt}(\cdot, \cdot)$ first alters the hue-saturation-value (HSV), keeping S and V unchanged while aligning H to that of $\tilde{c}$, then re-scales instance average RGB to that of $\tilde{c}$. This operation preserves surface texture and relative brightness. For achromatic colors like white, black, and gray, we reduce S to 0 and change V correspondingly. The answer a^τ is obtained by replacing the target color text with $\tilde{c}$, while $q^\tau = q$ remain unchanged to eliminate the possibility of language shortcuts. We explicitly ground the color field in the answer by prompting GPT-5.4 (OpenAI, 2025a), and discard ambiguous cases such as colors referring only to an irrelevant object or multicolored objects. Uniform sampling generates uniform distribution and prevents the formation of any exploitable object-class prior.

3.3 Counterfactual Left/Right Flip

The second bias-free intervention targets directional grounding. As established by Lemma 1, reflection across any vertical plane always reverses the left/right orientation, regardless of the ego pose and reflection axis. In practice, we reflect the scene across the yz-plane for convenience:

$$\mathcal{P}^\tau = \{(\text{diag}(-1, 1, 1)p_i, c_i, s_i)\}_{i=1}^N . \tag{4}$$

We retain only examples containing mirror-asymmetric expressions, including *left*, *right*, *leftmost*, *rightmost*, *lefthand*, *righthand*, *etc*, while discarding symmetric relations such as *near*, *above*, or *next to*. These expressions are grounded by prompting GPT-5.4 (OpenAI, 2025a). In theory, all mirror-asymmetric expressions should be swapped to maintain coordinate correctness:

$$q^\tau = \text{exchange_LR}(q), a^\tau = \text{exchange_LR}(a). \tag{5}$$

However, such change of question content still presents possibilities for language shortcuts. Therefore, the examples are further filtered to the subset with no directional words in questions where $q^\tau = q$. Besides, while ScanQA, Scan2Cap, and SQA3D contain similar frequencies of left/right in answers, this hides a strong data correlation between scenes and directional answers as displayed in Table 1. In contrast, our counterfactual pair contains two perfectly mirrored examples $(\mathcal{P}, q, a)$ and $(\mathcal{P}^\tau, q, a^\tau)$ for each scene $\mathcal{P}$, which naturally removes the data correlations between left/right occurrence with point cloud scenes, and removes language shortcuts through identical questions.

3.4 Counterfactual Object Interventions

We next construct more general geometric counterfactuals based on Hypo3D (Mao et al., 2025). Its source annotations describe hypothetical scene modifications in natural language, but many correspond to physical states that cannot be deterministically represented in a static point cloud. We therefore restrict the executable operation set of object interventions to *remove*, *move*, *add*, and *replace*. Free-form interventions such as opening, breaking, spilling, or changing illumination are discarded. Since Hypo3D did not provide pre-edit answers, we only provide intervened examples $(\mathcal{P}^\tau, q^\tau, a^\tau)$ as a secondary variation of the CounterGround3D subsets.

Object intervention realization. We probe the Hypo3D dataset and discover a unified pattern: After discarding the free-form operations, other change descriptions can be decomposed into a quadruple: the operation, a target object, an optional reference anchor, and a relation to the anchor or the room. We prompt GPT-5.4 to perform such decomposition and filter them with a feasibility check (OpenAI, 2025a). We then ground the quadruple to actual instance ids and use GPT-5.4 to resolve ambiguities. Such structured intervention is then materialized with a symbolic engine to implement different types of interventions. To obtain the newly added object point clouds in *add* and *replace*, we craft a donor bank from the training set indexed by class names, and perform a size-based rejection sampling to select a proper object. See Appendix A.3 for details.

Exhaustive manual labelling. Given 5.5k roughly realized object interventions from the realization pipeline, we conduct a round of exhaustive human screening and correction by experts to improve annotation quality. The screening process includes rectifying object placements, auditing the correctness of questions and answers, and deletion of erroneous pairs introduced in Hypo3D. For example, there are $\leq 1\%$ of the questions in Hypo3D where the scene implied by the question-answer pair does not match the counterfactual description. We discard 7.0% of the questions that contain irrecoverable placement artifacts or unrealistic changes. Approximately 56.0% and 13.8% of the questions and answers are corrected, respectively. In all, we acquire 5,110 clean examples for training and validation, fully validated by human experts. See Appendix A.4 for details.

Multi-intervention benchmark. After obtaining all six types of counterfactual invervention datasets, we search for the examples that fit standards of multiple different subsets to craft a multi-intervention benchmark. Since no question contains all three of color, direction, and object queries, the multi-intervention subset is composed of recolor+flip, recolor+object, and flip+object. These interventions are implemented and evaluated in the same procedure of single interventions and act as natural extensions of the CounterGround3D benchmark.

4 METHOD

In this section, we investigate how training on CounterGround3D improves the grounding capabilities of 3D LMMs. A straightforward baseline augments the original supervised fine-tuning (SFT) data with CounterGround3D, where samples are mixed with ScanQA, SQA3D, and Scan2Cap data. To explicitly optimize counterfactual grounding while preserving existing capabilities, we propose m^{2}DPO, which combines mDPO with symmetric loss designs and supervised learning.

Background on mDPO. In 2D VLMs, mDPO combines response preference with visual preference to reduce hallucination (Wang et al., 2024). Exploiting our scene notation, CoPO compares an intervened visual input $\mathcal{P}^\tau$ with an irrelevantly intervened input $\mathcal{P}^{\tau'}$ while holding (q^τ, a^τ) fixed; DPO prefers matched answer a^τ over an irrelevantly intervened $a^{\tau'}$ under the same input $(\mathcal{P}^\tau, q^\tau)$ (Rafailov et al., 2023). The respective loss functions are defined in Equations 6-9:

$$r(\mathcal{P}^\tau, q^\tau, a^\tau) = \beta\big(\log \pi_\theta(a^\tau \mid \mathcal{P}^\tau, q^\tau) - \log \pi_{\text{ref}}(a^\tau \mid \mathcal{P}^\tau, q^\tau)\big) \tag{6}$$

$$\mathcal{L}^{mdpo}_{\text{CoPO}} = -\log \sigma\big(r(\mathcal{P}^\tau, q^\tau, a^\tau) - r(\mathcal{P}^{\tau'}, q^\tau, a^\tau)\big) \tag{7}$$

$$\mathcal{L}^{mdpo}_{\text{DPO}} = -\log \sigma\big(r(\mathcal{P}^\tau, q^\tau, a^\tau) - r(\mathcal{P}^\tau, q^\tau, a^{\tau'})\big) \tag{8}$$

$$\mathcal{L}^{mdpo}_{\text{AncPO}} = -\log \sigma\big(r(\mathcal{P}^\tau, q^\tau, a^\tau) - \delta\big) \tag{9}$$

Here, $\beta > 0$ is the reward scaling strength, δ is the anchor threshold, and $\sigma(\cdot)$ is the sigmoid function. The distributions $\pi_\theta(\cdot \mid \cdot, \cdot)$ and $\pi_{\text{ref}}(\cdot \mid \cdot, \cdot)$ are the conditional answer distributions of the current model and a frozen reference model, the latter typically being a frozen copy before mDPO.

