WHEN DEBIASING RELOCATES SHORTCUTS: ORTHOGONAL INTERACTION FINE-TUNING FOR 3D LARGE MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite continued progress in 3D scene understanding, 3D large multimodal models (3D-LMMs) remain susceptible to language priors: predictions that should depend on the current scene can still be shaped by question patterns, co-occurrence statistics, and answer preferences. Existing prior-aware methods seek to suppress such shortcuts, but our analysis shows that gains on targeted debiasing objectives do not necessarily imply stronger scene dependence. Correction-induced changes can persist after explicit scene evidence is removed while systematically reshaping the answer policy. We term this phenomenon *Shortcut Relocation*: suppressing one shortcut can shift adaptation toward another shortcut-compatible policy rather than stronger scene-specific dependence. This finding motivates *Corrective Function Allocation*, a view of debiasing that explicitly restricts the functional components available to newly learned correction. We instantiate this principle with *Orthogonal Interaction Fine-Tuning* (OIF). OIF uses language-prior diagnostics only for risk localization, and restricts newly trainable corrective capacity to a centered bilinear interaction between linguistic context and the 3D scene; a policy-preservation objective limits unnecessary changes elsewhere. Under the specified reference measure, the added residual contains no first-order unimodal main effects. Matched-control experiments on Real-3DQA and its viewpoint-rotation variants, together with 3D-POPE, show that OIF reduces answer-policy drift while better preserving cross-benchmark robustness. These results show that prior-aware debiasing depends not only on identifying shortcut-prone predictions, but also on constraining where newly introduced corrective capacity can reside in function space. Code is included in the supplementary material.

1 INTRODUCTION

3D large multimodal models (3D-LMMs) are emerging as a general framework for unified 3D scene understanding, coupling 3D scene representations with large language models to support question answering, spatial reasoning, referring expression grounding, and scene description (Hong et al., 2023; Chen et al., 2024; Huang et al., 2024a;b; Deng et al., 2025). However, their predictions depend not only on the current 3D scene, but also on regularities inherited from language modeling and dataset construction, including question patterns, answer frequencies, scene knowledge, and object co-occurrences. Such regularities can remain predictive even when explicit 3D scene input is withheld (Goyal et al., 2017; Agrawal et al., 2018; Ramakrishnan et al., 2018; Ma et al., 2026). We refer to such scene-independent predictive regularities as *language priors*. The key concern is therefore not merely whether a prediction is correct, but whether it is actually grounded in the current 3D scene. Conventional evaluations can obscure this distinction, overestimating scene-understanding ability when predictions are supported by language shortcuts rather than scene-specific evidence. This mismatch becomes particularly apparent when scene conditions or viewpoints change, where shortcut-driven predictions are more likely to exhibit inconsistent spatial reasoning and scene hallucination (Teney et al., 2020; Ma et al., 2026; Yang et al., 2025).

Prior-aware methods attempt to close this gap by diagnosing predictions that remain accessible without explicit scene evidence. A common strategy uses a *blind model* that retains the linguistic context while withholding the 3D scene, providing a practical signal for identifying shortcut-prone exam-

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.

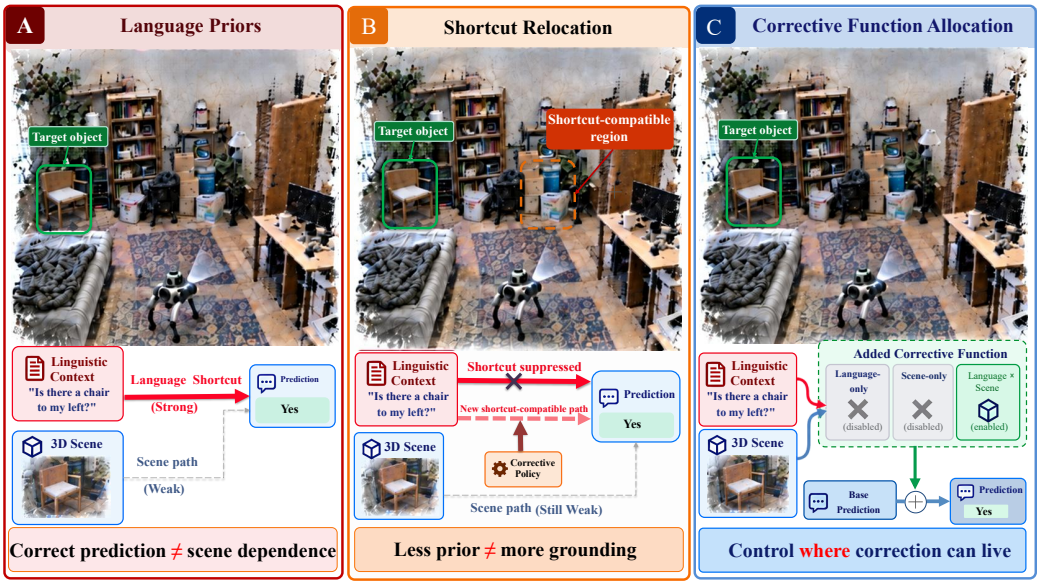


Figure 1: **From Shortcut Relocation to Corrective Function Allocation.** Prior-aware correction can suppress an existing language shortcut while the learned correction shifts toward another shortcut-compatible policy; OIF restricts newly introduced corrective capacity to the interaction between linguistic context and the current 3D scene.

ples (Ma et al., 2026). Training-time methods use such signals to adjust supervision or optimization (Ramakrishnan et al., 2018; Cadene et al., 2019; Clark et al., 2019; Ma et al., 2026), whereas inference-time methods use contrastive or calibration-based mechanisms to reduce reliance on predictions that remain strong without sufficient scene support (Niu et al., 2021; Leng et al., 2024; Ogunleye et al., 2026). Despite their differences, these approaches share the objective of weakening language shortcuts and increasing the influence of the current 3D scene on the final prediction. Crucially, however, prior diagnosis identifies *where* shortcut risk occurs; it does not determine *how* the model will realize the resulting correction. If alternative shortcut-compatible functions remain available, suppressing one predictive route need not force the model toward stronger scene dependence (Geirhos et al., 2020; Si et al., 2022; Wen et al., 2022).

Our analysis reveals precisely this failure mode. Under matched full-input and blind evaluations, correction-induced prediction changes often persist after explicit scene evidence is removed, while the marginal answer distribution and affirmative or directional tendencies shift systematically. These changes can yield local gains on targeted question types without improving overall viewpoint consistency, and on 3D-POPE can collapse existence judgments toward an almost constant affirmative policy. Together, these observations show that adaptation can satisfy a corrective objective by reorganizing the decision policy rather than increasing dependence on the current scene. We term this phenomenon *Shortcut Relocation*: suppressing one shortcut allows the corrective objective to be absorbed by another shortcut-compatible decision path instead of forcing stronger scene-specific dependence.

Shortcut Relocation exposes an underdetermination in prior-aware correction: the same corrective signal can be realized by qualitatively different input-dependent functions. A learned correction may depend primarily on linguistic context, exploit generic scene statistics, or arise from their joint dependence; all can reduce the training objective while generalizing differently. We therefore formulate *Corrective Function Allocation*, which constrains which functional components are allowed to carry newly introduced corrective capacity. For correction specific to the current language–scene pair, joint dependence on both inputs provides a minimum structural requirement, whereas explicit unimodal correction paths leave direct routes for Shortcut Relocation. Figure 1 illustrates this transition from suppressing a diagnosed shortcut to controlling where corrective capacity can reside.

Following this principle, we propose *Orthogonal Interaction Fine-Tuning* (OIF), a structured realization of Corrective Function Allocation. OIF freezes the base 3D-LMM and learns only an incremental corrective residual. A blind diagnostic localizes high-risk examples but does not specify an answer-side corrective target; newly trainable capacity is restricted to a centered bilinear interaction

between linguistic context and the explicit 3D scene, while a policy-preservation objective limits unnecessary changes elsewhere. Under the specified product reference measure, the added residual has zero first-order linguistic-only and scene-only main effects. OIF therefore constrains not merely where correction is applied, but the function class through which it can affect predictions.

We evaluate OIF along three complementary axes: *debiased question answering*, *viewpoint consistency*, and *object hallucination*, using Real-3DQA and its viewpoint-rotation variants together with 3D-POPE (Ma et al., 2026; Yang et al., 2025). Matched controls show that OIF reduces relocation signatures and answer-policy drift while maintaining a stronger cross-benchmark robustness profile than more permissive corrective alternatives. Restricting explicit correction paths can therefore mitigate Shortcut Relocation, while interaction structure alone does not ensure reliable identification of 3D evidence. These results support our central argument: generalization in prior-aware debiasing depends not only on the corrective signal, but also on the function space available for learning the correction.

Our main contributions are as follows:

- We identify *Shortcut Relocation* in prior-aware debiasing of 3D-LMMs, showing that prior-aware correction can reorganize shortcut-compatible behavior without a corresponding increase in scene dependence.
- We formulate *Corrective Function Allocation*, reframing prior-aware debiasing as controlling which input-dependent function components may carry newly learned corrective capacity.
- We propose *OIF*, which realizes this principle through risk localization, centered language–scene interaction, and base-policy preservation. Theoretical analysis and matched-control evaluations show that this structural restriction better preserves cross-benchmark robustness than more permissive corrective alternatives.

2 Related Work

3D scene-language research has evolved from task-specific question answering (Azuma et al., 2022; Ma et al., 2023) to pretrained and generalist 3D multimodal models (Hong et al., 2023; Chen et al., 2024; Huang et al., 2024b). Recent work has begun to examine whether these models actually rely on current 3D evidence. Real-3DQA exposes language-prior dependence through shortcut-filtered questions and viewpoint-consistency evaluation, while 3DR-FT uses a blind model to reweight training supervision (Ma et al., 2026). Related 2D work mitigates language bias through biased auxiliary branches, counterfactual training, or inference-time subtraction and calibration (Ramakrishnan et al., 2018; Cadene et al., 2019; Clark et al., 2019; Chen et al., 2020; Niu et al., 2021), and broader shortcut-learning studies show that suppressing one predictive cue need not force reliance on the intended evidence (Geirhos et al., 2020; Teney et al., 2020; Si et al., 2022; Wen et al., 2022). In 3D and multimodal inference, recent approaches contrast clean and perturbed visual inputs (Leng et al., 2024; Ogunleye et al., 2026) or intervene in decoder states and attention (Ji et al., 2026; Jo et al., 2026). Our work addresses a complementary question: rather than only diagnosing or suppressing a shortcut, we study how the model is allowed to realize the resulting correction. We formalize this as *Corrective Function Allocation* and instantiate it with OIF, which localizes prior risk while restricting newly learned corrective capacity to centered language–scene interactions.

3 When prior correction relocates shortcuts

A prior diagnostic can reveal predictions that remain accessible without explicit scene evidence, but it does not determine how the model will change once those predictions are corrected. We show that this unresolved freedom can relocate rather than remove shortcuts, motivating a structural constraint on the corrective function itself.

