# 3D-VCD: Hallucination Mitigation in 3D-LLM Embodied Agents through Visual Contrastive Decoding

Supplementary Material

## A. Implementation Details

We build on the released 3D-LLM [12] checkpoints from 3D-GRAND [43], which adopt a LLaMA-style causal decoder that ingests object-centric scene graphs and autoregressively generates natural-language answers. We do not finetune or modify any model parameters; instead, 3D-VCD operates purely at inference time by contrastively manipulating token logits. Each scene is serialized into text using object category, centroid $(x, y, z)$, and extent $(w, h, d)$ attributes extracted from the dataset JSON. To construct the distorted negative context required for contrastive decoding, we inject zero-mean Gaussian noise into object centroids and extents, varying the standard deviation across $\sigma \in \{0.01, 0.05, 0.15, 0.45\}$ for sensitivity analysis. We prepend both the original and distorted scene tokens to the user query and perform dual decoding passes to obtain logit sequences $z_{\text{orig}}$ and $z_{\text{dist}}$, computing the final logits as $z_{\text{vcd}} = (1 + \alpha)z_{\text{orig}} - \alpha z_{\text{dist}}$ with $\alpha = 1.0$. We decode using greedy search with temperature $T = 1.0$ and batch size 8. For HEAL experiments, we apply 3D-VCD to off-the-shelf language-only models (Llama-3-8B-Instruct and Qwen-14B-Instruct) by feeding them scene descriptions in the HEAL text format, requiring no spatial inference modules. Scene graphs are preprocessed once and cached for all runs.

## B. Additional Ablations

**Negative Transfer and Query Sensitivity.** A critical concern in contrastive decoding is whether the "negative" context accidentally remains consistent with the query, leading to the suppression of correct answers. We analyze this trade-off using our geometric noise ablation on the 3D-POPE benchmark in Table 3. We observe that minimal distortions ($\varepsilon = 0.01$) indeed fail to provide a sufficient contrastive signal, resulting in higher Yes-rates (86.05%) and lower F1 scores (73.81%) compared to moderate distortions. This supports the view that if the negative scene is too similar to the original, the penalty is ineffective. However, performance peaks at moderate noise levels ($\varepsilon = 0.05$), achieving the highest F1 (75.00%) and Accuracy (67.65%) on the *random* split. Performance degrades again at extreme noise levels ($\varepsilon = 0.45$), confirming that structural coherence is necessary for a meaningful contrastive signal.

**Sensitivity to Geometric Perturbations.** We perform an ablation study to examine how 3D-VCD responds to varying levels of geometric noise injected into the 3D scene graph. Specifically, we distort each object's centroid and extent using zero-mean Gaussian perturbations with increasing standard deviations $\varepsilon \in \{0.01, 0.05, 0.15, 0.45\}$. The goal is to determine whether 3D-VCD relies on exact spatial coordinates or whether its contrastive mechanism is robust to structural uncertainty.

In Table 3, and across most 3D-POPE splits, moderate distortions yield the highest gains: $\varepsilon = 0.05$ for instance, achieves F1 = 75.00 and Accuracy = 67.65 on the *random* split, while reducing Yes-Rate from roughly $86\%$ at $\varepsilon = 0.01$ and $\varepsilon = 0.45$ down to $78.41\%$. A slightly larger perturbation level ($\varepsilon = 0.15$) produces nearly identical performance, suggesting a performance plateau at moderate distortion levels.Conversely, the smallest and largest distortions ($\varepsilon = 0.01$ and $\varepsilon = 0.45$) degrade performance relative to the mid-range settings. Too little noise appears insufficient to decorrelate hallucination-inducing features, while excessive noise corrupts spatial information necessary for grounding. This pattern mirrors observations in the original 2D VCD literature [21], where blurring an image enough to remove high-frequency appearance features, but not erasing structural content, produces the strongest contrastive signal, whereas both minimal perturbation and extreme degradation lead to diminished gains. Overall, these findings suggest that our method performs best under moderate geometric perturbations that disrupt over-specific coordinate cues while maintaining the high-level spatial layout of the scene. This supports the view that 3D-VCD relies on coarse structural grounding rather than fine-grained 3D precision, making it robust to noisy or imperfect scene representations.