Definition of m^{2}DPO. In CounterGround3D, both $(\mathcal{P}, q, a)$ and $(\mathcal{P}^\tau, q^\tau, a^\tau)$ are valid examples. Therefore, we introduce four symmetric terms in Equations 10-11 to leverage paired structure:

$$\mathcal{L}_{\text{CoPO}} = -\log \sigma\big(r(\mathcal{P}, q, a) - r(\mathcal{P}^\tau, q, a)\big); \mathcal{L}^\tau_{\text{CoPO}} = -\log \sigma\big(r(\mathcal{P}^\tau, q^\tau, a^\tau) - r(\mathcal{P}, q^\tau, a^\tau)\big) \tag{10}$$

$$\mathcal{L}_{\text{DPO}} = -\log \sigma\big(r(\mathcal{P}, q, a) - r(\mathcal{P}, q, a^\tau)\big); \mathcal{L}^\tau_{\text{DPO}} = -\log \sigma\big(r(\mathcal{P}^\tau, q^\tau, a^\tau) - r(\mathcal{P}^\tau, q^\tau, a)\big) \tag{11}$$

We also incorporate supervised losses on the counterfactual pair as a stronger anchor, as defined in Equation 12. L and L^τ denote ground truth distributions, and $\text{CE}(\cdot, \cdot)$ is the cross entropy loss.

$$\mathcal{L}_{\text{SFT}} = \text{CE}\big(L, \pi_\theta(a \mid \mathcal{P}, q)\big), \quad \mathcal{L}^\tau_{\text{SFT}} = \text{CE}\big(L^\tau, \pi_\theta(a^\tau \mid \mathcal{P}^\tau, q^\tau)\big) \tag{12}$$

5 RESULTS

Datasets. Experiments are conducted on ScanNet, while 3RScan only appears during training to maintain evaluation consistency (Dai et al., 2017; Wald et al., 2019). For ScanQA, Scan2Cap, and SQA3D, we follow convention protocol for training-validation-test split Azuma et al. (2022); Chen et al. (2021); Ma et al. (2023a); Deng et al. (2025). We also report performance on the PoseAlign benchmark which aligns the point cloud $\mathcal{P}$ to an egocentric pose to express left/right direction consistently in the egocentric frame (Liu et al., 2026). PoseAlign is a slightly stronger setup and is ranked separately in all tables, marked with the **PA** suffix. Detailed description is in Appendix A.4.

Metrics. We include regular metrics including CiDEr (**C**), BLEU-4 (**B**), ROGUE-L (**R**), METEOR (**M**), EM, and EM-R to measure performance on ScanQA, Scan2Cap, and SQA3D. Furthermore, to precisely measure grounding capability and invariance to irrelevant interventions, we develop **G**, **I3**, and **IC** based on LLM-as-Judge accuracy (**L-A**) of Liu et al. (2026). LLM-as-Judge accuracy on original and edited data are noted as **L-A**$_{\mathrm{o}}$ and **L-A**$_{\mathrm{e}}$. When evaluated on answer-critical intervention examples, we define *grounded* (**G**) as the percentage of counterfactual pairs that are both judged as correct. When evaluated on answer-irrelevant examples, we define *irrelevant intervention invariance* (**I3**) as the percentage of answers that stay exactly the same, and further define *invariant correctness* (**IC**) as both passing I3 and L-A$_{\mathrm{o/e}}$ tests. See Appendix A.4 for details.

Baselines and models. We benchmark 5 baseline models with different architectures and point cloud tower implementations, including 3D-LLaVA, LL3DA, LEO, Grounded-3D-LLM, and DEER-3D (Deng et al., 2025; Chen et al., 2024a; Huang et al., 2023; Chen et al., 2024b; Zhang et al., 2026). Since DEER-3D is highly relevant but its edit code is not open-sourced, we re-implement it on 3D-LLaVA for evaluation protocol consistency and report performance of our trained model, marked as DEER-3D†. A blind baseline on 3D-LLaVA is also reported where the point cloud tokens are removed from input sequence. We also report two our models trained on CounterGround3D from 3D-LLaVA: **ours-composite** which adds our data into SFT stage directly, and **ours-m^{2}DPO** which leverages m^{2}DPO for better grounding. Please refer to Appendix A.4 for training specifications, and see Appendix A.6 for analysis of DEER-3D†.

5.1 MAIN RESULTS

Grounding capability. We evaluate counterfactual grounding performance on CounterGround3D, ScanQA, SQA3D, and Scan2Cap in Table 2, which reveals a substantial gap between conventional task performance and counterfactual grounding. Consistent with Zhang et al. (2025) and Ma et al. (2026), the blind baseline retains non-trivial conventional benchmark performance but exhibits floor-level grounding on recolor and flip, showing that CounterGround3D effectively separates visually insensitive models from genuinely grounded models. Among baselines, DEER-3D achieves the strongest recolor grounding (7.8%), while LEO performs best on flip grounding (2.9%); both remain low despite strong conventional benchmark performance, indicating that conventional benchmark scores do not necessarily reflect reliable usage of edited 3D evidence. Counterfactual training substantially improves multiple dimensions: Ours-m^{2}DPO achieves the best recolor (20.0%) and flip (3.1%) grounding, while Ours-composite performs best on replacement (36.6%) and remains competitive on add, remove, and multi-intervention. While our models achieve slightly worse original-task performance on ScanQA, SQA3D, and Scan2Cap than 3D-LLaVA, we interpret this phenomenon as a necessary cost because these benchmarks signal more shortcuts and biases than grounding capability, as achieved by the blind 3D-LLaVA model. Overall, CounterGround3D boosts sensitivity to answer-critical interventions, while also exposing grounding failures hidden by conventional benchmark metrics. Qualitative examples are shown in Figure 4.

Invariance to irrelevant interventions. We further evaluate the invariance to irrelevant interventions of all models in Table 3 as a complementary experiment to grounding evaluation. Recolor L-A$_{\mathrm{o}}$ differs slightly because invariance uses a larger evaluation set, as detailed in Appendix A.4. Since a visually insensitive model can remain invariant without using edited 3D evidence, we interpret I3 jointly with IC and G in Table 2. Flip is excluded because a global reflection is answer-critical by design. Among baselines, DEER-3D achieves the strongest recolor IC (40.6%) and the highest IC on move, add, and remove (38.2%, 31.4%, and 29.4%), while 3D-LLaVA has the highest raw

Table 2: Counterfactual grounding performance across different intervention types and original-task performance. Major metrics are highlighted with gray background. Blind 3D-LLaVA is incompatible with Scan2Cap, so related values are marked as brown and not comparable. LEO uses stronger setting on ScanQA and is marked as gray. The PoseAlign benchmark provides additional ego pose input and is ranked independently.

Model	Recolor			Flip			Move	Add	Replace	Remove	Multi-op	ScanQA	SQA3D	Scan2Cap
	G	L-A$_e$	L-A$_o$	G	L-A$_e$	L-A$_o$	L-A$_e$	L-A$_e$	L-A$_e$	L-A$_e$	L-A$_e$	C	EM	C@0.5
Blind 3D-LLaVA	0.0	5.2	48.8	0.0	38.3	39.2	35.4	29.0	26.2	16.5	18.2	82.3	51.8	–
Baseline Models:														
LL3DA	0.1	3.5	37.1	0.9	11.6	9.3	26.2	18.9	25.0	18.8	16.0	75.3	–	61.6
Grounded-3D-LLM	0.0	3.6	44.7	1.5	17.0	16.6	28.1	23.1	26.2	20.0	17.2	72.7	–	70.6
LEO	5.9	15.1	45.1	2.9	15.8	17.8	31.3	29.6	26.7	17.6	15.1	101.4	50.0	72.4
3D-LLaVA	6.1	12.6	56.1	1.9	16.8	19.1	39.0	29.0	32.0	25.9	21.9	**96.0**	**54.6**	76.3
DEER-3D†	7.8	14.4	**57.4**	1.7	17.1	18.0	**40.1**	**32.5**	30.2	**30.6**	**25.0**	94.6	54.5	**76.3**
Our Models:														
Ours-composite	17.4	32.5	51.3	2.7	20.2	21.3	38.1	32.0	**36.6**	29.4	22.9	93.5	53.9	73.2
Ours-m^{2}DPO	**20.0**	**38.3**	53.2	**3.1**	**24.2**	**24.1**	33.8	30.8	34.9	25.9	19.8	87.0	53.8	70.4
PoseAlign Benchmark:														
LL3DA+PA	0.0	3.3	39.4	0.5	14.1	13.6	31.6	27.2	27.9	22.4	18.8	76.7	–	63.2
3D-LLaVA+PA	7.2	12.8	**55.9**	9.9	23.3	22.4	40.6	33.1	**39.0**	24.7	21.4	**100.6**	54.4	**73.7**
Ours-composite+PA	23.6	40.5	52.9	**24.3**	**33.3**	31.9	44.1	37.3	38.4	**31.8**	40.1	98.6	57.3	72.5
Ours-m^{2}DPO+PA	**27.3**	**43.8**	55.0	24.1	33.0	**32.3**	**46.0**	**38.5**	**39.0**	28.5	**42.2**	95.3	**57.4**	70.7

Table 3: Invariance evaluation under irrelevant counterfactual interventions. Major metrics are highlighted with gray background. Blind 3D-LLaVA is incompatible with Scan2Cap, so related values are marked as brown and not comparable. The PoseAlign benchmark provides additional ego pose input and is ranked independently. High I3 and G together signal counterfactual grounding capability; high I3 with low G instead suggests visual insensitivity.