3.1 Outcome supervision underdetermines evidence use

Let L denote the linguistic context, including the situation description and question in SQA3D, X the explicit 3D scene representation, and Y the target answer sequence. Standard supervised

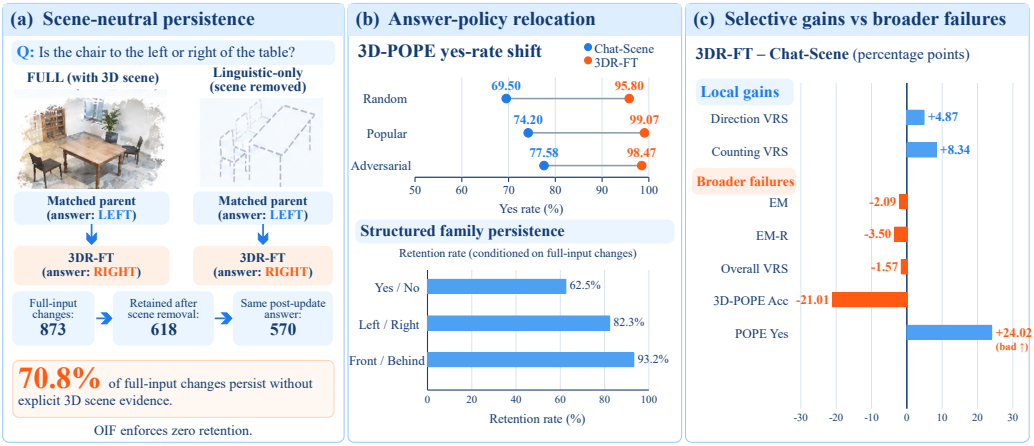


Figure 2: **Empirical signature of Shortcut Relocation.** (a) Full-input prediction changes induced by 3DR-FT that persist after explicit 3D evidence is removed. (b) Systematic shifts in answer policy, including affirmative and directional behavior. (c) Local gains on targeted question types alongside broader degradation in debiased accuracy, viewpoint consistency, and 3D-POPE calibration.

fine-tuning minimizes $\mathcal{L}_{\mathrm{SFT}} = -\log p_\theta(Y \mid L, X)$, which specifies the desired answer on each observed pair (L, X) but not the evidence on which that answer should depend. The same loss can be reduced through scene-specific evidence, linguistic regularities, or answer-distribution biases, so similar target accuracy can conceal different dependencies on the inputs.

A blind model, which withholds X, measures how predictable an answer remains from L alone and provides a diagnostic for prior-aware training (Ma et al., 2026). This diagnosis identifies shortcut-prone examples but does not constrain the form of the resulting correction: prior-aware adaptation may strengthen scene-conditioned reasoning or reshape a language-conditioned answer policy, and which of the two occurs is an empirical question.

3.2 Shortcut Relocation

We examine this question through matched full-input and linguistic-only evaluations. For each Real-3DQA example, we compare 3DR-FT with its parent checkpoint once using (L, X) and once after removing the explicit scene while keeping L fixed. Of the 873 prediction changes induced under full input, 618 (70.79%) persist without explicit scene evidence, and 570 reach the same post-update answer in both conditions.

This persistence is accompanied by systematic answer-marginal shifts and a mixed robustness profile: gains on direction and counting questions coexist with lower reduced-prior accuracy and a collapse of 3D-POPE existence judgments (Figure 2). Controlled comparisons appear in Section 5 and the complete paired analysis in Appendix D.5.

We call this failure mode *Shortcut Relocation*: corrective adaptation suppresses one shortcut but realizes its behavioral change through another shortcut-compatible policy rather than through stronger scene dependence. Because relocation concerns the change induced by adaptation rather than bias in the parent model alone, we diagnose it from converging evidence across scene-removal persistence, answer-policy shifts, and independent robustness tests.

3.3 Corrective Function Allocation

Let $\mathcal{C}_0$ be a pretrained 3D-LMM and let $\Delta_h(L, X) \in \mathbb{R}^{d_h}$ denote an incremental correction in its decoder hidden space. Under a product reference $Q = Q_L \otimes Q_X$, any square-integrable residual admits the two-input Hoeffding decomposition (Hoeffding, 1948; Efron & Stein, 1981)

$$\Delta_h(L, X) = \Delta_0 + \Delta_L(L) + \Delta_X(X) + \Delta_{LX}(L, X), \tag{1}$$

where Δ_0 is constant, Δ_L and Δ_X are centered unimodal main effects, and Δ_{LX} is the interaction component. The product reference defines this decomposition without assuming independence of L and X in the observed data (Hooker, 2007).

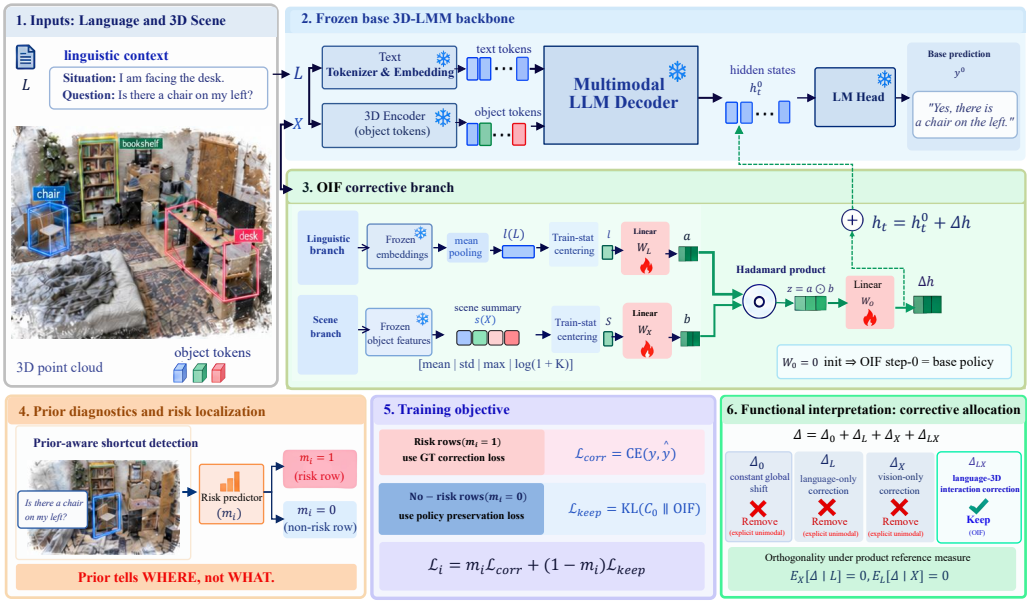


Figure 3: **Orthogonal Interaction Fine-Tuning.** OIF adds a centered bilinear language–scene residual to a frozen 3D-LMM. A blind diagnostic only localizes high-risk training examples: these receive ground-truth correction, while the remaining examples preserve the base policy. Only W_L, W_X, and W_O are trained.

The main-effect terms provide direct shortcut-compatible routes: Δ_L can reshape language-conditioned answer policies, while Δ_X can encode scene-level frequency effects independently of the current request. To remove these direct routes from the newly learned correction, we restrict the residual to the interaction subspace.

$$\mathcal{H}_{LX}(Q) = \{g \in L^2(Q; \mathbb{R}^{d_h}) : \mathbb{E}_Q[g \mid L] = 0,\ \mathbb{E}_Q[g \mid X] = 0\}. \tag{2}$$

Under Q, this subspace is orthogonal to the constant and unimodal main-effect components. The constraint removes newly trainable unimodal routes without claiming causal grounding. Section 4 constructs a parameterization that satisfies it by design; Appendix A gives the full decomposition and proof.

4 Orthogonal Interaction Fine-Tuning

Orthogonal Interaction Fine-Tuning (OIF) adapts a frozen 3D-LMM through a structurally constrained residual. It combines centered language–scene interaction with prior-aware risk localization and base-policy preservation, as illustrated in Figure 3.

4.1 Residual Interface

Let $h_t^0(L, X, y_{<t})$ and ℓ_t^0 denote the hidden state and logits produced by the frozen base model $\mathcal{C}_0$ for answer prefix $y_{<t}$. OIF adds a sample-conditioned residual immediately before the frozen language-model head:

$$h_t^{\mathrm{OIF}} = h_t^0 + \Delta_h(L, X), \qquad \ell_t^{\mathrm{OIF}} = \ell_t^0 + W_{\mathrm{LM}}\Delta_h(L, X). \tag{3}$$

The base encoder, 3D interface, decoder, and language-model head remain frozen. The same $\Delta_h(L, X)$ is broadcast across answer positions, while the base hidden states retain their token-dependent computation.

4.2 Centered Language–Scene Interaction

Both inputs to the corrective branch are derived from frozen base representations. We summarize the linguistic context by mean-pooling its frozen token embeddings, denoted $l(L)$, and summarize

the valid frozen object features by their coordinatewise mean, standard deviation, maximum, and object count, denoted $s(X)$. Exact constructions are given in Appendix B.

To prevent a multiplicative interaction from inducing first-order main effects through nonzero marginal means, we standardize both summaries with statistics fixed from the SQA3D training split:

$$\widetilde{l}(L) = \frac{l(L) - \mu_L}{\sigma_L + \epsilon}, \qquad \widetilde{s}(X) = \frac{s(X) - \mu_X}{\sigma_X + \epsilon}. \tag{4}$$

Their row-weighted empirical marginals define the product reference $\widehat{Q}_{\mathrm{ref}} = \widehat{P}_L^{\mathrm{row}} \otimes \widehat{P}_X^{\mathrm{row}}$, which fixes the functional decomposition; optimization still uses the observed language–scene pairs.

OIF restricts the newly trainable correction to a bias-free low-rank bilinear interaction:

$$\Delta_h(L, X) = W_O \left[\left(W_L \widetilde{l}(L) \right) \odot \left(W_X \widetilde{s}(X) \right) \right]. \tag{5}$$

No additional nonlinear or token-adaptive module is introduced. Only $\Theta_{\mathrm{OIF}} = \{W_L, W_X, W_O\}$ is optimized. We initialize $W_O = 0$, so OIF exactly reproduces the frozen base model before the first update. Architectural dimensions and initialization details are reported in Appendix B.

4.3 Risk-localized correction and policy preservation

Before training, a frozen blind model assigns each example a fixed binary risk mask $m_i \in \{0, 1\}$. Risk examples are trained toward the ground-truth answer, whereas the remaining examples preserve the frozen base policy through forward distillation (Hinton et al., 2015):

$$\mathcal{L}^{(i)} = m_i \, \mathcal{L}_{\mathrm{CE}}^{(i)} + (1 - m_i) \, \mathcal{L}_{\mathrm{KL}}^{(i)}, \qquad \mathcal{L}_{\mathrm{KL}}^{(i)} = D_{\mathrm{KL}} \left(\mathrm{sg}[p_{\mathcal{C}_0}^{(i)}] \,\|\, p_{\mathrm{OIF}}^{(i)} \right). \tag{6}$$

Both terms are averaged over answer-token positions under the same teacher-forced prefix. The exact risk cohort and token-level objective are specified in Appendix B. At inference, neither the blind model nor the training-time mask is required.

4.4 Orthogonality of the corrective function

The parameterization in Equation 5 satisfies the Corrective Function Allocation constraint in Equation 2 by construction. Under the empirical product reference $\widehat{Q}_{\mathrm{ref}}$, the centered, bias-free residual satisfies

$$\mathbb{E}_{\widehat{Q}_{\mathrm{ref}}}[\Delta_h \mid L] = 0, \qquad \mathbb{E}_{\widehat{Q}_{\mathrm{ref}}}[\Delta_h \mid X] = 0, \tag{7}$$

so $\Delta_h \in \mathcal{H}_{LX}(\widehat{Q}_{\mathrm{ref}})$ contains no first-order linguistic-only or scene-only main effect under this reference. The identities hold because the centered projected marginals have zero expectation and the fixed linear maps W_O and W_{LM} preserve this property, so the logit residual satisfies the same constraint throughout optimization. The guarantee concerns the added residual under the declared reference measure and does not by itself establish causal grounding. Appendix A states the general theorem and its proof.

5 Experiments

5.1 Experimental setup

Training and evaluation. Chat-Scene (Huang et al., 2024a) is our primary backbone. OIF and all newly trained controls start from its released checkpoint and are trained for one epoch on the SQA3D training split (Ma et al., 2023); only the corrective branch is updated. 3DR-FT (Ma et al., 2026) follows its official recipe. Real-3DQA (Ma et al., 2026) evaluates debiased question answering and viewpoint consistency, and 3D-POPE (Yang et al., 2025) evaluates object-existence hallucination. Appendix C specifies the datasets, metrics, and pooling convention.

Comparisons and controls. We compare OIF with 3DR-FT; with three inference-time interventions adapted to Chat-Scene, namely 3D-VCD (Ogunleye et al., 2026), CausalLens (Ji et al., 2026), and ACG (Jo et al., 2026); with the structural controls MainEffect and Full-BM; and with matched random and All-rows training cohorts (Appendix C.2). OIF is additionally transferred to 3D-LLaVA (Deng et al., 2025).