**Robustness to Perturbation Strength.** Table 4 presents detailed results across 3D-POPE splits under varying perturbation strengths, combining both semantic and geometric distortions. We evaluate two $\alpha$ configurations: $\alpha = 1.0$ (strong contrastive signal), and $\alpha = 0.5$ (moderate), Notably, the Random split yields the strongest performance, achieving an F1 of 74.48% under *Low-SemSub-Geom* with $\alpha = 1.0$. This shows that 3D-VCD's robustness extends to unseen, diverse prompts. Overall, increasing $\alpha$ (contrastive strength) produces small but consistent improvements in F1, demonstrating that stronger contrastive supervision amplifies grounding consistency without sacrificing generalization. These results collectively validate that combining semantic and geometric distortions provides the most balanced and stable form of regularization for visual contrastive decoding in 3D embodied environments.

## C. Results on the HEAL Probing Set

We evaluate the efficacy of our proposed 3D-VCD method on the HEAL benchmark, specifically focusing on adversarial probes designed to induce hallucinations through scene-task inconsistencies (Table 5). We apply our method by contrasting the adversarial prompts (*e.g.*, Distractor Injection) against

Table 3. **Effect of Geometric Distortion Strength** ($\varepsilon$) on 3D-POPE Performance. Lower yes-rate indicates reduced hallucination.

| 3D-POPE | Geometric Distortions ($\varepsilon$) | Precision ↑ | Recall ↑ | F1 ↑ | Accuracy ↑ | Yes % ↓ |
|---|---|---|---|---|---|---|
| *Random* | 3D-VCD $\varepsilon$=0.01 | 59.20 | 98.02 | 73.81 | 63.86 | 86.05 |
| | 3D-VCD $\varepsilon$=0.05 | 61.89 | 95.17 | 75.00 | 67.65 | 78.41 |
| | 3D-VCD $\varepsilon$=0.15 | 61.68 | 95.28 | 74.88 | 67.39 | 78.82 |
| | 3D-VCD $\varepsilon$=0.45 | 59.21 | 98.25 | 73.89 | 63.79 | 86.55 |
| *Popular* | 3D-VCD $\varepsilon$=0.01 | 51.54 | 97.66 | 67.47 | 52.45 | 95.69 |
| | 3D-VCD $\varepsilon$=0.05 | 52.35 | 94.09 | 67.27 | 53.34 | 91.61 |
| | 3D-VCD $\varepsilon$=0.15 | 52.48 | 94.44 | 67.47 | 53.46 | 91.96 |
| | 3D-VCD $\varepsilon$=0.45 | 51.67 | 97.55 | 67.56 | 52.63 | 95.46 |
| *Adversarial* | 3D-VCD $\varepsilon$=0.01 | 50.37 | 97.26 | 66.37 | 51.32 | 95.35 |
| | 3D-VCD $\varepsilon$=0.05 | 51.50 | 94.39 | 66.64 | 53.30 | 90.57 |
| | 3D-VCD $\varepsilon$=0.15 | 51.60 | 94.04 | 66.64 | 53.28 | 90.42 |
| | 3D-VCD $\varepsilon$=0.45 | 50.43 | 97.25 | 66.42 | 51.39 | 95.33 |

Table 4. **Ablation on Semantic and Geometric Distortions** under varying contrastive strengths $\alpha$ on 3D-POPE.

| 3D-POPE | Method | Precision ↑ | Recall ↑ | F1 ↑ | Accuracy ↑ | Yes % ↓ |
|---|---|---|---|---|---|---|
| *Random* | Low-SemSub-Geom ($\alpha$=1.0) | **62.16** | 92.90 | **74.48** | **67.99** | **75.15** |
| | Low-SemSub-Geom ($\alpha$=0.5) | 59.19 | 96.22 | 73.29 | 64.39 | 82.56 |
| | High-SemSub-Geom ($\alpha$=1.0) | 59.52 | 95.65 | 73.38 | 64.58 | 81.99 |
| | High-SemSub-Geom ($\alpha$=0.5) | 59.42 | 95.95 | 73.39 | 64.56 | 82.26 |
| *Popular* | Low-SemSub-Geom ($\alpha$=1.0) | **52.35** | 92.86 | 66.95 | **54.00** | 89.02 |
| | Low-SemSub-Geom ($\alpha$=0.5) | 51.39 | 96.23 | 66.99 | 52.72 | 93.39 |
| | High-SemSub-Geom ($\alpha$=1.0) | 51.84 | 95.65 | **67.24** | 53.42 | 92.22 |
| | High-SemSub-Geom ($\alpha$=0.5) | 51.73 | 95.84 | 67.19 | 53.23 | 92.58 |
| *Adversarial* | Low-SemSub-Geom ($\alpha$=1.0) | **52.90** | 92.59 | 67.33 | **54.92** | **87.82** |
| | Low-SemSub-Geom ($\alpha$=0.5) | 51.65 | 96.17 | 67.21 | 53.08 | 93.08 |
| | High-SemSub-Geom ($\alpha$=1.0) | 52.22 | 95.76 | **67.58** | 53.89 | 92.04 |
| | High-SemSub-Geom ($\alpha$=0.5) | 52.03 | 95.91 | 67.46 | 53.71 | 92.24 |

their corresponding clean baseline prompts. This setup tests whether 3D-VCD can recover the agent's original grounding capabilities even when presented with misleading textual cues.