Model	Recolor			Move			Add			Replace			Remove		
	I3	L-A$_o$	IC	I3	L-A$_e$	IC	I3	L-A$_e$	IC	I3	L-A$_e$	IC	I3	L-A$_e$	IC
Blind 3D-LLaVA	100.0	36.0	36.0	100.0	35.4	35.4	100.0	29.0	29.0	100.0	26.2	26.2	100.0	16.5	16.5
Baseline Models:															
LL3DA	61.0	14.3	13.5	88.0	26.2	23.7	88.2	18.9	18.3	83.1	25.0	23.3	88.2	18.8	16.5
Grounded-3D-LLM	69.4	33.4	24.0	91.6	28.1	26.4	91.7	23.1	23.1	90.1	26.2	24.4	90.6	20.0	18.8
LEO	62.2	36.3	24.3	85.3	31.3	28.9	87.6	29.6	27.2	83.1	26.7	25.6	84.7	17.6	14.1
3D-LLaVA	90.9	43.5	40.2	**93.7**	39.0	37.3	**94.1**	29.0	28.4	**93.6**	32.0	31.4	**97.6**	25.9	25.9
DEER-3D†	**92.6**	42.8	**40.6**	91.0	**40.1**	**38.2**	92.3	**32.5**	**31.4**	86.6	30.2	29.7	94.1	**30.6**	**29.4**
Our Models:															
Ours-composite	83.7	40.8	35.0	92.9	38.1	36.2	92.9	32.0	30.2	92.4	**36.6**	**34.3**	95.3	29.4	28.2
Ours-m^{2}DPO	81.0	**43.7**	36.5	92.6	33.8	32.7	87.6	30.8	27.2	91.9	34.9	33.1	95.3	25.9	24.7
PoseAlign Benchmark:															
LL3DA+PA	**93.7**	20.7	20.0	95.6	31.6	31.3	**94.7**	27.2	26.0	89.5	27.9	26.7	89.4	22.4	21.2
3D-LLaVA+PA	92.0	42.5	39.7	**95.9**	40.6	39.8	92.9	33.1	33.1	89.5	**39.0**	36.6	90.6	24.7	24.7
Ours-composite+PA	91.2	40.8	38.6	93.7	44.1	43.1	92.3	37.3	35.5	89.0	38.4	35.5	**95.3**	**31.8**	**31.8**
Ours-m^{2}DPO+PA	92.2	**42.9**	**40.7**	95.1	**46.0**	**44.7**	**94.7**	**38.5**	**37.3**	**90.1**	**39.0**	**37.2**	**95.3**	28.5	27.3

I3 on all four geometry-changing interventions. However, their weaker grounding scores show that high invariance can coexist with visual insensitivity. Our variants maintain comparable invariance and correctness with baselines: Ours-composite achieves the best replace L-A$_e$ and IC (36.6% and 34.3%) and remains competitive on move, add, and remove. The invariance of our models are further improved when combined with the PoseAlign setup. Overall, counterfactual training preserves stability under irrelevant interventions while improving responsiveness to relevant ones.

PoseAlign benchmark analysis. The PoseAlign setup uses the same set of training and evaluation data as regular experiments, but resolves directional ambiguity by rigidly transforming each scene to an egocentric reference frame (Liu et al., 2026), and Tables 2–3 show that this substantially strengthens both counterfactual grounding and invariance performances. We first highlight that PoseAlign alone does not provide counterfactual grounding, with the best PoseALign baseline 3D-LLaVA+PA only achieving 7.2% and 9.9% recolor and flip grounding, respectively. In contrast, Ours-m^{2}DPO+PA achieves the strongest recolor grounding (27.3%) and second flip grounding (24.1%), with the best move (46.0%), add (38.5%), multi-op (42.2%), and replace (39.0%), substantially outperforming all compared models. The invariance results also reveal a consistent trend: Ours-m^{2}DPO+PA achieves the best IC on recolor, move, add, and replace, and the second-best IC on remove. Overall, CounterGround3D training with aligned egocentric pose from PoseAlign yields substantially stronger counterfactual grounding capability, increasing sensitivity to relevant interventions while maintaining high invariance to irrelevant changes.

Table 4: Dissected dataset-wise performance on different question sources of CounterGround3D. Invariance is not reported for flip, since the global reflection is never irrelevant by construction. Val sample numbers are reported in parentheses after metric **G** or **I3**.

Model	Recolor · ScanQA		Recolor · Scan2Cap		Flip · ScanQA	Flip · SQA3D	Flip · Scan2Cap
	G (730)	I3 (1606)	G (153)	I3 (1857)	G (160)	G (156)	G (922)
Baseline Models:							
LL3DA	0.1	**97.9**	0.0	29.1	2.5	4.5	0.0
Grounded-3D-LLM	0.0	92.5	0.0	49.4	1.3	1.9	1.4
LEO	3.3	81.0	18.3	45.9	**7.5**	3.2	2.1
3D-LLaVA	5.3	95.5	9.8	87.0	1.3	4.5	1.6
DEER-3D†	7.7	94.8	8.5	**90.6**	1.9	**7.7**	0.7
Our Models:							
Ours-composite	17.5	89.7	17.0	78.5	5.6	4.5	1.8
Ours-m^{2}DPO	**19.6**	86.8	**22.2**	75.9	4.4	5.1	**2.5**
PoseAlign Benchmark:							
LL3DA+PA	0.0	**98.2**	0.0	**89.8**	1.3	0.6	0.3
3D-LLaVA+PA	7.7	96.1	5.2	88.5	6.3	23.1	8.4
Ours-composite+PA	24.7	94.5	18.3	88.3	**46.3**	66.7	13.3
Ours-m^{2}DPO+PA	**28.9**	95.3	**19.6**	89.5	43.1	**67.3**	**13.5**

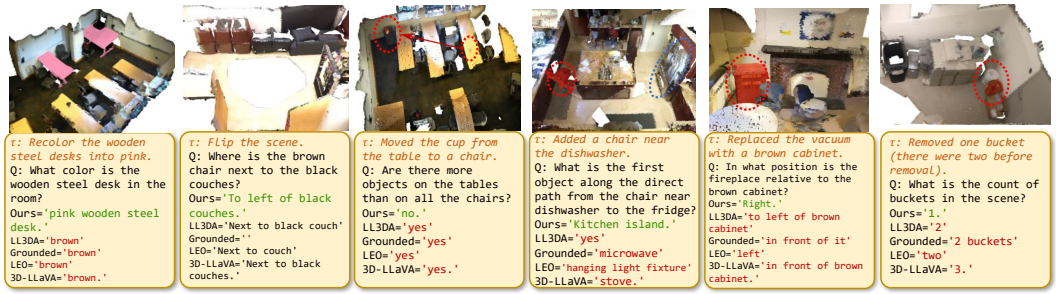


Figure 4: **Qualitative results** on all six subsets. Object interventions are highlighted.

Dataset-wise analysis. We provide a dataset-wise breakdown in Table 4 to localize gains across different source tasks in CounterGround3D. On recoloring, counterfactual training consistently improves grounding, where Ours-m^{2}DPO achieves the best performance on both ScanQA and Scan2Cap, reaching 19.6% and 22.2%, respectively. The flip results are more task-dependent without ego poses from PoseAlign: LEO performs best on ScanQA (7.5%), DEER-3D on SQA3D (7.7%), and Ours-m^{2}DPO on Scan2Cap (2.5%). All flip grounding remain low compared to the PoseAlign setup: Ours-composite+PA reaches 46.3% grounding on ScanQA flip, while Ours-m^{2}DPO+PA achieves the strongest SQA3D and Scan2Cap flip performance at 67.3% and 13.5%, respectively, and also leads recolor grounding on both ScanQA (28.9%) and Scan2Cap (19.6%). Overall, the results show that counterfactual supervision transfers consistently to attribute grounding, whereas directional grounding depends more strongly on the source task and benefits substantially from an explicit egocentric reference frame in the PoseAlign setup.