Table 1: **Main results (%).** Real-3DQA reports EM, EM-R, VRS, and Dir-VRS; 3D-POPE reports pooled precision, F1, accuracy, and affirmative rate. Generalist rows are context; bold and underline mark the best and second-best results among the Chat-Scene reference and task-adapted methods.

Method	Real-3DQA				3D-POPE			
	EM ↑	EM-R ↑	VRS ↑	Dir-VRS ↑	P ↑	F1 ↑	Acc. ↑	Yes ↓
Generalist models								
Chat-Scene (Huang et al., 2024a)	17.91	22.49	<u>11.90</u>	19.69	64.92	<u>77.38</u>	72.01	73.76
3D-LLM (Hong et al., 2023)	7.50	10.40	9.60	24.60	–	–	–	–
LEO (Huang et al., 2024b)	14.30	19.10	14.30	23.20	–	–	–	–
Task-adapted mitigation methods								
3DR-FT (Ma et al., 2026)	15.82	18.99	10.33	**24.56**	50.51	66.84	51.00	97.78
3D-VCD (Ogunleye et al., 2026)	<u>18.65</u>	23.03	**13.00**	<u>21.68</u>	<u>65.09</u>	**77.41**	<u>72.14</u>	<u>73.37</u>
CausalLens (Ji et al., 2026)	18.59	<u>23.43</u>	11.53	19.47	63.56	76.78	70.68	76.28
ACG (Jo et al., 2026)	16.30	19.80	7.10	20.80	57.37	71.91	62.33	83.01
OIF	**19.60**	**24.44**	11.57	20.58	**65.42**	77.20	**72.19**	**71.97**

Table 2: **3D-POPE results by subset (%).** Random, Popular, and Adversarial contain 6,856 rows each; all methods use the Chat-Scene backbone. Yes ↓ is the affirmative rate and is read jointly with precision. Bold marks column-best and underline second-best.

Method	Random				Popular				Adversarial			
	P↑	F1↑	Acc.↑	Yes↓	P↑	F1↑	Acc.↑	Yes↓	P↑	F1↑	Acc.↑	Yes↓
Chat-Scene (Huang et al., 2024a)	68.90	80.14	76.27	69.50	<u>64.54</u>	**77.11**	<u>71.57</u>	74.20	61.72	**75.07**	<u>68.19</u>	77.58
3DR-FT (Ma et al., 2026)	51.55	67.75	52.98	95.80	49.85	66.26	49.71	99.07	50.16	66.53	50.31	98.47
3D-VCD (Ogunleye et al., 2026)	<u>69.42</u>	**80.40**	<u>76.72</u>	<u>68.79</u>	64.45	<u>76.96</u>	71.41	<u>74.10</u>	<u>61.83</u>	**75.07**	**68.28**	<u>77.23</u>
CausalLens (Ji et al., 2026)	67.25	79.42	74.87	72.10	63.15	76.48	70.19	76.78	60.62	74.60	66.99	79.97
ACG (Jo et al., 2026)	57.27	71.83	62.14	83.15	58.38	72.69	63.79	81.58	56.50	71.22	61.05	84.29
OIF	**69.81**	<u>80.22</u>	**76.75**	**67.53**	**64.98**	76.83	**71.66**	**72.30**	**61.94**	<u>74.76</u>	68.17	**76.08**

Mechanism measurements. Scene-neutralized retention, R_{retain}, measures how often a prediction change induced by adaptation under full input persists after scene neutralization, and Policy-TV measures marginal answer-policy drift relative to the parent model. Appendix C defines both quantities together with the intervention and bootstrap procedures.

5.2 Overall effectiveness

5.2.1 Debiased question answering and viewpoint consistency

Within the matched mitigation block, OIF attains the highest EM and EM-R in Table 1, together with the highest pooled 3D-POPE precision and accuracy and the lowest affirmative rate. Relative to the released checkpoint, it improves EM by 1.68 points and EM-R by 1.95 points (paired-bootstrap 95% CIs $[0.13, 3.30]$ and $[0.27, 3.64]$; Appendix D.2) and raises Dir-VRS from 19.69 to 20.58, while overall VRS moves from 11.90 to 11.57. The VRS change comes from distance questions, while existence and counting VRS rise.

The competing interventions are less balanced. 3DR-FT attains the highest Dir-VRS but lowers EM, EM-R, and overall VRS below the released checkpoint and answers affirmatively on 97.8% of 3D-POPE queries; 3D-VCD attains the highest overall VRS through a second decoding pass, with smaller gains than OIF in EM, EM-R, precision, and accuracy; CausalLens improves EM and EM-R but not Dir-VRS or existence calibration; and ACG raises Dir-VRS while worsening every other column. Among single-pass alternatives, OIF exhibits the most balanced profile: gains in EM, EM-R, Dir-VRS, precision, and accuracy accompany a lower affirmative rate, while overall VRS remains close to the released checkpoint.

5.2.2 Object hallucination

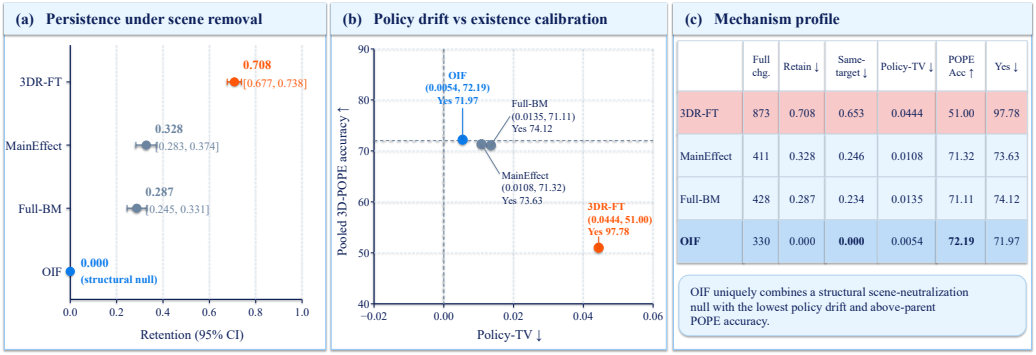
Table 2 shows that the pooled advantage in Table 1 holds within every subset: OIF attains the highest precision and the lowest affirmative rate on all three subsets, and the highest accuracy on Random and Popular. Against the released checkpoint, precision rises and the affirmative rate falls on every

subset. CausalLens and ACG raise the affirmative rate above the released checkpoint with lower precision and accuracy than OIF on every subset, and 3D-VCD approaches OIF's operating point only via a second decoding pass.

5.3 Behavioral validation of Corrective Function Allocation

Figure 4 tests whether restricting the corrective function changes the relocation signature established in Section 3. 3DR-FT retains 70.8% of its full-input prediction changes after scene neutralization; MainEffect and Full-BM retain 32.8% and 28.7%. OIF has zero scene-neutralized corrective changes, a structural null: neutralizing the centered scene summary makes the interaction residual vanish identically, which verifies the allocation constraint without implying causal grounding.

The same restriction also limits broader policy reorganization. Among learned corrective variants, OIF has the smallest Policy-TV (0.0054) while improving pooled 3D-POPE accuracy over its parent (72.19 vs. 72.01) and reducing the affirmative rate (71.97 vs. 73.76). MainEffect and Full-BM exhibit larger policy shifts and below-parent accuracy, while 3DR-FT combines the largest drift with near-constant affirmative prediction. Function allocation therefore changes the behavioral pathway of correction, not only task performance.



	Full chg.	Retain ↓	Same-target ↓	Policy-TV ↓	POPE Acc ↑	Yes ↓
3DR-FT	873	0.708	0.653	0.0444	51.00	97.78
MainEffect	411	0.328	0.246	0.0108	71.32	73.63
Full-BM	428	0.287	0.234	0.0135	71.11	74.12
OIF	330	0.000	**0.000**	0.0054	**72.19**	71.97

Figure 4: **Behavioral validation of Corrective Function Allocation.** (a) Retention of adaptation-induced changes after scene neutralization. (b) Answer-policy drift versus pooled 3D-POPE calibration. (c) Mechanism scorecard; OIF's zero retention is a structural null.

5.4 Effects of function allocation and risk localization

5.4.1 Corrective Function Allocation

The allocation controls share OIF's frozen base, summaries, risk mask, and objective (Appendix B). With $a = W_L\tilde{l}(L)$ and $b = W_X\tilde{s}(X)$, MainEffect (denoted ADD in the formulation) replaces the multiplicative interaction with centered unimodal main effects, and Full-BM adds both main effects to the interaction branch:

$$\Delta_h^{\text{ADD}} = W_O(a+b), \qquad \Delta_h^{\text{Full}} = U_I(a \odot b) + U_L a + U_X b. \tag{8}$$

Full-BM lowers the rank of the Full residual to match OIF's parameter budget (Appendix A.7).

Table 3 shows that, at equal parameter count, interaction-only allocation is the best arm on every reported 3D-POPE measure. OIF exceeds MainEffect on every pooled measure, with paired-bootstrap 95% CIs that exclude zero on pooled F1, accuracy, and affirmative rate (Appendix D.2), and Full-BM, which reopens both unimodal branches at the same budget, trails OIF on all ten measures, with pooled intervals again excluding zero, and shows a nonzero relocation signature in Figure 4. The result concerns where corrective capacity is allocated, not how much is added.

5.4.2 Prior-aware risk localization

The prior-aware selector exceeds the matched random-mask mean by 0.23 points in pooled F1 and 0.35 points in pooled accuracy, with an affirmative rate 0.34 points lower (paired 95% CIs $[0.03, 0.42]$, $[0.11, 0.60]$, and $[-0.59, -0.09]$), and it outperforms every individual mask on the pooled panel and on the Random and Popular subsets; on Adversarial, one mask is ahead by less

Table 3: **Function allocation and risk localization on 3D-POPE (%).** All arms stay within OIF's 5.77M parameter budget and start from the released checkpoint; Overall pools the three subsets. Allocation arms keep OIF's prior-aware cohort (19.39% of the training examples) and vary the subspace that may carry the correction; localization arms keep OIF's interaction branch and change the training cohort. Matched random reports the mean over three seeds at the same active fraction (Appendix D.1); All rows applies correction to every example. Bold marks the column-best score.

Method	Function subspace	Overall				Random		Popular		Adversarial	
		P ↑	F1 ↑	Acc. ↑	Yes ↓	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑
Function allocation											
MainEffect	$\mathcal{H}_L \oplus \mathcal{H}_X$	64.48	76.80	71.32	73.63	79.18	75.06	76.82	71.30	74.54	67.61
Full-BM	$\mathcal{H}_L \oplus \mathcal{H}_X \oplus \mathcal{H}_{LX}$	64.24	76.73	71.11	74.12	79.06	74.82	76.70	70.97	74.54	67.53
Risk localization											
Matched random	$\mathcal{H}_{LX}$	65.10	76.98	71.84	72.31	79.62	75.91	76.78	71.53	74.70	68.08
All rows	$\mathcal{H}_{LX}$	64.73	76.77	71.46	72.84	79.34	75.41	76.70	71.31	74.41	67.66
OIF	$\mathcal{H}_{LX}$	**65.42**	**77.20**	**72.19**	**71.97**	**80.22**	**76.75**	**76.83**	**71.66**	**74.76**	**68.17**

than the spread among the three masks (Appendix D.1). Applying correction to all rows reaches 76.77 pooled F1 and 71.46 accuracy, below the prior-aware result by margins whose paired intervals exclude zero. Matching supervision volume and coarse label composition is therefore insufficient; the prior diagnostic contributes sample-level localization information.