Our inference-time intervention significantly enhances the grounding capabilities of the Qwen-14B-Instruct model across all challenging splits. Under the *Distractor Injection* setting, which tests resilience to irrelevant text mentions, 3D-VCD reduces the State Hallucination Rate ($C_S$) by approximately 70% (from 16.5% to 5.0%), effectively steering the generation away from ungrounded textual cues and towards the physical scene reality. Furthermore, in the *Scene-Object Synonymous* probe, 3D-VCD achieves the lowest Object Hallucination Rate ($C_O$) of 1.0% among all evaluated models, outperforming even the robust Llama-3-8B baseline. This indicates that our contrastive decoding strategy reinforces semantic consistency, ensuring the agent grounds its reasoning in the object's physical presence rather than its lexical token. Most notably, under *Scene-Task Contradiction*, where the base model hallucinates objects 53.9% of the time to satisfy impossible goals, 3D-VCD drastically mitigates this behavior ($C_O$ 1.5%), demonstrating a strong refusal to fabricate non-existent items.

## D. Additional Qualitative Examples

Across both 3D-POPE and HEAL, our qualitative examples highlight the central deficiencies of current 3D-aware language models. Baseline models often fail at the most fundamental level by producing an incorrect "yes" or "no" response, as shown in Figures 7 and 8. These examples indicate that the model relies heavily on linguistic priors rather than true 3D evidence, particularly in cluttered scenes or when relational reasoning is involved. In the HEAL example (Figure 7), the baseline 3D-LLM hallucinates the presence of a bed and incorrectly answers "yes," despite the scene information indicating otherwise. By contrast, 3D-VCD correctly answers by contrasting logits under the original and perturbed 3D scene graphs, suppressing object activations that are not grounded in the scene. In the 3D-POPE example (Figure 8), the baseline

Table 5. **Comparative Hallucination Rates (CHAIR) on HEAL Probes.** We report Object Hallucination ($C_O$) and State Hallucination ($C_S$) rates. Our proposed 3D-VCD method (applied to Qwen-14B) shows reduced hallucination, particularly under Distractor Injection.

| Env | Models | Base Prompt | | TaskDescMod | | | | SceneMod | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | DistInj | | TaskObjRem | | SynonymSub | | SceneTaskCon | |
| | | $C_O \downarrow$ | $C_S \downarrow$ | $C_O \downarrow$ | $C_S \downarrow$ | $C_O \downarrow$ | $C_S \downarrow$ | $C_O \downarrow$ | $C_S \downarrow$ | $C_O \downarrow$ | $C_S \downarrow$ |
| EHAVIOR | Gemma-2-9b-it | 0.87 | 1.80 | 11.60 | 5.70 | 10.8 | 4.8 | 11.4 | 2.8 | 73.1 | 8.2 |
| | DS-R1-Distil-LLaMA-8B | 1.26 | 8.16 | 13.90 | 27.20 | 6.8 | 13.3 | 20.4 | 10.1 | 60.0 | 19.6 |
| | Llama-3-8B-Instruct | 1.93 | 2.15 | 2.58 | 9.49 | 5.0 | 6.7 | 2.2 | 7.4 | 17.8 | 7.7 |
| | Qwen-14B-Instruct | 1.1 | 11.0 | 4.1 | 16.5 | 3.5 | 10.0 | 3.9 | 10.0 | 53.9 | 5.7 |
| | Qwen-14B-Instruct-VCD | – | – | **3.6** | **5.00** | 4.7 | **7.2** | **1.0** | **5.7** | **1.5** | 7.2 |

similarly hallucinates a desk and falsely predicts its presence, whereas 3D-VCD correctly determines that no desk exists. Together, these qualitative findings demonstrate that 3D-VCD strengthens the model's reliance on authentic visual evidence, yielding more faithful and reliable predictions across diverse 3D reasoning tasks.