6 Conclusion

We have introduced **CounterGround3D**, a counterfactual intervention pair benchmark for evaluating visual grounding capability of 3D LMMs. By physically realizing controlled interventions over color, direction, object existence, location, and spatial relationship, CounterGround3D enables joint evaluation of grounding under answer-critical interventions and invariance under answer-preserving interventions. Our experiments show that strong performance on conventional 3D benchmarks does not necessarily translate to reliable counterfactual grounding, exposing substantial insensitivity to visual input across several existing models. To further exploit the paired structure of the benchmark, we proposed **m^{2}DPO**, which combines multimodal preference optimization with counterfactual pairs and supervised counterfactual training. The resulting models achieve substantially stronger grounding especially in the PoseAlign setting, while largely preserving their original task performance. We hope CounterGround3D provides a useful foundation for developing trustworthy 3D LMMs whose predictions are faithfully grounded in the underlying 3D world.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision – ECCV 2020*. Springer, 2020. doi: 10.1007/978-3-030-58452-8_25.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19129–19139, 2022.

Trishna Chakraborty, Udita Ghosh, Xiaopan Zhang, Fahim Faisal Niloy, Yue Dong, Jiachen Li, Amit K Roy-Chowdhury, and Chengyu Song. Heal: An empirical study on hallucinations in embodied agents driven by large language models. *arXiv preprint arXiv:2506.15065*, 2025.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision – ECCV 2020*, pp. 202–221. Springer, 2020. doi: 10.1007/978-3-030-58565-5_13.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: Visual interactive instruction tuning for omni-3D understanding reasoning and planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26428–26438, 2024a.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3D-LLM with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024b.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2Cap: Context-aware dense captioning in RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3193–3203, 2021.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3D-LLaVA: Towards generalist 3d lmms with omni superpoint transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3772–3782. IEEE, 2025.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 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 (CVPR)*, pp. 14375–14385, 2024. doi: 10.1109/CVPR52733.2024.01363.

Peizheng Guo, Jingyao Wang, Wenwen Qiang, Jiahuan Zhou, Changwen Zheng, and Gang Hua. COPO: Causal-oriented policy optimization for hallucinations of MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11051–11063, 2026.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. *arXiv preprint arXiv:2311.12871*, 2023.

Jiangyong Huang, Xiaojian Ma, Xiongkun Linghu, Junchao He, Qing Li, Song-Chun Zhu, Yixin Chen, Baoxiong Jia, and Siyuan Huang. Leo-vl: Efficient scene representation for scalable 3d vision-language learning. *arXiv preprint arXiv:2506.09935*, 2025.

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

Guanghao Jin, Jingpei Wu, Tianpei Guo, Yiyi Niu, Weidong Zhou, Linyi Yang, and Guoyang Liu. KnowDR-REC: Auditing knowledge-conditioned visual grounding in referring expression comprehension. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38607–38629. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1923.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. VLind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.231.

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

Xintong Li, Chuhan Wang, Junda Wu, Rohan Surana, Tong Yu, Julian McAuley, and Jingbo Shang. Importance sampling for multi-negative multimodal direct preference optimization. In *International Conference on Learning Representations*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Quan Liu, Weihao Xuan, Junjue Wang, Naoto Yokoya, Ling Shao, and Shijian Lu. Direction-aware 3d large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9668–9678, 2026.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934. IEEE, 2025.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Prisacariu. Do 3d large language models really understand 3d spatial relationships? In *International Conference on Learning Representations*, volume 2026, pp. 32591–32610, 2026.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *Proceedings of the International Conference on Learning Representations*, 2023a.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023b.

Ye Mao, Weixun Luo, Junpeng Jing, Anlan Qiu, and Krystian Mikolajczyk. Hypo3d: Exploring hypothetical reasoning in 3d. *arXiv preprint arXiv:2502.00954*, 2025.

Giorgio Mariani, Alessandro Raganato, Simone Melzi, and Gabriella Pasi. Glue3d: General language understanding evaluation for 3d point clouds. *Information Fusion*, pp. 104007, 2025.

OpenAI. ChatGPT (GPT-5), 2025a. URL https://chat.openai.com/. Large language model.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025b. URL https://arxiv.org/abs/2508.10925.

Nhi Pham and Michael Schott. H-POPE: Hierarchical polling-based probing evaluation of hallucinations in large vision-language models. *arXiv preprint arXiv:2411.04077*, 2024.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1437.

Ayan Sar, Pranav Puri, Anurag Kaushish, Sampurna Roy, Sumit Aich, Tanupriya Choudhury, and Ajith Abraham. VLM Reality Check: A causal counterfactual benchmark for diagnosing cognitive biases in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 4361–4375, 2026.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7657–7666. IEEE, 2019.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An llm-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Shuguang Cui, and Zhen Li. Comprehensive visual question answering on point clouds through compositional scene manipulation. *IEEE Transactions on Visualization and Computer Graphics*, 30(12):7473–7485, 2023.

Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3D-GRAND: A million-scale dataset for 3D-LLMs with better grounding and less hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29501–29512. IEEE, 2025a.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10610–10620, 2025b.

Dokyoon Yoon, Youngsook Song, and Woomyoung Park. Stop learning it all to mitigate visual hallucination, focus on the hallucination target. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4200–4208, 2025.

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

Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1791–1800, 2021.

Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-Trans2Cap: Cross-modal knowledge transfer using transformer for 3D dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8563–8573, 2022.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022.

Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14141–14151, 2024. doi: 10.1109/CVPR52733.2024.01341.

Weichen Zhang, Ruiying Peng, Xin Zeng, Jianjie Fang, Ziyou Wang, Kaiyuan Li, Heng Dong, Wei Li, Chen Gao, Xin Wang, et al. The point, the vision and the text: Does point cloud boost spatial reasoning of large language models? a bias-controlled study. *arXiv preprint arXiv:2504.04540*, 2025.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3DRefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15225–15236, 2023.

Yue Zhang, Zun Wang, Han Lin, Jialu Li, Jianing Yang, Yonatan Bitton, Idan Szpektor, and Mohit Bansal. Error-driven scene editing for 3d grounding in large language models. In *European Conference on Computer Vision (ECCV)*, 2026. arXiv:2511.14086.

A Appendix

A.1 AI Use Statement

In this work, we used generative AI tools for some tasks requiring disclosure, primarily in dataset generation and curation. As described in Section 3, GPT-5.4 was used to assist with grounding change-related spans in questions and answers, disambiguating object and attribute fields, and processing Hypo3D counterfactual descriptions, including parsing, feasibility checking, and ambiguity resolution. These uses contributed to the generation and curation of the resulting counterfactual dataset. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine research hypotheses, design experiments or research methodology, implement the proposed method, perform qualitative or thematic analysis, or interpret experimental results; other required disclosure categories are not applicable to this work.

Additionally, we used generative AI tools for tasks for which disclosure is recommended. GPT-5.6 Sol was used throughout the manuscript to improve readability and presentation, including restructuring, word choice, typo detection, and formatting. The official ICLR PAT system was also used to provide feedback during final manuscript polishing. All AI-assisted dataset outputs were reviewed and validated by the authors as part of the dataset construction process, and all AI-assisted manuscript edits were reviewed and approved by the authors. We take responsibility for the final content of this work, including all text, claims, data, and other artifacts produced with the aid of generative AI.

A.2 Proof of Correctness of an Arbitrary Flip

Let the scene use vertical axis $\mathbf{u}$ and let $\mathbf{n}$ denote the horizontal normal of the selected reflection plane.