5.5 Cross-backbone generalization

Table 4 shows that on 3D-LLaVA, OIF improves EM by 2.69 points and EM-R by 2.09 points over the released checkpoint, raises overall VRS from 14.50 to 17.60 and Dir-VRS from 26.80 to 35.40, and increases pooled 3D-POPE accuracy while lowering the affirmative rate. These gains, larger than on Chat-Scene, show that the allocation principle is not tied to Chat-Scene's visual representation. OIF adds 5.77M trainable parameters on Chat-Scene and 10.49M on 3D-LLaVA, and retains a single decoding path; efficiency measurements are reported in Appendix D.6 and sample-level comparisons in Appendix E.

Table 4: **Cross-backbone results (%).** Released checkpoints are the reference; 3D-POPE metrics pool the three subsets.

Backbone	Method	Real-3DQA				3D-POPE		
		EM ↑	EM-R ↑	VRS ↑	Dir-VRS ↑	Acc. ↑	F1 ↑	Yes ↓
Chat-Scene (Huang et al., 2024a)	Released	17.91	22.49	**11.90**	19.69	72.01	**77.38**	73.76
	+OIF	**19.60**	**24.44**	11.57	**20.58**	**72.19**	77.20	**71.97**
3D-LLaVA (Deng et al., 2025)	Released	20.40	25.66	14.50	26.80	73.56	**78.17**	71.13
	+OIF	**23.10**	**27.74**	**17.60**	**35.40**	**74.01**	77.93	**67.79**

6 Conclusion

We identify *Shortcut Relocation*, a failure mode in which prior-aware debiasing suppresses an existing language shortcut while the learned correction is absorbed by another shortcut-compatible policy. This motivates *Corrective Function Allocation*: debiasing should constrain not only where correction is applied but also the function space through which corrective capacity can affect predictions. OIF instantiates this principle by keeping the base 3D-LMM fixed and restricting the added residual to a centered language–scene interaction, which removes first-order unimodal correction paths under the specified reference measure. Across Real-3DQA, its viewpoint-rotation variants, and 3D-POPE, on both Chat-Scene and 3D-LLaVA, this structural restriction preserves robustness during prior-aware adaptation. Requiring cross-modal interaction does not by itself establish causal grounding; OIF provides a controlled step toward preventing debiasing from introducing new shortcut-compatible corrections.

ETHICS STATEMENT

This work studies shortcut behavior and hallucination in 3D multimodal models on existing, publicly available indoor-scene benchmarks, which are used under their published licenses. It introduces no new human-subject data and no new annotation, and it does not release any personal information beyond what the benchmarks already contain. The proposed diagnostics should not be interpreted as a guarantee of safe deployment: interaction-restricted correction does not establish causal grounding, and benchmark improvements do not by themselves validate a model for embodied use.

REPRODUCIBILITY STATEMENT

The architecture, reference measure, objective, and structural controls are specified in Section 4, with proofs and parameter counts in Appendix A. Appendix B gives the layer configuration, the risk-cohort rule, the optimizer and schedule for both backbones, and the step-by-step training procedure; Appendix C gives the data splits, benchmark metrics, comparison-method settings, the scene-removal intervention, the paired measurements, and the efficiency protocol. Every trained model starts from a released checkpoint identified in Section 5, uses seed 42, and follows a fixed one-epoch budget, and Appendix D reports the matched random masks for seeds 42, 43, and 44. The run configurations, checkpoint identities, raw predictions, and plotting inputs behind every table and figure are retained with the reported results.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, presentation refinement, candidate-literature identification, and the preparation of figure specifications. Generative AI was not used to generate experimental data or measurements, execute experiments, implement the proposed method, or serve as an authoritative source for mathematical claims, proofs, or scientific conclusions. All research hypotheses, method definitions, mathematical arguments, experimental protocols, analyses, and interpretations were reviewed, verified, and finalized by the authors. All cited works were checked against primary or bibliographic records. The authors take full responsibility for the final content of this work.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018. doi: 10.1109/CVPR.2018.00522. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Azuma_ScanQA_3D_Question_Answering_for_Spatial_Scene_Understanding_CVPR_2022_paper.html.

Rémi Cadene, Corentin Dancette, Hedi Ben-Younes, Matthieu Cord, and Devi Parikh. RUBi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/51d92be1c60d1db1d2e5e7a07da55b26-Abstract.html.

Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10800–10809, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_Counterfactual_Samples_Synthesizing_for_Robust_Visual_Question_Answering_CVPR_2020_paper.html.

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, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_LL3DA_Visual_Interactive_Instruction_Tuning_for_Omni-3D_Understanding_Reasoning_and_CVPR_2024_paper.html.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4069–4082, 2019. doi: 10.18653/v1/D19-1418. URL https://aclanthology.org/D19-1418/.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3D-LLaVA: Towards generalist 3D LMMs with omni superpoint transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3772–3782, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Deng_3D-LLaVA_Towards_Generalist_3D_LMMs_with_Omni_Superpoint_Transformer_CVPR_2025_paper.html.

Bradley Efron and Charles Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9(3): 586–596, 1981. doi: 10.1214/aos/1176345462.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z.

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. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. doi: 10.48550/arXiv.1503.02531. URL https://arxiv.org/abs/1503.02531.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196. URL https://doi.org/10.1214/aoms/1177730196.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0900. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/413885e70482b95dcbeeddc1daf39177-Abstract-Conference.html.

Giles Hooker. Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 16(3):709–732, 2007. doi: 10.1198/106186007X237892.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-Scene: Bridging 3D scene and large language models with object identifiers. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. doi: 10.52202/079017-3620. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/cebbd24f1e50bcb63d015611fe0fe767-Abstract-Conference.html.

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. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20413–20451, 2024b. URL https://proceedings.mlr.press/v235/huang24ae.html.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. CausalLens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026.

Yujin Jo, Sangyoon Bae, and Taesup Kim. Attention-space contrastive guidance for efficient hallucination mitigation in LVLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9706–9715, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Jo_Attention-Space_Contrastive_Guidance_for_Efficient_Hallucination_Mitigation_in_LVLMs_CVPRF_2026_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X. Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3D large language models really understand 3D spatial relationships? In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3vlMiJwo8b.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12700–12710, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Niu_Counterfactual_VQA_A_Cause-Effect_Look_at_Language_Bias_CVPR_2021_paper.html.

Makanjuola Adekunmi Ogunleye, Eman Abdelrahman, and Ismini Lourentzou. 3D-VCD: Hallucination mitigation in 3D-LLM embodied agents through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40197–40207, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ogunleye_3D-VCD_Hallucination_Mitigation_in_3D-LLM_Embodied_Agents_through_Visual_Contrastive_CVPR_2026_paper.html.

Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/67d96d458abdef21792e6d8e590244e7-Paper.pdf.

Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3698–3712, 2022. doi: 10.18653/v1/2022.findings-emnlp.271. URL https://aclanthology.org/2022.findings-emnlp.271/.

Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of Goodhart's law. In *Advances in Neural Information Processing Systems*, volume 33, pp. 407–417, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/045117b0e0a11a242b9765e79cbf113f-Abstract.html.

Chuan Wen, Jianing Qian, Jierui Lin, Jiaye Teng, Dinesh Jayaraman, and Yang Gao. Fighting fire with fire: Avoiding DNN shortcuts through priming. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23723–23750, 2022. URL https://proceedings.mlr.press/v162/wen22d.html.

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*, pp. 29501–29512, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_3D-GRAND_A_Million-Scale_Dataset_for_3D-LLMs_with_Better_Grounding_and_CVPR_2025_paper.html.

Appendix for *When Debiasing Relocates Shortcuts: Orthogonal Interaction Fine-Tuning for 3D Large Multimodal Models*

A Formal analysis of Corrective Function Allocation

This appendix establishes the function-space formulation in Section 3.3 and proves the orthogonality result in Section 4.4. We first characterize the interaction subspace, then show that the OIF parameterization belongs to it. Table A.1 summarizes the notation.

A.1 Function spaces and orthogonal decomposition

Let Q_L and Q_X be probability measures on the linguistic and scene input spaces, and set $Q = Q_L \otimes Q_X$. Consider the Hilbert space $\mathcal{H} = L^2(Q; \mathbb{R}^{d_h})$, with inner product and norm

$$\langle f, g \rangle_Q = \mathbb{E}_Q[f(L,X)^\top g(L,X)], \qquad \|f\|_Q^2 = \langle f, f \rangle_Q. \tag{9}$$

Functions are identified up to equality Q-almost everywhere. Define the following subspaces of $\mathcal{H}$:

$$\mathcal{H}_0 = \{(L,X) \mapsto c : c \in \mathbb{R}^{d_h}\}, \tag{10}$$

$$\mathcal{H}_L = \{(L,X) \mapsto g(L) : g \in L^2(Q_L; \mathbb{R}^{d_h}),\ \mathbb{E}_{Q_L}[g] = 0\}, \tag{11}$$

$$\mathcal{H}_X = \{(L,X) \mapsto r(X) : r \in L^2(Q_X; \mathbb{R}^{d_h}),\ \mathbb{E}_{Q_X}[r] = 0\}, \tag{12}$$

$$\mathcal{H}_{LX} = \{u \in \mathcal{H} : \mathbb{E}_Q[u \mid L] = 0,\ \mathbb{E}_Q[u \mid X] = 0\}. \tag{13}$$

We omit the argument Q when the reference is fixed. Conditional expectation is a bounded projection in L^2, so these subspaces are closed.

For any $f \in \mathcal{H}$, the two-input Hoeffding decomposition is obtained by setting (Hoeffding, 1948; Efron & Stein, 1981)

$$f_{\text{const}} = \mathbb{E}_Q[f], \tag{14}$$

$$f_L(L) = \mathbb{E}_Q[f \mid L] - f_{\text{const}}, \tag{15}$$

$$f_X(X) = \mathbb{E}_Q[f \mid X] - f_{\text{const}}, \tag{16}$$

$$f_{LX}(L,X) = f(L,X) - f_{\text{const}} - f_L(L) - f_X(X). \tag{17}$$

The first three components belong to $\mathcal{H}_0, \mathcal{H}_L, \mathcal{H}_X$, respectively. Independence under Q gives $\mathbb{E}_Q[f_X(X) \mid L] = \mathbb{E}_{Q_X}[f_X] = 0$, hence

$$\mathbb{E}_Q[f_{LX} \mid L] = \mathbb{E}_Q[f \mid L] - f_{\text{const}} - f_L(L) = 0. \tag{18}$$

The symmetric calculation gives $\mathbb{E}_Q[f_{LX} \mid X] = 0$.

The components are pairwise orthogonal. Constants are orthogonal to each centered component, and independence yields $\langle f_L, f_X \rangle_Q = 0$. For the interaction,

$$\langle f_{LX}, f_L \rangle_Q = \mathbb{E}_{Q_L}\left[\mathbb{E}_Q[f_{LX} \mid L]^\top f_L(L)\right] = 0, \tag{19}$$

with the same result for f_X. Consequently,

$$L^2(Q; \mathbb{R}^{d_h}) = \mathcal{H}_0 \oplus \mathcal{H}_L \oplus \mathcal{H}_X \oplus \mathcal{H}_{LX}. \tag{20}$$

Orthogonality also establishes uniqueness: the squared norm of the difference between two such decompositions is the sum of the squared norms of their component differences.

A.2 The corrective interaction space

Write $\mathcal{H}_{\text{main}} = \mathcal{H}_0 \oplus \mathcal{H}_L \oplus \mathcal{H}_X$. Equation 20 gives

$$\mathcal{H}_{LX} = \mathcal{H}_{\text{main}}^\perp. \tag{21}$$

Thus $u \in \mathcal{H}_{LX}$ is orthogonal to every function $c + g(L) + r(X)$ whose components are square-integrable, including those with nonzero means. Conversely, if u is orthogonal to all such functions, choose $g(L) = \mathbb{E}_Q[u \mid L]$. Then

$$0 = \mathbb{E}_Q[u^\top g(L)] = \mathbb{E}_{Q_L}\left[\|\mathbb{E}_Q[u \mid L]\|_2^2\right], \tag{22}$$

Table A.1: Notation for Corrective Function Allocation.