## E. Broader Impact

This work aims to improve the reliability of 3D embodied agents by reducing hallucinations at inference time. More trustworthy grounding can benefit applications such as household robotics, assistive systems, and embodied decision-making, where outputs must remain consistent with the observed environment. Because 3D-VCD is training-free and can be applied without retraining, it also offers a practical way to improve existing systems without the cost of collecting new data or updating model parameters. At the same time, reducing hallucinations does not fully solve reliability in embodied AI. Grounding errors are only one source of failure, and safer perception does not guarantee safe downstream planning or action under noise, distribution shift, or adversarial conditions. As a result, methods such as 3D-VCD should be deployed together with additional safeguards, including uncertainty estimation, human oversight, and rigorous real-world testing.

**Question:** *"Are there any beds in the room?"*
**Ground truth: No**

**3D-LLM (Baseline)** ✗ Incorrect
**Answer: Yes**, There is a bed in the room.

**3D-VCD (Ours)** ✓ Correct
**Answer: No**, There are no beds in the room.

**Scene Objects:**
Lamp, coffee table, guitar, plant, TV stand, bean bag, pillow, TV, couch, picture, recliner chair

Figure 7. **Qualitative comparison on HEAL.** The baseline 3D-LLM hallucinates the presence of a bed. In contrast, 3D-VCD correctly answers *No* by contrasting logits under perturbed 3D scene graphs, effectively suppressing hallucinated object activations. The right panel shows the object-level scene segmentation for reference.

**Question:** *"Is there a desk in the room?"*
**Ground truth: No**

**3D-LLM (Baseline)** ✗ Incorrect
**Answer: Yes**, There is a desk in the room.

**3D-VCD (Ours)** ✓ Correct
**Answer: No**, there is no desks in the room.

**Scene Objects:**
Chair, floor carpet, wall, refrigerator, wall panel, table, tissue box, shelf, closet wall, kitchen cabinet