A general-form global reflection is written as

$$p_i^\tau = \left(I - 2\mathbf{n}\mathbf{n}^\top\right)(p_i - o) + o, \tag{13}$$

where o is a point on the reflection plane.

Following the definition of Equation 13, we show that any normal vector $\mathbf{n} = (\cos\theta, \sin\theta, 0)$ flips the left/right relation.

Lemma 1. *Let*

$$R_\mathbf{n} = I - 2\mathbf{n}\mathbf{n}^\top \tag{14}$$

be the reflection matrix associated with any vertical plane whose unit normal $\mathbf{n}$ is horizontal, i.e., $\mathbf{n}^\top\mathbf{u} = 0$. Let $\mathbf{f}$ be a horizontal forward direction defining the reference frame, and let

$$\mathbf{d} = p_t - p_r \tag{15}$$

denote the displacement from a reference point p_r to a target point p_t. Define the signed left/right relation as

$$\ell(\mathbf{f}, \mathbf{d}) = \mathbf{u}^\top(\mathbf{f} \times \mathbf{d}). \tag{16}$$

If the scene and its reference direction are transformed jointly by $R_\mathbf{n}$, then

$$\ell(R_\mathbf{n}\mathbf{f}, R_\mathbf{n}\mathbf{d}) = -\ell(\mathbf{f}, \mathbf{d}). \tag{17}$$

Therefore, every non-degenerate left/right relation is reversed by reflection about any vertical plane, independently of the azimuth θ of its horizontal normal $\mathbf{n} = (\cos\theta, \sin\theta, 0)$.

Proof. Because the reflection plane is vertical, its normal is orthogonal to the vertical axis:

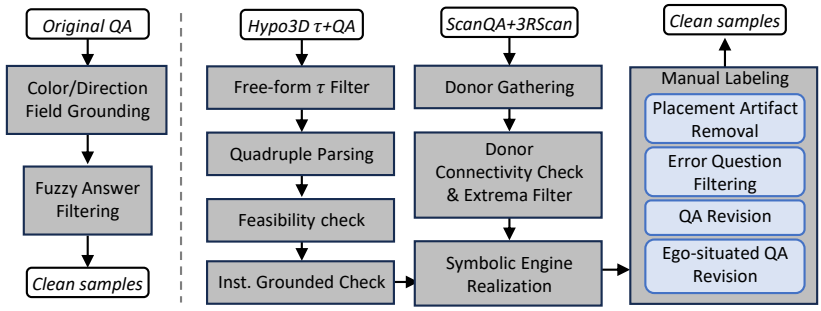
$$\mathbf{n}^\top\mathbf{u} = 0. \tag{18}$$

Hence

$$\mathbf{u}^\top R_\mathbf{n} = \mathbf{u}^\top\left(I - 2\mathbf{n}\mathbf{n}^\top\right) = \mathbf{u}^\top. \tag{19}$$

Moreover, $R_\mathbf{n}$ is an orthogonal reflection, and therefore

$$R_\mathbf{n}^\top R_\mathbf{n} = I, \qquad \det(R_\mathbf{n}) = -1. \tag{20}$$



(a) Recolor & Flip (b) Move, Add, Replace, Remove

Figure 5: The dataset generation pipeline for CounterGround3D.

For any orthogonal transformation R and vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$, the scalar triple product satisfies

$$(R\mathbf{a})^\top \big((R\mathbf{b}) \times (R\mathbf{c})\big) = \det(R)\, \mathbf{a}^\top (\mathbf{b} \times \mathbf{c}). \tag{21}$$

Applying this identity and applying $\mathbf{u}^\top R_{\mathbf{n}} = \mathbf{u}^\top$ yields

$$\begin{aligned} \ell(R_{\mathbf{n}}\mathbf{f}, R_{\mathbf{n}}\mathbf{d}) &= \mathbf{u}^\top \left((R_{\mathbf{n}}\mathbf{f}) \times (R_{\mathbf{n}}\mathbf{d})\right) && (22)\\ &= \det(R_{\mathbf{n}})\, \mathbf{u}^\top R_{\mathbf{n}}(\mathbf{f} \times \mathbf{d}) && (23)\\ &= \det(R_{\mathbf{n}})\, \mathbf{u}^\top (\mathbf{f} \times \mathbf{d}) && (24)\\ &= -\ell(\mathbf{f}, \mathbf{d}). && (25) \end{aligned}$$

Thus, the sign of the left/right predicate is reversed. The degenerate case $\ell(\mathbf{f}, \mathbf{d}) = 0$ corresponds to a target lying exactly on the forward–backward axis and is therefore neither strictly left nor strictly right. Since the proof depends only on $\mathbf{n}^\top \mathbf{u} = 0$ and $\det(R_{\mathbf{n}}) = -1$, and not on the specific value of θ, the result holds for any vertical reflection plane. □

A.3 Details for Object Intervention Realization

The detailed dataset generation pipeline is displayed in Figure 5(b). In the following sections, We will introduce the modules for object inervention sequentially.

Edit parsing and feasibility. Given a free form description h from Hypo3D, we map it to a quadruple (o, r_t, r_a, ρ), where o is the operation, r_t and r_a are textual references to the target and optional anchor, and ρ is a normalized spatial relation. An LLM performs extraction under a fixed schema (OpenAI, 2025b), followed by a feasibility check. Relations are mapped to a closed vocabulary, while descriptions requiring unresolved scene-dependent reasoning are rejected rather than guessed from text.

Instance grounding. The target and anchor references are subsequently mapped to scene instances,

$$\Gamma(r, \mathcal{P}) \to j, \qquad j \in \{1, \ldots, K\}, \tag{26}$$

using a qualifier-aware cascade. We first resolve unique semantic-label matches, followed by positional qualifiers, attributes, room-region descriptors, and egocentric extrema. Object centers and the estimated room footprint provide the geometric quantities required by positional and regional references. If no unique instance can be established, the edit is discarded. Thus, only examples satisfying both

$$\text{Feasible}(h) = 1, \qquad \text{Groundable}(h, \mathcal{P}) = 1 \tag{27}$$

enter the executor.

Physical realization. Given grounded target j_t and anchor j_a, the executor applies the corresponding transformation directly to the point cloud. For `remove`, all points assigned to j_t are deleted. For `move`, the target is translated from center μ_t to a relation-specific destination $\hat{\mu}_t$:

$$p_i^{\tau} = p_i + (\hat{\mu}_t - \mu_t), \qquad i \in M_{j_t}, \tag{28}$$

where $\hat{\mu}_t$ is selected under in-room and least-collision constraints. Vertical relations are realized through bounding box stacking and a slow iterative descent which aborts upon contact, while horizontal and distance relations use anchor-relative placement on the room occupancy model.

Donor-bank curation. The `add` and `replace` operations may require geometry not present in the current point cloud. We therefore construct a donor bank of approximately 65K real object instances mined from training split in ScanNet and 3RScan. Raw instance segments, however, frequently contain severe occlusion, disconnected fragments, over-segmentation, or sparse geometry, which can produce implausible artifacts after transplantation. We consequently curate each donor using complementary geometric and category-level filters. First, we identify fragmented instances using a radius-based connectivity test: points within $0.10\,\mathrm{m}$ are connected, and a donor is rejected if its largest connected component contains fewer than 85% of its points. Second, within each semantic class, we reject segmentation outliers according to robust statistics of log-volume and log-density. Specifically, instances whose median/MAD-based robust z-score exceeds 3.5 in either quantity are removed. We deliberately exclude aspect ratio from this criterion, as several valid categories, such as doors, doorframes, and whiteboards, are naturally thin. To preserve class coverage, at least five donors are retained for every insertable category. This curation removes 7,384 of 65,340 candidate donors (11.3%), including 6,040 fragmented instances, 1,200 volume outliers, and 144 low-density outliers, while maintaining coverage over all 192 insertable object classes. After filtering, donors are selected by semantic class, with preference for same-dataset and size-compatible instances. Together, these steps reduce donor-specific geometric artifacts before the subsequent orientation, placement, and room-overflow checks.