Symbol	Meaning
L, X, Y	linguistic context, explicit 3D scene, target answer
$\mathcal{C}_0$	frozen base 3D-LMM
h_t^0, ℓ_t^0	base state and logits for a given answer prefix
$\Delta_h, \Delta\ell$	added hidden-state and logit residuals
$l(L), s(X)$	frozen linguistic and scene summaries
$\widetilde{l}, \widetilde{s}$	centered and scaled summaries
d_L, d_X, d_h, R	summary dimensions, hidden dimension, interaction rank
$Q = Q_L \otimes Q_X$	product reference measure
$\mathcal{H}_0, \mathcal{H}_L, \mathcal{H}_X, \mathcal{H}_{LX}$	constant, linguistic, scene, and interaction subspaces
m_i	binary prior-risk mask

so $\mathbb{E}_Q[u \mid L] = 0$ almost everywhere. Choosing $r(X) = \mathbb{E}_Q[u \mid X]$ establishes the other condition.

The orthogonal projection onto the interaction subspace is therefore

$$\Pi_{LX} f = f - \mathbb{E}_Q[f \mid L] - \mathbb{E}_Q[f \mid X] + \mathbb{E}_Q[f]. \tag{23}$$

Corrective Function Allocation restricts Δ_h to the range of this projection. The pretrained predictor $\mathcal{C}_0$ remains fixed, retaining its existing main effects and interactions.

A.3 Membership of the OIF residual

The following result applies to general product references as well as the finite empirical reference used by OIF.

Theorem A.1 (OIF membership in the interaction space)**.** *Let $Q = Q_L \otimes Q_X$. Suppose the deterministic summary maps $\widetilde{l} \in L^2(Q_L; \mathbb{R}^{d_L})$ and $\widetilde{s} \in L^2(Q_X; \mathbb{R}^{d_X})$ satisfy $\mathbb{E}_{Q_L}[\widetilde{l}] = 0$ and $\mathbb{E}_{Q_X}[\widetilde{s}] = 0$. For any finite matrices $W_L \in \mathbb{R}^{R \times d_L}$, $W_X \in \mathbb{R}^{R \times d_X}$, and $W_O \in \mathbb{R}^{d_h \times R}$, the residual*

$$\Delta_h(L, X) = W_O\Big[(W_L \widetilde{l}(L)) \odot (W_X \widetilde{s}(X))\Big] \tag{24}$$

belongs to $\mathcal{H}_{LX}(Q)$.

Proof. First, the residual is square-integrable. The inequality $\|a \odot b\|_2 \le \|a\|_2 \|b\|_2$ and independence under Q imply

$$\begin{aligned} \mathbb{E}_Q\|\Delta_h\|_2^2 \le \|W_O\|_{\mathrm{op}}^2 \|W_L\|_{\mathrm{op}}^2 \|W_X\|_{\mathrm{op}}^2 \\ \times \mathbb{E}_{Q_L}\|\widetilde{l}\|_2^2 \, \mathbb{E}_{Q_X}\|\widetilde{s}\|_2^2 < \infty. \end{aligned} \tag{25}$$

Next define $a(L) = W_L \widetilde{l}(L)$ and $b(X) = W_X \widetilde{s}(X)$. Their marginal means vanish by linearity. Conditioning on L yields

$$\mathbb{E}_Q[\Delta_h \mid L] = W_O \mathbb{E}_Q[a(L) \odot b(X) \mid L] \tag{26}$$

$$= W_O[a(L) \odot \mathbb{E}_{Q_X}[b(X)]] = 0. \tag{27}$$

Similarly, $\mathbb{E}_Q[\Delta_h \mid X] = W_O[\mathbb{E}_{Q_L}[a(L)] \odot b(X)] = 0$. These two identities establish membership in $\mathcal{H}_{LX}(Q)$. □

The theorem holds for every finite parameter value. Since training changes only the matrices, the interaction constraint remains invariant throughout optimization. Equation 7 follows by choosing the empirical product reference $\widehat{Q}_{\mathrm{ref}}$ of Section 4.2, for which finite feature values ensure the required second moments.

A.4 From hidden-state residuals to logit residuals

For any fixed prefix $y_{<t}$, the frozen language-model head gives

$$\Delta\ell(L, X) \equiv \ell_t^{\mathrm{OIF}}(L, X, y_{<t}) - \ell_t^0(L, X, y_{<t}) \tag{28}$$

$$= W_{\mathrm{LM}} \Delta_h(L, X). \tag{29}$$

The head's bias cancels in the difference. The fixed linear map preserves square integrability and commutes with conditional expectation:

$$\mathbb{E}_Q[\Delta\ell \mid L] = W_{\mathrm{LM}}\mathbb{E}_Q[\Delta_h \mid L] = 0, \qquad \mathbb{E}_Q[\Delta\ell \mid X] = 0. \tag{30}$$

Hence $\Delta\ell$ lies in the corresponding interaction subspace of $L^2(Q; \mathbb{R}^{|\mathcal{V}|})$, where $\mathcal{V}$ is the vocabulary.

The added logit term is constant across answer positions for a given (L, X). Independently generated sequences can acquire different prefixes, at which point their base hidden states also differ. After softmax, the probability difference

$$\mathrm{softmax}(\ell_t^0 + \Delta\ell) - \mathrm{softmax}(\ell_t^0) \tag{31}$$

depends nonlinearly on the base logits and the residual, so its conditional means can be nonzero even when those of $\Delta\ell$ vanish.

A.5 The empirical product reference

Let $\{(L_i, X_i, Y_i)\}_{i=1}^N$ be the SQA3D training rows. Define measures on the input spaces,

$$\widehat{P}_L^{\mathrm{row}} = \frac{1}{N}\sum_{i=1}^N \delta_{L_i}, \qquad \widehat{P}_X^{\mathrm{row}} = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}. \tag{32}$$

Repeated scenes receive mass in proportion to their number of rows. Their product, $\widehat{Q}_{\mathrm{ref}}$, corresponds to independently selecting two row indices and combining the linguistic input from one with the scene from the other. For any integrable g,

$$\mathbb{E}_{\widehat{Q}_{\mathrm{ref}}}[g] = \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N g(L_i, X_j). \tag{33}$$

The paired empirical distribution instead assigns mass to the original (L_i, X_i) pairs.

Write $l_i = l(L_i)$ and $s_i = s(X_i)$. The stored means are $\mu_L = N^{-1}\sum_i l_i$ and $\mu_X = N^{-1}\sum_i s_i$. With fixed positive diagonal scale matrices D_L, D_X, centering gives

$$\frac{1}{N}\sum_i D_L^{-1}(l_i - \mu_L) = 0, \qquad \frac{1}{N}\sum_i D_X^{-1}(s_i - \mu_X) = 0. \tag{34}$$

In Equation 4, the diagonal entries are the empirical standard deviations plus ϵ. Positivity handles constant coordinates, and fixed scaling preserves the zero means. The resulting residual obeys the exact Cartesian identities

$$\frac{1}{N}\sum_j \Delta_h(L_i, X_j) = 0, \qquad \frac{1}{N}\sum_i \Delta_h(L_i, X_j) = 0 \tag{35}$$

for every training L_i and X_j, in exact arithmetic.

A different weighting defines a different reference. For instance, equal total weight per scene changes the marginals when scenes have unequal row counts. Recomputing both means under the new weighting yields a centered parameterization to which Theorem A.1 applies. Evaluating an existing OIF residual with its original means under a scene-balanced reference instead measures sensitivity to that change of measure. Conditional means on the observed paired distribution can also be nonzero because its linguistic and scene inputs are dependent (Hooker, 2007).

A.6 Structural null and initialization

The multiplicative form gives

$$\widetilde{l}(L) = 0 \text{ or } \widetilde{s}(X) = 0 \quad \Longrightarrow \quad \Delta_h(L, X) = 0. \tag{36}$$

These null inputs have summaries equal to their reference means. The scene-removal intervention of Appendix C.3 realizes this null for the scene branch: it removes the object-token span from the prefix and sets the centered scene summary to its neutral point $\widetilde{s}(X) = 0$, so the residual vanishes under that intervention. An input intervention that alters the scene tokens without resetting the centered summary does not fall under this identity and in general yields a nonzero residual.

Initializing $W_O^{(0)} = 0$ makes the residual vanish for every input, so the initial hidden states, logits, and token distributions equal those of $\mathcal{C}_0$ at every shared prefix. Identical deterministic decoding then produces identical sequences. At initialization, only W_O can receive a nonzero gradient from the training objective. Gradients to the input projections become available as W_O moves away from zero.

Interaction membership specifies the statistical form of the added correction under Q. A function in $\mathcal{H}_{LX}(Q)$ can encode either a relevant spatial relation or a spurious cross-modal association. Evidence identification therefore depends on the learned interaction, which the viewpoint, hallucination, and intervention experiments evaluate.

A.7 Control classes and parameter counts

Centering separates the additive and interaction controls algebraically. For projected features $a(L)$ and $b(X)$, the ADD residual in Equation 8 belongs to $\mathcal{H}_L \oplus \mathcal{H}_X$, whereas OIF belongs to $\mathcal{H}_{LX}$. Their representable functions intersect only at zero under Q. When both classes contain nonzero functions, neither contains the other.

At a common rank R, the Full residual in Equation 8 contains OIF by choosing $U_I = W_O$ and $U_L = U_X = 0$. The inclusion is strict whenever a nonzero main effect can be represented. The Full residual thus relaxes the allocation constraint at fixed rank, and Full-BM additionally lowers the rank so that the comparison controls parameter budget.

The parameter counts follow directly from the matrix dimensions:

$$P_{\mathrm{OIF}}(R) = R(d_L + d_X + d_h), \tag{37}$$

$$P_{\mathrm{ADD}}(R) = R(d_L + d_X + d_h), \tag{38}$$

$$P_{\mathrm{Full}}(R) = R(d_L + d_X + 3d_h). \tag{39}$$

With unit rank granularity, the largest Full rank within OIF's rank-512 budget is

$$R_{\mathrm{BM}} = \left\lfloor \frac{512(d_L + d_X + d_h)}{d_L + d_X + 3d_h} \right\rfloor. \tag{40}$$

With $d_L = 4096$, $d_X = 3073$, and $d_h = 4096$, this gives $R_{\mathrm{BM}} = 296$ and 5.76M trainable parameters, against 5.77M for OIF and MainEffect at $R = 512$.

The bias-free construction also identifies how main effects can reappear. Adding input-projection biases gives

$$(a + c_L) \odot (b + c_X) = a \odot b + a \odot c_X + c_L \odot b + c_L \odot c_X. \tag{41}$$

The additional terms are a linguistic main effect, a scene main effect, and a constant. Similarly, a nonlinear transform of centered features generally requires recentering to restore zero means. OIF uses fixed centering followed by linear projections and one multiplicative interaction, preserving the conditions of Theorem A.1.

B Implementation details

This section specifies the trainable correction path, the construction of the fixed risk cohort, the loss routing used during fine-tuning, and the matched structural controls. All backbone parameters, object features, and the language-model head remain frozen throughout OIF training.

B.1 Architecture and initialization

Let $e_1, \ldots, e_{|L|}$ denote the frozen token embeddings of the concatenated situation and question. The linguistic summary is their mean,

$$l(L) = \frac{1}{|L|} \sum_{j=1}^{|L|} e_j. \tag{42}$$

Table B.1: OIF implementation. All trainable maps are bias-free.