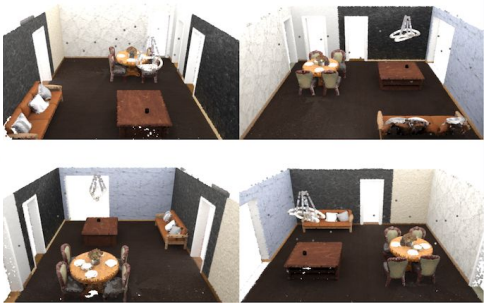

Figure 8. **Qualitative comparison on 3D-POPE.** The baseline 3D-LLM hallucinates a desk object and incorrectly predicts its presence. In contrast, 3D-VCD correctly determines that no desk exists by suppressing spurious category matches through contrastive decoding aligned with the object-centric scene graph.

```
Model Prompt: [INST] <<SYS>>
You are a helpful language and vision assistant that helps human reason about a 3D scene. You will be provided with a 3D scene context that contains a set of objects. You are able to understand the 3D scene, and answer the user's questions truthfully with respect to the 3D scene. In the 3D scene context, each object is represented by a unique identifier, such as <obj_0>, <obj_1>, etc. In all of your responses, you should explicitly ground each object noun phrase to the corresponding object identifier in the 3D scene.For example, when the user asks 'Describe this room.', you should respond with this format: 'This is a living room with a <p>brown sofa</p>[<obj_2>] and <p>three chairs</p>[<obj_0>, <obj_1>, <obj_3>].'Note that your answer should always enclose object noun phrases with <p> and </p> tags, followed by the corresponding object identifiers in square brackets.
<</SYS>>

 %%%% Object-centric context: <obj_0>: {'category': 'trash can', 'centroid': '[-0.25, -2.57, 0.12]', 'extent': '[0.43, 0.27, 0.38]'}; <obj_1>: {'category': 'pillow', 'centroid': '[-0.90, -1.77, 0.47]', 'extent': '[0.47, 0.42, 0.22]'}; <obj_2>: {'category': 'chair', 'centroid': '[-0.29, 1.94, 0.36]', 'extent': '[0.72, 0.70, 0.83]'}; <obj_3>: {'category': 'chair', 'centroid': '[0.29, 1.32, 0.32]', 'extent': '[0.59, 0.59, 0.69]'}; <obj_4>: {'category': 'couch', 'centroid': '[-0.98, -1.17, 0.41]', 'extent': '[1.03, 2.42, 0.94]'}; <obj_5>: {'category': 'ceiling', 'centroid': '[1.32, 0.61, 2.67]', 'extent': '[0.76, 0.77, 0.12]'}; <obj_6>: {'category': 'shelf', 'centroid': '[-1.35, -0.77, 1.77]', 'extent': '[0.34, 1.28, 0.70]'}; <obj_7>: {'category': 'pillow', 'centroid': '[-0.92, -0.91, 0.44]', 'extent': '[0.48, 0.54, 0.21]'}; <obj_8>: {'category': 'doorframe', 'centroid': '[0.04, -2.76, 0.98]', 'extent': '[0.66, 0.20, 2.07]'}; <obj_9>: {'category': 'shelf', 'centroid': '[1.33, 1.50, 1.09]', 'extent': '[0.60, 0.97, 0.65]'}; <obj_10>: {'category': 'mini fridge', 'centroid': '[0.36, 0.10, 0.47]', 'extent': '[0.46, 0.52, 0.90]'}; <obj_11>: {'category': 'ceiling', 'centroid': '[0.61, -2.29, 2.50]', 'extent': '[1.25, 1.28, 0.13]'}; <obj_12>: {'category': 'shelf', 'centroid': '[-1.32, 1.55, 1.05]', 'extent': '[0.47, 0.92, 0.67]'}; <obj_13>: {'category': 'shoe', 'centroid': '[0.18, 0.65, -0.02]', 'extent': '[0.35, 0.37, 0.12]'}; <obj_14>: {'category': 'shoe', 'centroid': '[-0.66, 0.09, -0.03]', 'extent': '[0.24, 0.35, 0.10]'}; <obj_15>: {'category': 'bookshelf', 'centroid': '[-0.83, -2.61, 0.55]', 'extent': '[1.03, 0.34, 1.25]'}; <obj_16>: {'category': 'fan', 'centroid': '[0.27, 2.52, 1.36]', 'extent': '[0.55, 0.24, 0.42]'}; <obj_17>: {'category': 'clothes', 'centroid': '[0.62, -1.48, 1.29]', 'extent': '[0.70, 1.64, 1.26]'}; <obj_18>: {'category': 'desk', 'centroid': '[-0.83, 1.11, 0.60]', 'extent': '[1.11, 2.20, 1.03]'}; <obj_19>: {'category': 'backpack', 'centroid': '[-0.71, 1.25, 0.13]', 'extent': '[0.40, 0.46, 0.34]'}; <obj_20>: {'category': 'chair', 'centroid': '[-0.08, -2.04, 0.10]', 'extent': '[0.51, 0.55, 0.47]'}; <obj_21>: {'category': 'book', 'centroid': '[-0.58, -2.63, 1.02]', 'extent': '[0.31, 0.21, 0.17]'}; <obj_22>: {'category': 'cabinet', 'centroid': '[1.05, -1.00, 0.40]', 'extent': '[0.89, 2.24, 0.99]'}; <obj_23>: {'category': 'desk', 'centroid': '[0.89, 1.30, 0.59]', 'extent': '[1.18, 1.99, 0.84]'}; <obj_24>: {'category': 'blanket', 'centroid': '[-0.83, 1.37, 1.60]', 'extent': '[0.90, 2.15, 1.08]'}; <obj_25>: {'category': 'shoe', 'centroid': '[0.38, 0.48, -0.01]', 'extent': '[0.40, 0.43, 0.13]'}; <obj_26>: {'category': 'window', 'centroid': '[0.12, 2.44, 1.72]', 'extent': '[2.64, 0.46, 1.35]'}; <obj_27>: {'category': 'bed', 'centroid': '[-0.81, 0.85, 1.25]', 'extent': '[1.20, 2.43, 1.84]'}; <obj_28>: {'category': 'clothes', 'centroid': '[-0.16, -1.74, 0.28]', 'extent': '[0.66, 0.57, 0.82]'}; <obj_29>: {'category': 'trash can', 'centroid': '[0.53, 2.20, 0.24]', 'extent': '[0.46, 0.48, 0.27]'}; <obj_30>: {'category': 'bottle', 'centroid': '[-0.66, -2.56, 0.48]', 'extent': '[0.10, 0.14, 0.22]'};

Ground these sentences: <refer_expression>Are there any lamps in the room?<refer_expression> [/INST]
```

```
Model Response:
"raw_output": <detailed_grounding>Yes, there is a <p>lamp</p>[<obj_16>] in the room. It is located on the <p>ceiling</p>[<obj_11> </detailed_grounding>
```

Figure 9. **3D-VCD model input and output.**