Quality control. We distinguish interventions by their potential to introduce editing artifacts. Recoloring, object removal, and flipping modify existing points without synthesizing geometry and are therefore the cleanest interventions. Moving an object may expose previously occupied regions, while addition and replacement rely on donor geometry. We retain the latter to broaden the range of interventions, but apply semantic donor filtering, orientation correction, collision checks, room-overflow rejection, and human expert correction to minimize edit-specific cues.

A.4 Detailed Experiment Setup

Additional definition of metrics. For the invariance tests, IC in theory should be defined as the percentage passing IC and L-A_o; However, since object interventions do not have pre-edit answer, we can only calculate their invariance to a second irrelevant intervention τ' applied upon the first answer-critical intervention τ, and compute IC as both passing I3 and L-A_e. The LLM-as-Judge uses GPT-OSS-20B from OpenAI (2025b) for local deployment, and uses the same prompt template as Liu et al. (2026). Besides, the LLM-as-Judge protocol is not directly suited to evaluating CounterGround3D, because ground-truth answers to the same question may differ in whether they explicitly specify color. For example, ScanQA includes both "small stool" and "tiny black stool" as ground-truth answers to the same question. Under such a protocol, a prediction that identifies the correct object but assigns it an incorrect color may still be judged correct. To address this issue, we augment the in-context examples provided to the judging LLM to emphasize the scene property affected by each intervention, as listed in Appendix A.7. These examples clarify that correctly identifying the object alone is insufficient when the prediction misrepresents the intervention-relevant property.

Training protocol. The ours-composite model is trained with the same set of parameters as the SFT stage of 3D-LLaVA (Deng et al., 2025). Specifically, the projector and LoRA component of the LLM is trained on ScanQA, SQA3D, Scan2Cap, ScanRefer, Multi3DRefer, Nr3D caption, and our CounterGround3D data with plain dataset concatenation. The parameters are trained using AdamW with 2×10^{-4} learning rate and cosine annealing for one epoch. The ours-m^{2}DPO model inherits its weights from ours-composite and is further tuned with m^{2}DPO with 4×10^{-5} learning rate for another epoch. Because the object interventions including move, add, replace, and remove do not

Table 5: Manual refinement statistics for the object-level counterfactual subsets. "PC changed" denotes target items whose class/placement/instance-selection are refined; Q/A changed denotes revised questions/answers. The revision targets object instance correction, placement alignment, artifact elimination, and removal of time-related description from Hypo3D.

Operation	Original	Remaining	Discarded	PC changed	Q changed	A changed
Move	2,589	2,338	251 (9.7%)	1,847 (71.3%)	1,222 (47.2%)	446 (17.2%)
Add	1,178	1,123	55 (4.7%)	979 (83.1%)	814 (69.1%)	132 (11.2%)
Replace	1,004	942	62 (6.2%)	822 (81.8%)	678 (67.5%)	179 (17.8%)
Remove	722	707	15 (2.1%)	0 (0.0%)	362 (50.1%)	0 (0.0%)
All	**5,493**	**5,110**	**383 (7.0%)**	**3,648 (66.4%)**	**3,076 (56.0%)**	**757 (13.8%)**

Table 6: Dataset statistics of CounterGround3D across different intervention types.

Split	Recolor	Flip	Move	Add	Replace	Remove
Training	14,069	10,570	1,971	954	770	622
Validation	883 / 3463	1,238	367	169	172	85

contain a pre-intervention answer, only two eligible loss terms including $\mathcal{L}^{\tau}_{\mathrm{CoPO}}$ and $\mathcal{L}^{\tau}_{\mathrm{SFT}}$ are used for these subsets during m^{2}DPO. In addition, since CoPO prefers $(\mathcal{P}^{\tau}, q^{\tau}, a^{\tau})$ over $(\mathcal{P}, q^{\tau}, a^{\tau})$, and prefers $(\mathcal{P}^{\tau}, q, a)$ over $(\mathcal{P}, q, a)$ which are all correct samples given an irrelevant intervention, irrelevent examples plus CoPO loss leads to incorrect preference signals which penalize invariance and consequently decreased invariance performance numbers. Therefore, we drop irrelevant interventions during m^{2}DPO training to maintain supervision regularity. All experiments are conducted on 4 H100 GPUs with a batch size of 8 and gradient accumulation of 2.

Different amount of validation samples for *G* and *I*. As listed in Table 4, the amount of samples for grounding validation is significantly less than those for invariance validation. This is intended because grounding needs a clean and non-ambiguous answer set which should contain one color or one direction only, while discarding the fuzzy descriptions where the color or the direction is not mentioned or inconsistent across the ground-truth answer list. For example, a ground truth answer list of ['black', 'brown'] is inappropriate for accurate color grounding; ['left of the bench', 'near bench'] is inappropriate for precise direction grounding. After screening the fuzzy examples, the remaining set is the clean eval set which is used for all grounding experiments. The invariance experiment, on the other hand, is not affected by this limitation and can use the whole evaluation set and the original ground truth answer list. Therefore, invariance evaluation samples are super-sets of grounding samples and naturally contain more samples in them.

Manual refinement statistics for object interventions. While the object placements generated by the object intervention realization pipeline is almost perfect, we further conduct expert manual labeling and filtering to minimize risks of placement artifacts, patch the mismatched question and edit inherited from Hypo3D, purge cases where placement is unrealistic or full of artifacts, and unify the direction description in Hypo3D into a unified egocentric situated description. Specifically, we move or re-designate the target/anchor objects, refine their placement location and orientation, and revise the question and answer texts. The change-related descriptions, e.g., "current", "after the move", "the moved object", "what...now", "newly added", "original position", "now that", *etc.,* are all deleted with the remaining question texts manually re-normalized. Table 5 shows that manual refinement modifies a substantial section of the automatically generated object interventions. Of the 5,493 original items, 5,110 are retained, corresponding to an overall discard rate of 7.0%. Among all samples, 66.4% refined the object location/selection, 56.0% receive a modified question, and 13.8% with an improved answer. These statistics motivate manual verification as a critical step for constructing high-quality counterfactual 3D benchmarks. The labeling program will be open-sourced alongside the codebase.

Dataset statistics. Table 6 summarizes the scale of CounterGround3D across the six intervention types. Recolor and flip constitute the largest subsets, with 14,069/883 and 10,570/1,238 training/validation samples, respectively, while the object-level interventions are smaller due to stricter geometric realization and manual refinement. Among them, move contains the most examples, fol-

Table 7: Ablation of the m^{2}DPO components at a learning rate of 4×10^{-5} on the recolor training set. C, A, L, P, and S denote CoPO($^\tau$), AncPO, DPO($^\tau$), SFT on original data, and SFT$^\tau$ on edited data, respectively. mDPO only calculates on the preferred data ($^\tau$) which is asymmetric, which is denoted by '*'. Metrics are based on recolor EM for efficient implementation.

Configuration	Training Component					Recolor		
	C	A	L	P	S	G	EM_o	EM_e
Ours-composite	–	–	–	–	–	14.1	32.4	21.7
mDPO	✓*	✓*	✓*	–	–	0.8	3.7	23.0
C	✓	–	–	–	–	0.0	0.4	1.2
CA	✓	✓	–	–	–	7.5	17.3	24.7
CAS	✓	✓	–	–	✓	6.7	16.3	**25.8**
CAP	✓	✓	–	✓	–	**15.8**	<u>32.5</u>	25.4
CALP	✓	✓	✓	✓	–	14.5	30.3	25.3
CAPS	✓	✓	–	✓	✓	15.3	31.0	25.7
CAPLS	✓	✓	✓	✓	✓	13.8	28.3	25.4
CASL	✓	✓	✓	–	✓	3.7	9.2	25.5
CLP	✓	–	✓	✓	–	14.8	30.9	25.0
CPS	✓	–	–	✓	✓	<u>15.6</u>	31.6	<u>25.8</u>
SP	–	–	–	✓	✓	15.5	**32.7**	24.3

lowed by add, replace, and remove. Recolor validation contains 883 clean samples for grounding and 3,463 samples for invariance as discussed in A.4. Flip does not calculate invariance, while object interventions are all single-answer clean samples through manual labeling, so they have the same amount of samples for grounding and invariance. This distribution reflects the different availability and construction difficulty of each intervention type while maintaining dedicated validation sets for all six operations.