Component	Specification
Base model and LM head	Frozen end to end
Linguistic input	Situation and question
Linguistic pooling	Mean frozen token embedding
Scene input	Frozen per-object features
Scene pooling	Mean, std. (with 10^{-6} inside the root), max, and $\log(1+K)$
Centering statistics	SQA3D-train mean and std. (Equation 4); $\epsilon = 10^{-6}$
Interaction rank	$R = 512$
Fusion	$W_L\widetilde{l} \odot W_X\widetilde{s}$
Nonlinearity / norm / dropout	None / none / none
Output location	Final decoder state before the frozen LM head
Temporal form	One residual broadcast to answer positions
Initialization	Xavier for W_L, W_X; zero for W_O
Trainable parameters	5.77M

For frozen per-object scene features $x_1, \ldots, x_K$, the scene summary concatenates the coordinate-wise mean, standard deviation, and maximum with an object-count statistic:

$$s(X) = \left[\bar{x}; \hat{\sigma}_x; \max_k x_k; \log(1+K)\right], \tag{43}$$

where $\bar{x} = K^{-1}\sum_k x_k$ and $\hat{\sigma}_x = \sqrt{K^{-1}\sum_k (x_k - \bar{x})^2 + 10^{-6}}$ are per-scene statistics, distinct from the training-split centering statistics μ_X and σ_X of Equation 4. Both summaries are standardized with coordinate-wise statistics estimated once on the SQA3D training split. These statistics remain fixed during training and evaluation. Table B.1 summarizes the configuration.

The residual is added to the final decoder hidden state before the frozen language-model head. Only W_O is initialized to zero, so the initial residual vanishes exactly while gradients can immediately reach the two input projections after the first update.

B.2 Risk cohort and loss routing

The risk mask is computed before corrective training and stored with the training manifest. The prior diagnostic first identifies 13,159 prior-conflict rows among the 26,623 SQA3D training examples, namely the rows on which the answer generated by the blind model disagrees with the ground truth. The frozen parent then scores each such row twice, once under linguistic-only input and once under full input, by the margin between the mean token log-probability of the ground-truth answer and that of the blind model's incorrect answer. A row whose two margins are both non-positive enters the `STRICT_PRIOR_PERSISTENT` cohort, since the parent prefers the incorrect prior answer even when the scene is present. A non-positive linguistic-only margin together with a positive full-input margin marks a row that the scene already resolves, the reverse sign pattern marks full-input interference, and two positive margins mark a row on which the current policy is already correct. OIF activates the `STRICT_PRIOR_PERSISTENT` cohort, which contains 5,162 examples, or 19.389% of the training split; the other three categories contain 1,900, 502, and 5,595 rows, respectively. The diagnostic therefore determines where corrective supervision is applied; its predicted answer is never used as a correction target.

For an active row, OIF minimizes ground-truth cross-entropy over the response tokens. For every other row, it minimizes forward KL divergence from the stop-gradient distribution of the frozen parent. The two branches are combined without temperature scaling, oversampling, class rebalancing, or an additional loss coefficient. The risk manifest and parent distributions are fixed for the entire run. Table B.2 lists the full procedure.

Optimization on Chat-Scene uses AdamW with a peak learning rate of 10^{-4}, weight decay 0.01, and gradient-norm clipping at 1.0 on the three trainable matrices, in bf16 precision with seed 42; two further training seeds are reported in Appendix D.3. The global batch is 128 rows, so one epoch takes 208 updates, with a short linear warmup followed by cosine decay. The structural controls

Table B.2: OIF training procedure. The parent model, scene features, risk manifest, and centering statistics are fixed before optimization.

Step	Operation
1	Load the frozen parent, frozen object features, risk manifest, and SQA3D-train centering statistics.
2	Compute $l(L)$ and $s(X)$ without gradients through the parent, then obtain $\widetilde{l}$ and $\widetilde{s}$ using the fixed statistics.
3	Compute $a = W_L\widetilde{l}$, $b = W_X\widetilde{s}$, $z = a \odot b$, and $\Delta_h = W_O z$.
4	Broadcast Δ_h to the answer-token hidden states and apply the frozen language-model head.
5	Apply response-token cross-entropy when $m_i = 1$ and forward KL from the stop-gradient parent distribution when $m_i = 0$.
6	Backpropagate only through W_L, W_X, W_O and update them for one epoch.
7	Retain the run configuration, seed, checkpoint identity, raw predictions, and the logits required by the paired analyses.

and matched random masks share this recipe. The 3D-LLaVA transfer keeps the same trainable set, rank, seed, and one-epoch budget, and uses AdamW with a peak learning rate of 2×10^{-5}, no weight decay, the same warmup-then-cosine schedule, gradient clipping at 1.0, and an effective batch of 32 rows.

B.3 Structural controls

The controls preserve the same frozen parent, linguistic and scene summaries, centering statistics, risk cohort, loss routing, output location, and one-epoch budget. ***MainEffect*** is the additive control denoted ADD in the formulation; it changes only the fusion operator from $a \odot b$ to $a + b$. ***Full-BM*** retains the interaction branch and adds separate linguistic and scene main-effect branches, with its rank reduced to $R = 296$ (5.76M parameters) to match OIF's 5.77M trainable-parameter budget (Appendix A.7). Each control uses separate output maps for its active branches, and every output map is zero-initialized to preserve the matched parent's predictions at step zero.

C Evaluation and measurement protocols

C.1 Data, selectors, and benchmark metrics

Training uses the 26,623-example SQA3D training split (Ma et al., 2023). The one-epoch budget applies to OIF and all newly trained corrective variants; 3DR-FT follows its official recipe. The prior-aware selector activates 5,162 examples, or 19.389% of the training split. Each matched random selector preserves this active count and the joint answer-family $\times$ question-type composition; seeds 42, 43, and 44 identify the three sampled masks. Table D.1 reports all three runs and their mean.

Real-3DQA (Ma et al., 2026) evaluates debiased question answering on 1,485 examples and viewpoint consistency on their rotated variants, using exact match (EM), refined exact match (EM-R), and the official Viewpoint Rotation Score. 3D-POPE (Yang et al., 2025) evaluates object-existence hallucination with binary existence queries on its Random, Popular, and Adversarial subsets. For a group g containing the original viewpoint and three rotated variants, let $c_g \in \{0, 1, 2, 3, 4\}$ count correct predictions. With G groups,

$$P_{\geq k} = \frac{100}{G} \sum_{g=1}^{G} \mathbb{1}[c_g \geq k], \qquad \mathrm{VRS} = \frac{1}{4} \sum_{k=1}^{4} P_{\geq k}. \tag{44}$$

Dir-VRS applies the same calculation to direction questions. Table 1 reports pooled precision, F1, accuracy, and affirmative-response rate; Table 2 reports precision, F1, accuracy, and affirmative-response rate separately for each subset. In the pooled panel and in the cross-backbone and mechanism summaries, predictions from the three equal-size subsets are pooled before the confusion-count

metrics are recomputed. This convention explains why the reported pooled F1 need not equal the arithmetic mean of the three subset F1 scores.

C.2 Comparison methods and controls

3DR-FT (Ma et al., 2026) is the prior-aware fine-tuning baseline. 3D-VCD (Ogunleye et al., 2026), CausalLens (Ji et al., 2026), and ACG (Jo et al., 2026) are inference-time interventions adapted to Chat-Scene; 3D-VCD requires a second decoding pass, whereas the other methods and OIF retain a single pass. The distorted pass of 3D-VCD permutes the object-level 3D and 2D features across object slots with a fixed per-scene derangement and adds Gaussian noise of standard deviation 0.05 to the object locations, and the two passes are contrasted with weight $\alpha = 1.0$. CausalLens applies its scene-sensitivity-weighted attention intervention in decoder layers 10 to 20 with strength 0.15, and ACG propagates its conditional and scene-masked trajectories through all 32 decoder layers with guidance scale $\gamma = 2.4$. Every method decodes with the native configuration of the target model, namely beam search with five beams on Real-3DQA and greedy decoding on 3D-POPE. MainEffect allows linguistic and scene main-effect corrections, whereas Full-BM adds both main effects to the interaction branch at OIF's parameter budget; Appendix B specifies both controls. Three matched random masks preserve the size and answer-family $\times$ question-type composition of the prior-aware cohort, and All rows applies correction to every training example. The cross-backbone experiment transfers OIF to 3D-LLaVA (Deng et al., 2025) without altering its native visual pipeline.

C.3 Reference checkpoints and scene interventions

The main benchmark tables use the released Chat-Scene checkpoint as their visible reference. Paired mechanism measurements compare each corrected model with the checkpoint from which that run was initialized, which is the same released checkpoint for every method. For the 3DR-FT diagnosis, M_0 is the released checkpoint and the corrected model is $M_1 =$ 3DR-FT, so the mechanism study and the main tables share one reference.

Each paired example is evaluated under full input and under one scene-removal intervention that is defined once and applied identically to every arm. The intervention consists of the same pair of operations for every model: the object-token span is removed from the prefix, so that the language model receives the linguistic context alone, and, for every arm whose residual reads a scene summary, the centered scene summary is set to its neutral point $\widetilde{s} = 0$. The linguistic context, the frozen parent, and the decoding configuration are unchanged, so a cross-arm difference reflects the arm's own residual algebra rather than the operator. Because every projection is bias-free, this operator gives $b = W_X \widetilde{s} = 0$, and the OIF residual $W_O(a \odot b)$ vanishes for every example, whereas the additive and full-budget controls retain their linguistic branches, $W_O a$ and $U_L a$ respectively. The resulting zero linguistic-only change count for OIF is a structural null: it verifies the allocation constraint, but does not by itself establish that each full-input prediction is causally grounded in the intended scene evidence.

C.4 Paired prediction-change measurements

For a corrective model M and its parent M_0, with generated prediction $\hat{y}^c_{M,i}$ under condition $c \in \{F, T\}$, the change indicators

$$F_i = \mathbb{1}[\hat{y}^F_{M,i} \neq \hat{y}^F_{M_0,i}], \qquad T_i = \mathbb{1}[\hat{y}^T_{M,i} \neq \hat{y}^T_{M_0,i}] \tag{45}$$

compare the corrected model and its matched parent within each input condition. Let $N_F = \sum_i F_i$, $N_T = \sum_i T_i$, and $N_B = \sum_i F_i T_i$. Scene-neutralized retention conditions on full-input changes,

$$R_{\text{retain}} = \frac{\sum_i F_i T_i}{\sum_i F_i} = \frac{N_B}{N_F}. \tag{46}$$

It records that an update effect occurs under both conditions, not that the two corrected predictions are identical.

To measure agreement in the resulting target, define

$$S_i = F_i T_i \mathbb{1}[\hat{y}^F_{M,i} = \hat{y}^T_{M,i}], \tag{47}$$

with

$$R^F_{\text{target}} = \frac{\sum_i S_i}{N_F}, \qquad R^B_{\text{target}} = \frac{\sum_i S_i}{N_B}. \tag{48}$$

For the detailed 3DR-FT diagnosis, $N_F = 873$, $N_T = 786$, $N_B = 618$, and $\sum_i S_i = 570$. Hence $R_{\text{retain}} = 70.79\%$, $R^F_{\text{target}} = 65.29\%$, and $R^B_{\text{target}} = 92.23\%$. The remaining joint counts are $N_{F\text{-only}} = 255$, $N_{T\text{-only}} = 168$, and $N_{\text{neither}} = 444$, for a union of 1,041 changed examples. Family-wise rates use N_F restricted to the relevant family as their denominator.

C.5 Marginal answer-policy drift

Policy-TV is the total variation distance between the answer-family distributions of M and M_0,

$$D_{\text{TV}}(M, M_0) = \frac{1}{2} \sum_{a \in \mathcal{A}} |p_M(a) - p_{M_0}(a)|, \tag{49}$$

computed from normalized answer frequencies over six predeclared families $\mathcal{A}$: yes/no, left/right, front/behind, up/down, counting, and other. For every corrective variant, the reference distribution is produced by its matched parent on the same evaluation rows and under the same input condition. The parent therefore has Policy-TV zero by definition. Comparisons of the smallest nonzero drift are restricted to the learned corrective variants; Policy-TV is interpreted jointly with task and robustness metrics rather than as an independent objective.

C.6 Teacher-forced ground-truth scores

For a ground-truth answer with A_i scored tokens, we use its mean teacher-forced token log-probability,

$$s^c_M(i) = \frac{1}{A_i} \sum_{t=1}^{A_i} \log p_M(y_{i,t} \mid y_{i,<t}, L_i, X^c_i), \quad X^F_i = X_i, \quad X^T_i = \varnothing. \tag{50}$$

The answer-token mask and length normalization follow the Chat-Scene scoring implementation. The update-induced shift is

$$g_c(i) = s^c_{M_1}(i) - s^c_{M_0}(i), \qquad c \in \{F, T\}. \tag{51}$$

A positive value raises the ground-truth answer's mean log-probability; a negative value lowers it. For the 3DR-FT diagnosis, the observed means are $\mathbb{E}[g_F] = -12.3254$ and $\mathbb{E}[g_T] = -11.9867$, with paired difference $\mathbb{E}[g_F - g_T] = -0.3387$. Pearson and Spearman correlations between the two shifts are 0.9840 and 0.9761 across the 1,485 examples.

C.7 Bootstrap protocol

Confidence intervals use 10,000 paired row-bootstrap resamples with seed 42. Each resample draws 1,485 row indices with replacement, and all checkpoint-condition records for a selected row are resampled together. We report percentile 95% intervals and no p-values. For the detailed 3DR-FT analysis, the resampled estimators are R_{retain}, R^F_{target}, R^B_{target}, $\mathbb{E}[g_F]$, $\mathbb{E}[g_T]$, $\mathbb{E}[g_F - g_T]$, Pearson correlation, and Spearman correlation. The same paired procedure is applied separately to the retention statistic of each learned control in Figure 4. OIF's zero is not a sampled estimate and is labeled as a structural null rather than a zero-width empirical interval.

C.8 Inference-efficiency protocol

Measurements use NVIDIA A800 80GB PCIe GPUs at batch size 1. Each method is evaluated on the same 256 Real-3DQA sample IDs with native beam-5 decoding and on the same 258 3D-POPE sample IDs with greedy decoding; the latter contain 86 examples from each subset. Table D.11 reports median end-to-end latency over three complete repetitions and peak allocated GPU memory.

Per-sample latency includes each method's native generation length. Relative to OIF, 3DR-FT produces approximately $1.65\times$ as many decoding steps per example on Real-3DQA and $3.65\times$ as many on 3D-POPE. The measured Real-3DQA latencies per generated step are 86.38 ms for OIF, 73.83 ms for 3DR-FT, and 594.75 ms for 3D-VCD. We therefore interpret the reported latency as end-to-end deployment cost rather than kernel-level decoder speed.

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 MATCHED RANDOM MASKS BY SEED

Table D.1 lists the three matched random-mask runs whose mean appears in Table 3. The prior-aware selector exceeds every individual mask on the pooled panel and on the Random and Popular subsets; on Adversarial, the seed-42 mask is ahead by 0.03 F1 and 0.08 accuracy, which is smaller than the spread among the three masks (at most 0.15 pooled F1).

Table D.1: Matched random masks by seed on 3D-POPE (%). Each mask keeps the active count (19.39%) and the answer-family × question-type composition of the prior-aware mask. Bold marks the column-best score.

Selector	Seed	Overall				Random			Popular			Adversarial		
		P ↑	F1 ↑	Acc. ↑	Yes ↓	F1 ↑	Acc. ↑	Yes ↓	F1 ↑	Acc. ↑	Yes ↓	F1 ↑	Acc. ↑	Yes ↓
Matched random	42	65.24	77.03	71.97	72.06	79.67	76.04	67.87	76.81	71.62	72.37	**74.79**	**68.25**	**75.95**
Matched random	43	65.16	77.02	71.90	72.25	79.83	76.18	68.07	76.80	71.57	72.54	74.60	67.96	76.15
Matched random	44	64.91	76.88	71.65	72.62	79.36	75.51	68.63	76.74	71.41	72.90	74.70	68.04	76.33
Matched random	Mean	65.10	76.98	71.84	72.31	79.62	75.91	68.19	76.78	71.53	72.60	74.70	68.08	76.14
OIF	–	**65.42**	**77.20**	**72.19**	**71.97**	**80.22**	**76.75**	**67.53**	**76.83**	**71.66**	**72.30**	74.76	68.17	76.08

D.2 PAIRED CONFIDENCE INTERVALS FOR OIF DIFFERENCES

Tables D.2 and D.3 give paired row-bootstrap 95% confidence intervals (10,000 resamples, seed 42) for OIF minus each comparison run, resampling the 20,568 pooled 3D-POPE rows and the 1,485 Real-3DQA questions with the same index draw for both runs. Every interval against the allocation arms, the all-rows arm, and the matched random-mask mean excludes zero on pooled F1, accuracy, and affirmative rate; against the individual seed-42 and seed-43 masks the F1 and accuracy intervals include zero, consistent with the margins in Table D.1.

Table D.2: OIF minus comparison run on pooled 3D-POPE (points), with paired-bootstrap 95% confidence intervals.

OIF minus	P	F1	Acc.	Yes
3DR-FT	$+14.91\ [14.45, 15.37]$	$+10.36\ [9.93, 10.79]$	$+21.20\ [20.53, 21.86]$	$-25.81\ [-26.44, -25.18]$
MainEffect	$+0.94\ [0.72, 1.17]$	$+0.40\ [0.17, 0.63]$	$+0.88\ [0.58, 1.18]$	$-1.66\ [-1.96, -1.37]$
Full-BM	$+1.18\ [0.93, 1.44]$	$+0.48\ [0.23, 0.73]$	$+1.08\ [0.76, 1.41]$	$-2.15\ [-2.48, -1.83]$
All rows	$+0.69\ [0.43, 0.95]$	$+0.44\ [0.19, 0.68]$	$+0.73\ [0.41, 1.06]$	$-0.87\ [-1.20, -0.54]$
Matched random (mean)	$+0.32\ [0.12, 0.51]$	$+0.23\ [0.03, 0.42]$	$+0.35\ [0.11, 0.60]$	$-0.34\ [-0.59, -0.09]$
Matched random (seed 42)	$+0.18\ [-0.06, 0.42]$	$+0.17\ [-0.07, 0.41]$	$+0.23\ [-0.07, 0.53]$	$-0.09\ [-0.39, 0.20]$
Matched random (seed 43)	$+0.26\ [0.03, 0.50]$	$+0.19\ [-0.05, 0.42]$	$+0.29\ [-0.01, 0.59]$	$-0.28\ [-0.58, 0.01]$
Matched random (seed 44)	$+0.51\ [0.27, 0.74]$	$+0.32\ [0.08, 0.55]$	$+0.54\ [0.23, 0.84]$	$-0.65\ [-0.95, -0.35]$

Table D.3: OIF minus comparison run on Real-3DQA (points), with paired-bootstrap 95% confidence intervals.

OIF minus	EM	EM-R
Released checkpoint	$+1.68\ [0.13, 3.30]$	$+1.95\ [0.27, 3.64]$
3DR-FT	$+3.77\ [1.48, 6.06]$	$+5.45\ [3.03, 7.88]$

D.3 TRAINING-SEED REPLICATES OF OIF

The OIF row in the main text is the seed-42 run. To measure run-to-run variation, we retrained OIF with training seeds 43 and 44 under an otherwise identical protocol: the same released checkpoint, the same three trainable matrices, the same prior-aware cohort and mask, and the same one-epoch schedule of 208 updates; only the seed that governs initialization and data order changes. Table D.4 reports the three runs. The spread across seeds is small: the sample standard deviation is at most 0.46 points on every column, and the mean over seeds reproduces the ordering reported in the main text. Averaged over the three seeds, OIF improves EM by 1.93 points and EM-R by 2.07 points over the released checkpoint, and raises Dir-VRS from 19.69 to 21.09; both replicate seeds obtain a higher

Dir-VRS than the seed-42 run. Table D.5 gives paired-bootstrap intervals for the seed mean minus the released checkpoint on Real-3DQA and minus the budget-matched arms on pooled 3D-POPE, computed as in Appendix D.2 with the seed-mean statistic recomputed inside every resample. Every interval excludes zero: the Real-3DQA gains over the released checkpoint and the pooled precision, F1, accuracy, and affirmative-rate advantages over MainEffect, Full-BM, and the all-rows arm hold for the seed mean, not only for the single reported run.

Table D.4: OIF under three training seeds (%). Real-3DQA reports EM, EM-R, VRS, and Dir-VRS; 3D-POPE reports precision, F1, accuracy, and affirmative rate pooled over the three subsets. Seed 42 is the run reported in the main text; the last row gives the mean and sample standard deviation over the three seeds.

	Real-3DQA				3D-POPE (pooled)			
Training seed	EM ↑	EM-R ↑	VRS ↑	Dir-VRS ↑	P ↑	F1 ↑	Acc. ↑	Yes ↓
42 (main text)	19.60	24.44	11.57	20.58	65.42	77.20	72.19	71.97
43	19.60	24.38	11.50	21.46	65.25	77.08	72.00	72.15
44	20.34	24.85	11.97	21.24	64.86	76.79	71.56	72.56
Mean$_{\pm s.d.}$	$19.85_{\pm 0.43}$	$24.56_{\pm 0.26}$	$11.68_{\pm 0.25}$	$21.09_{\pm 0.46}$	$65.18_{\pm 0.29}$	$77.02_{\pm 0.21}$	$71.92_{\pm 0.32}$	$72.23_{\pm 0.30}$

Table D.5: Seed-mean OIF minus comparison run (points), with paired-bootstrap 95% confidence intervals (10,000 resamples, seed 42). Real-3DQA differences are against the released checkpoint over the 1,485 questions; 3D-POPE differences are against the budget-matched arms over the 20,568 pooled rows.

Seed-mean OIF minus	Real-3DQA	
	EM	EM-R
Released checkpoint	$+1.93\ [0.65, 3.21]$	$+2.07\ [0.72, 3.43]$

Seed-mean OIF minus	3D-POPE (pooled)			
	P	F1	Acc.	Yes
MainEffect	$+0.70\ [0.52, 0.87]$	$+0.22\ [0.04, 0.40]$	$+0.60\ [0.37, 0.83]$	$-1.41\ [-1.64, -1.18]$
Full-BM	$+0.93\ [0.74, 1.13]$	$+0.30\ [0.10, 0.50]$	$+0.81\ [0.56, 1.07]$	$-1.90\ [-2.16, -1.64]$
All rows	$+0.44\ [0.22, 0.66]$	$+0.26\ [0.04, 0.47]$	$+0.46\ [0.18, 0.74]$	$-0.62\ [-0.90, -0.34]$

D.4 Cross-method mechanism measurements

Table D.6 provides the exact values summarized in Figure 4. Full-input and linguistic-only change counts measure the size of an update relative to the released Chat-Scene checkpoint from which every method starts. These descriptive counts are not objectives. Retention, Policy-TV, and the pooled 3D-POPE accuracy and affirmative rate jointly characterize whether a correction remains active without scene evidence, reorganizes the marginal answer policy, and preserves calibrated existence judgments on the 20,568-row 3D-POPE panel. Table D.7 reports the paired-bootstrap intervals for the retention estimates.

Table D.6: Cross-method mechanism scorecard. Full and linguistic-only changes compare each method with the released Chat-Scene checkpoint. Same-target is normalized by full-input changes. POPE Acc. and Yes pool the three 3D-POPE subsets (20,568 rows). OIF's zero linguistic-only change count and retention follow from the scene-neutralized structural intervention; among the learned variants, OIF has the lowest policy drift and affirmative rate and is the only one whose accuracy exceeds the released checkpoint.

Method	Full chg.	Ling.-only chg.	Retain ↓	Same target ↓	Policy-TV ↓	POPE Acc. ↑	POPE Yes ↓
Released Chat-Scene	–	–	–	–	0.0000	72.01	73.76
3DR-FT	873	786	0.708	0.653	0.0444	51.00	97.78
MainEffect	411	269	0.328	0.246	0.0108	71.32	73.63
Full-BM	428	249	0.287	0.234	0.0135	71.11	74.12
OIF	330	0	**0.000**	**0.000**	**0.0054**	**72.19**	**71.97**

Table D.7: Retention estimates with paired-bootstrap 95% confidence intervals. OIF is a structural null rather than an empirical zero.