Integration with PoseAlign. CounterGround3D naturally extends to the direction-aware setting of PoseAlign, which addresses the directional ambiguity of existing 3D benchmarks by explicitly aligning point clouds to an ego reference frame (Liu et al., 2026). The PoseAlign-Transform method maps the input scene into the recovered camera coordinate system so that directional relations such as "left" and "right" are expressed consistently in the egocentric frame, without requiring downstream architectural changes. SQA3D directly reuses its official annotated ego pose, with the camera translated to a human viewing height according to the action context: 1.3 m for sitting, 1.7 m for standing, and 1.5 m for unmatched cases. For the recolor subset, we reuse the pose associated with the original ScanQA, SQA3D, or Scan2Cap query provided in the paper. For the flip subset, since PoseAlign leverages an egocentric front-left-up coordinate system corresponding to xyz axes, we apply the flip intervention on egocentric y axis instead of global x axis to avoid disrupting front direction definition. For object interventions, we reuse PoseRecover pipeline to obtain the ego poses that inspect the task-relevant object. Specifically, `move`, `add`, and `replace` use the post-intervention object, falling back to the anchor object; `remove` uses the removed target as if it is still there. Then, the pipeline in PoseRecover takes these target objects and filters all camera extrinsics to find the best-match ego poses. For composite edits, we prioritize the first geometric sub-operation under the same rule, then the recolor target, and finally a manual designation for a few out-of-sight objects. This integration preserves CounterGround3D's intervention semantics while providing a well-defined egocentric coordinate frame for evaluating counterfactual directional grounding.

A.5 Additional Experiments

Ablation of m^{2}DPO. In this experiment, we explore the usage of different loss components of m^2-DPO in Table 7. Symmetric terms of $(\mathcal{L}_{\text{CoPO}}, \mathcal{L}^{\tau}_{\text{CoPO}})$ and $(\mathcal{L}_{\text{DPO}}, \mathcal{L}^{\tau}_{\text{DPO}})$ are tuned with the same set of parameters, w_{CoPO} and w_{DPO}, except for mDPO setting which only trains its original asymmetric losses on counterfactual samples with $\mathcal{L}^{mdpo}_{\text{CoPO}}, \mathcal{L}^{mdpo}_{\text{DPO}}, \mathcal{L}^{mdpo}_{\text{AncPO}}$. Metrics are EM-based to speed up experiments. CoPO alone collapses grounding to 0.0, and adding AncPO only partially recovers performance (G=7.5%). Multimodal DPO (mDPO) fails catastrophically on original unedited data (Wang et al., 2024) due to the lack of symmetric loss design and additional SFT loss.

Table 8: Loss weight tuning experiment around the CALPS configuration ($w_{\mathrm{CoPO}} = 1$, $w_{\mathrm{AncPO}} = 1$, $w_{\mathrm{DPO}} = 1$, $w_{\mathrm{SFT_o}} = 0.5$, $w_{\mathrm{SFT_e}} = 0.5$, and $\beta = 0.1$). G denotes strict grounded-pair accuracy. All experiments use a learning rate of 4×10^{-5}. Metrics are based on recolor EM for efficient implementation.

Hyperparameter	Weight	G	$\mathrm{EM_o}$	$\mathrm{EM_e}$
Ours-m^{2}DPO (CALPS)	default	16.1	29.8	30.6
w_{CoPO}	0.5	16.2	29.8	30.6
	2.0	15.6	29.8	30.3
w_{AncPO}	0.5	15.8	29.5	30.2
	2.0	15.8	29.4	30.3
w_{DPO}	0.0	16.9	31.4	30.2
	0.5	16.6	30.5	30.3
	2.0	14.4	28.3	29.5
$w_{\mathrm{SFT_o}}$	0.25	14.5	27.0	30.3
	1.0	**17.2**	**31.6**	29.7
$w_{\mathrm{SFT_e}}$	0.25	15.9	30.5	30.1
	1.0	16.0	28.8	**30.7**
β	0.05	<u>17.1</u>	<u>30.9</u>	30.3
	0.20	15.9	29.0	30.6

Table 9: Native ScanRefer grounding performance of our DEER-3D† re-implementation. Acc@0.25 and Acc@0.5 are computed using instance mask IoU. R1 and R2 represent two rounds of iterative self-training from 3D-LLaVA weights.

Model	mIoU	Acc@0.25	Acc@0.5
3D-LLaVA	42.6	60.6	41.5
DEER-3D† R1	45.0	<u>63.0</u>	<u>44.9</u>
DEER-3D† R2	**47.8**	**65.1**	**49.0**

In contrast, introducing original data SFT yields a substantial improvement, where CAP reaches the highest grounding score of 15.8, exceeding the 14.1 baseline while retaining comparable $\mathrm{EM_o}$. Pure supervised training (SP, equivalent of Ours-composite) attains the highest $\mathrm{EM_o}$ but is slightly weaker in grounding, suggesting that preference supervision contributes additional counterfactual sensitivity beyond SFT alone. Notably, adding the language preference term consistently degrades performance, e.g., CAP $\succ$ CALP and CAPS $\succ$ CAPLS, indicating limited benefit from DPO language optimization for visually grounding objective. Adding SFT to the preservation-based formulation provides a more balanced configuration: CAPS maintains strong grounding and edited-scene accuracy while combining counterfactual preference and supervised signals. Therefore, we designate CAPS as default in subsequent experiments.

Loss-weight ablation. Table 8 studies each loss weight independently. While CAPS is used for m^{2}DPO by default, Table 8 selects the anchor as CALPS because all loss terms are needed to display tuning effect, otherwise we cannot examine the L dimension. In this experiment, the The results show that w_{CoPO} and w_{AncPO} are relatively insensitive within the tested range, with only minor changes in grounding and exact-match accuracy. In contrast, $w_{\mathrm{SFT_o}}$ provides the strongest control over the grounding-accuracy tradeoff: increasing it from 0.5 to 1.0 yields the highest grounding score of 17.2% and $\mathrm{EM_o}$ of 31.6%, while reducing it to 0.25 substantially degrades both metrics. The language term exhibits a similar sensitivity, with a large weight $w_{\mathrm{DPO}} = 2$ reducing grounding to 14.4%, while weaker language supervision improves both grounding and $\mathrm{EM_o}$. A smaller DPO temperature, $\beta = 0.05$, also provides a favorable balance where 17.1% grounding and 30.9% $\mathrm{EM_o}$. Overall, the sweep indicates that SFT strength, language weighting, and β are the primary performance levers, while the CoPO and AncPO weights are less sensitive.

A.6 Reimplementation of DEER-3D

DEER-3D improves upon 3D-LLaVA, but inconsistently. Our re-implemented DEER-3D† improves over 3D-LLaVA on several geometry-changing interventions, but the gains are uneven across

Table 10: Transfer of DEER-3D† to CounterGround3D. Recolor and flip are evaluated on the clean subsets, while object operations are evaluated on the manually refined validation sets. G denotes strict grounded-pair accuracy, and L-A_o / L-A_e denote LLM-as-Judge accuracy on the original / edited scenes, respectively.

Intervention	Metric	3D-LLaVA	DEER-3D R1	DEER-3D R2-cont
Recolor	G	6.1	5.9	**7.8**
	L-A_o	56.1	54.4	**57.4**
	L-A_e	12.6	13.1	**14.4**
Flip	G	1.9	**2.6**	1.7
	L-A_o	19.1	**19.4**	18.0
	L-A_e	16.8	15.8	**17.1**
Move	L-A_e	39.0	38.1	**40.1**
Add	L-A_e	29.0	32.0	**32.5**
Replace	L-A_e	**32.0**	**32.0**	30.2
Remove	L-A_e	25.9	29.4	**30.6**
Multi-op	L-A_e	21.9	20.3	**25.0**

operation types as listed in Table 2-3. In particular, DEER-3D† raises recolor grounding from 6.1% to 7.8%, add from 29.0% to 32.5%, remove from 25.9% to 30.6%, and multi-op from 21.9% to 25.0%. In contrast, it underperforms 3D-LLaVA on flip and replace. This pattern because DEER-3D primarily manipulates surrounding objects through cloning, surrounding-object replacement, recoloring, and rotation, and therefore transfer more naturally to perturbations resembling movement, addition, removal, and color changes. By comparison, CounterGround3D directly intervenes on answer-critical targets with global reflection, object addition, and target replacement, creating stronger interventions that are not covered by the edit profile of DEER-3D.

DEER-3D transfer analysis. Tables 9 and 10 show that stronger native ScanRefer grounding does not translate uniformly to CounterGround3D. On ScanRefer, DEER-3D† improves monotonically across self-training rounds, with R2 reaching 47.8 mIoU and $65.1/49.0$ Acc@0.25/0.5, clearly surpassing the 3D-LLaVA initialization. On CounterGround3D, however, transfer remains intervention-dependent. R2-cont improves over 3D-LLaVA on clean recolor ($G = 7.8\%$), move (40.1%), add (32.5%), remove (30.6%), and multi-op (25.0%), while R1 attains the strongest clean-flip grounding at 2.6%. Replace is the main exception, where 3D-LLaVA and R1 both achieve 32.0%, above R2-cont at 30.2%. Overall, DEER-3D† demonstrates that task-specific grounding improvements can partially transfer to broader counterfactual reasoning, but the gains are neither monotonic across self-training rounds nor consistent across intervention types. This suggests that exposure to a restricted family of counterfactual edits is insufficient for general grounding, motivating the broader operation coverage and paired supervision adopted in CounterGround3D.

A.7 LLM-AS-JUDGE SCRIPTS

Improved template for recolor subset. The following prompt is for stricter LLM-as-Judge evaluation on the recolor subset on ScanQA. Other datasets use the same structure but different examples, which will be revealed in the code release.

```
Your job is to look at a question, a ground-truth answer, and a predicted answer, and then
    assign a grade of either ["CORRECT", "INCORRECT"].
This is a COLOUR-critical task: the question is about the colour of an object, so the grade
    turns on whether the predicted COLOUR is right.
First, I will give examples of each grade, and then you will grade a new example.


The following are examples of CORRECT predicted answers.
```
Question 1: What color is the chair by the tv?
Ground truth: "green", "green office chair"
Predicted answer 1: green.
Predicted answer 2: the chair is green.
Predicted answer 3: dark green chair.

Question 2: What color is the counter below the two windows?
Ground truth: "black", "long black counter"
```
```

```
Predicted answer 1: black.
Predicted answer 2: black kitchen counter.

Question 3: What is on the desk?
Ground truth: "a mug", "grey mug"
Predicted answer 1: mug.
Predicted answer 2: a mug.
```
These predicted answers are all CORRECT because:
    - The colour named in the prediction matches a colour in one of the ground truths. The object name, aliases, phrasing, shade qualifiers, capitalization, punctuation and word order do not matter, as long as the base colour is right.
    - "grey"/"gray" are the same colour, and a shade like "dark green" still counts as "green".
    - Question 3 is the ONLY sparing case: one ground truth ("a mug") contains NO colour word, and the prediction also contains no colour word and matches its meaning exactly. A colourless prediction is acceptable ONLY in this situation, and it should be rare.

The following are examples of INCORRECT predicted answers.
```
Question 1: What color is the chair by the tv?
Ground truth: "green", "green office chair"
Predicted answer 1: brown.
Predicted answer 2: blue chair.
Predicted answer 3: office chair.
Predicted answer 4: the chair.

Question 2: What color is the counter below the two windows?
Ground truth: "black", "long black counter"
Predicted answer 1: kitchen counter.
Predicted answer 2: long counter.
Predicted answer 3: metallic.
```
These predicted answers are all INCORRECT because:
    - The colour is wrong: the prediction names a different colour than every ground truth (e.g. "brown" or "blue" for a "green" object).
    - Or the prediction omits the colour while the applicable ground truth requires one: naming the object ("office chair", "the chair", "kitchen counter") without its colour is INCORRECT on this colour-critical subset. Having the object right is not enough; the colour is the point.
    - A non-colour word such as "metallic" is not a colour and does not exactly match a colourless ground truth, so it is INCORRECT.

Also note the following things:
- Do not reward getting the object right when the colour is wrong or missing.
- If several colours are named, the ground-truth colour must be present and none of the named colours may contradict it.
- The colourless-match exception (see Question 3 above) applies ONLY when a ground-truth answer itself has no colour word AND the prediction has no colour word AND they match in meaning. Use it sparingly; when in doubt, require the colour.

Here is a new example.
```
Question: {question}
Ground truth: {ground_truth}
Predicted answer: {predicted_answer}
```

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT

Just return the letters "A" or "B", with no text around it.

Improved template for flip subset. The following prompt is for stricter LLM-as-Judge evaluation on the flip subset on ScanQA. Other datasets use the same structure but different examples, which will be revealed in the code release.

```
Your job is to look at a question, a ground-truth answer, and a predicted answer, and then assign a grade of either ["CORRECT", "INCORRECT"].
This is a DIRECTION-critical task: the answer hinges on a left/right direction, so the grade turns on whether the predicted left/right side is right.
First, I will give examples of each grade, and then you will grade a new example.
```

```
The following are examples of CORRECT predicted answers.
```
Question 1: How is the chair situated in relation to the fan now?
Ground truth: "left", "to the left of the fan"
Predicted answer 1: left.
Predicted answer 2: to the left.
Predicted answer 3: on the left of the fan.

Question 2: Where is the whiteboard with respect to the shelf after the change?
Ground truth: "front left", "to the front-left of the shelf"
Predicted answer 1: front left.
Predicted answer 2: on the left, in front of the shelf.

Question 3: What is nearest in front of the couch now?
Ground truth: "office chair", "a chair"
Predicted answer 1: office chair.
Predicted answer 2: a chair.
```
These predicted answers are all CORRECT because:
    - The left/right side named in the prediction matches the left/right side in one of the
        ground truths. Extra words (front/back, object names), aliases, phrasing,
        capitalization, punctuation and word order do not matter, as long as left versus
        right is right.
    - Question 3 is the ONLY sparing case: the ground truths contain NO left/right word, and
        the prediction also contains no left/right word and matches the meaning exactly. A
        prediction with no left/right term is acceptable ONLY in this situation, and it
        should be rare.


The following are examples of INCORRECT predicted answers.
```
Question 1: How is the chair situated in relation to the fan now?
Ground truth: "left", "to the left of the fan"
Predicted answer 1: right.
Predicted answer 2: to the right of the fan.
Predicted answer 3: next to the fan.
Predicted answer 4: in front of the fan.

Question 2: Where is the whiteboard with respect to the shelf after the change?
Ground truth: "front left", "to the front-left of the shelf"
Predicted answer 1: front right.
Predicted answer 2: in front of the shelf.
```
These predicted answers are all INCORRECT because:
    - The side is flipped: the prediction names the opposite left/right of every ground truth
        (e.g. "right" for a "left" answer).
    - Or the prediction omits the left/right term while the ground truth requires one: "next
        to the fan", "in front of the shelf" give no left/right and are INCORRECT on this
        direction-critical subset. Getting the object or the front/back right is not enough;
        the left/right side is the point.
    - front/back alone do not satisfy a ground truth that specifies left or right.


Also note the following things:
- Do not reward getting the object or the front/back right when the left/right side is wrong
    or missing.
- If a side is combined with other directions (e.g. "front left"), the left/right part must
    match the ground truth's left/right part.
- The no-left/right exception (see Question 3 above) applies ONLY when a ground-truth answer
    itself has no left/right word AND the prediction has no left/right word AND they match in
     meaning. Use it sparingly; when in doubt, require the correct left/right side.


Here is a new example.
```
Question: {question}
Ground truth: {ground_truth}
Predicted answer: {predicted_answer}
```

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT

Just return the letters "A" or "B", with no text around it.
```