Method	R_{retain}	95% CI
3DR-FT	0.708	[0.677, 0.738]
MainEffect	0.328	[0.283, 0.374]
Full-BM	0.287	[0.245, 0.331]
OIF	**0.000**	structural null

D.5 Detailed Shortcut Relocation diagnosis

Figure D.1 expands the 3DR-FT case using the released Chat-Scene checkpoint as its parent. It separates the geometry of paired prediction changes, their concentration across answer families, and the alignment of correction-induced ground-truth score shifts under full and linguistic-only input.

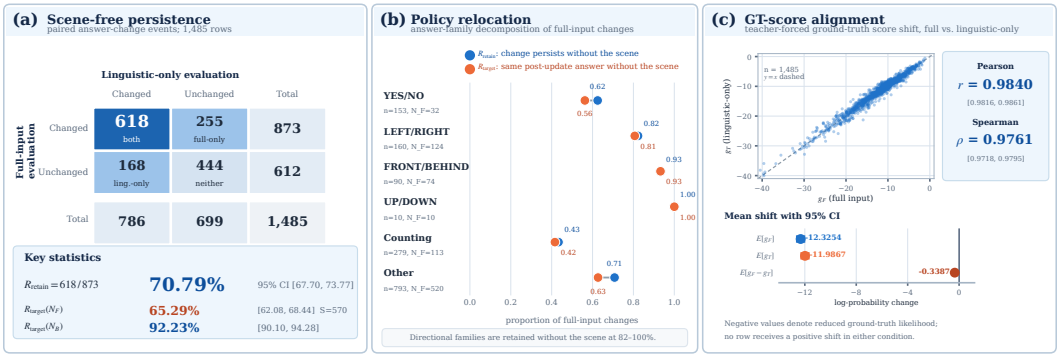


Figure D.1: Complete 3DR-FT scene-removal diagnosis. The parent and corrected checkpoints are evaluated on the same 1,485 examples under full and linguistic-only input. Panel (a) reports paired change events and conditional persistence. Panel (b) decomposes persistence and same-target agreement across answer families. Panel (c) compares teacher-forced ground-truth score shifts across the two conditions. Negative score shifts denote reduced ground-truth likelihood.

D.5.1 Event geometry and uncertainty

Table D.8 gives the exact joint counts behind panel (a). The 618 changes observed under both conditions form the numerator of conditional persistence, while the 873 full-input changes form its denominator. Thus $R_{\text{retain}} = 618/873 = 70.79\%$. Table D.9 reports paired-bootstrap intervals for every estimator in this subsection.

Table D.8: Joint answer-change counts for 3DR-FT relative to the released checkpoint across 1,485 paired examples.

Full-input prediction	Linguistic-only prediction		Total
	Changed	Unchanged	
Changed	**618**	255	873
Unchanged	168	444	612
Total	786	699	1,485

For full input, the parent and 3DR-FT teacher-forced means are -6.5475 and -18.8729, giving $\mathbb{E}[g_F] = -12.3254$. Under linguistic-only input, the corresponding means are -7.4221 and -19.4088, giving $\mathbb{E}[g_T] = -11.9867$. The update lowers average ground-truth likelihood by nearly the same amount with and without the scene (paired difference -0.34), no example receives a positive shift in either condition, and the two shifts are almost perfectly correlated. The shift is therefore carried by the linguistic pathway rather than by scene evidence. The high correlations show that similar examples receive similar score shifts; they do not indicate shared improvement.

Table D.9: Paired row-bootstrap estimates and percentile 95% confidence intervals for the detailed 3DR-FT diagnosis. We use 10,000 resamples with seed 42.

Estimator	Estimate	95% CI
R_{retain}	0.7079	[0.6770, 0.7377]
R^F_{target}	0.6529	[0.6208, 0.6844]
R^B_{target}	0.9223	[0.9010, 0.9428]
$\mathbb{E}[g_F]$	-12.3254	$[-12.6087, -12.0449]$
$\mathbb{E}[g_T]$	-11.9867	$[-12.2669, -11.7068]$
$\mathbb{E}[g_F - g_T]$	-0.3387	$[-0.3889, -0.2867]$
Pearson(g_F, g_T)	0.9840	[0.9816, 0.9861]
Spearman(g_F, g_T)	0.9761	[0.9718, 0.9795]

D.5.2 Answer-family and question-type decompositions

Table D.10 conditions both rates on full-input changes within each group. Left/right, front/behind, and up/down questions exhibit high persistence and high agreement in the resulting answer, so directional changes are almost entirely retained without the scene. Yes/no changes are few and persist at a lower rate, and counting changes persist least.

Table D.10: Shortcut Relocation decomposition by answer family and question type. n is the number of examples, N_F the number of full-input changes, and N_B the number of changes under both conditions. Rates are proportions; $\Delta g = \mathbb{E}[g_F - g_T]$.

(a) Answer family

Family	n	N_F	N_B	R_{retain}	R^F_{target}	Δg
Yes/no	153	32	20	0.6250	0.5625	-0.6758
Left/right	160	124	102	0.8226	0.8065	-0.3508
Front/behind	90	74	69	0.9324	0.9324	-0.0677
Up/down	10	10	10	1.0000	1.0000	0.1326
Counting	279	113	49	0.4336	0.4159	-0.9636
Other	793	520	368	0.7077	0.6269	-0.0881

(b) Question type

Type	n	N_F	N_B	R_{retain}	R^F_{target}	Δg
Distance	5	1	1	1.0000	1.0000	-1.2136
Direction	633	450	349	0.7756	0.7133	-0.2771
Counting	297	117	49	0.4188	0.4017	-0.9474
Existence	81	19	16	0.8421	0.7895	-0.5918
Other	469	286	203	0.7098	0.6503	0.0166

D.6 Parameter and inference efficiency

Table D.11 compares trainable parameters and measured end-to-end deployment cost for representative training- and inference-time methods.

Table D.11: Parameter and deployment costs on A800 80GB PCIe GPUs at batch size 1. Latencies are medians of three repetitions on matched examples.

Method	Trainable params (M) ↓	Decode paths ↓	Latency (ms/sample) ↓ Real-3DQA	Latency (ms/sample) ↓ 3D-POPE	Peak memory (MiB) ↓
3DR-FT (Ma et al., 2026)	345.43	**1**	776.40	422.93	21,050
3D-VCD (Ogunleye et al., 2026)	**0**	2	3,415.16	831.85	23,772
OIF	5.77	**1**	**549.64**	**150.93**	**20,504**

OIF updates 5.77M parameters, approximately 1.67% of the 345.43M parameters updated by 3DR-FT, and retains a single decoding path. It has the lowest measured end-to-end latency and peak allocated memory among the three methods. The latency comparison includes each method's native generation length: 3DR-FT generates approximately $1.65\times$ as many steps per example as OIF on Real-3DQA and $3.65\times$ as many on 3D-POPE. Its Real-3DQA latency per generated step is therefore lower than OIF's, 73.83 versus 86.38 ms, whereas 3D-VCD requires 594.75 ms. The reported latency measures end-to-end deployment cost, not kernel-level decoder speed.

E QUALITATIVE RESULTS AND INTERPRETATION BOUNDARIES

Figures E.1 and E.2 provide sample-level comparisons between OIF and 3DR-FT. These examples illustrate the form of successful corrections; aggregate performance is reported in Section 5. Each card shows the scene, the question, the two model predictions, and the ground truth. A green check marks the OIF prediction (Ours), a red cross marks the 3DR-FT prediction (Baseline), and a gray square marks the ground truth.

E.1 INTERPRETATION BOUNDARIES

The results support a claim about where newly learned corrective capacity is allocated, not a causal account of how a prediction was formed. Dependence on both L and X does not establish that the model used the scene for the intended reason. Likewise, the exact orthogonality result holds under the declared product reference $Q_L \otimes Q_X$ and need not transfer unchanged to the empirical paired distribution or to an arbitrary distribution shift. Finally, KL preservation limits unnecessary policy movement on non-risk rows but can also preserve mistakes made by the parent model.

OIF inherits several architectural limitations. Its global scene summary discards object identity, and its static residual is less expressive than token-dependent or multi-step interaction mechanisms. The method also relies on a prior-derived training cohort and on fixed feature scaling. Our experiments therefore establish controlled gains for short-answer indoor 3D question answering and object-existence judgments; they do not establish causal grounding or performance on open-ended dialogue, navigation, or manipulation.

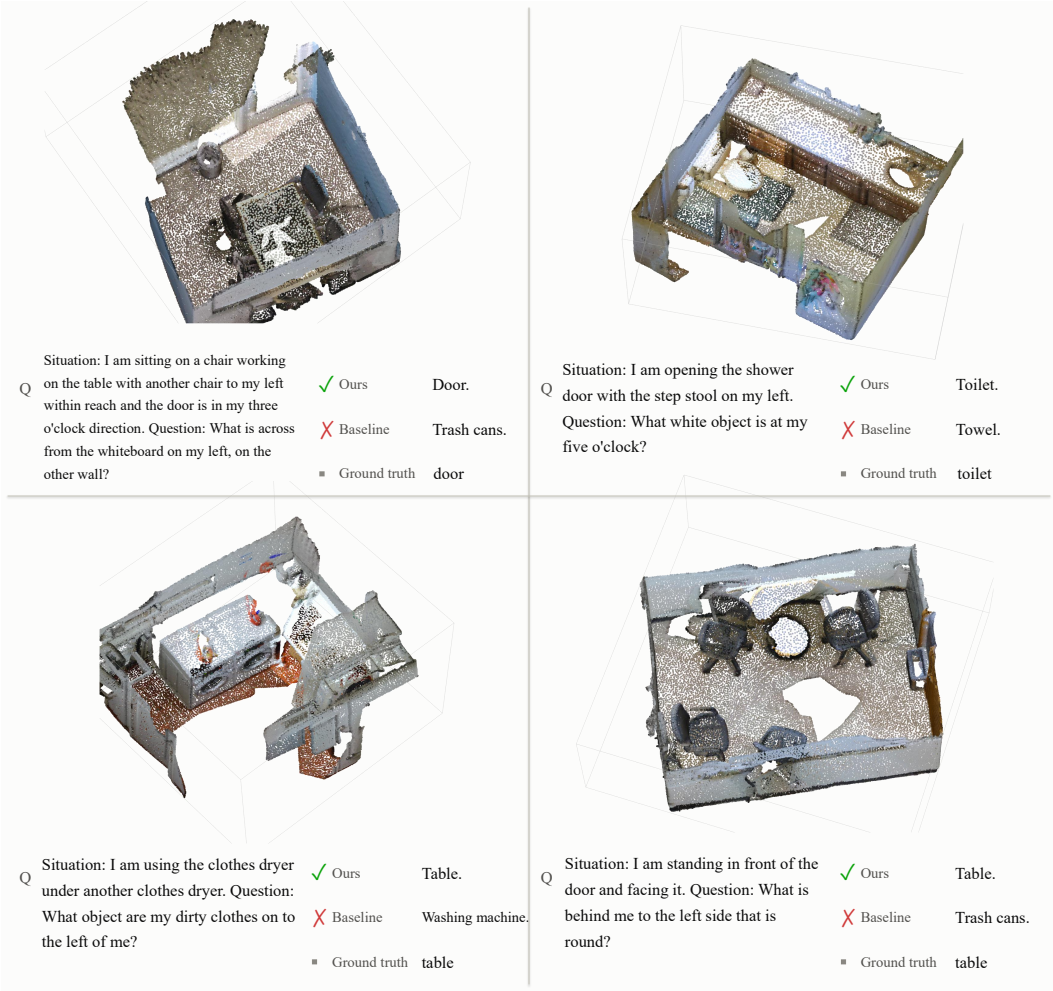


Figure E.1: Qualitative Real-3DQA comparisons. Four fixed cases show the scene, question, predictions, and ground truth together. OIF gives the ground-truth answer in each displayed case, whereas 3DR-FT selects a different answer. The cases are selected from the admissible pool described above and are not used as an additional aggregate evaluation.

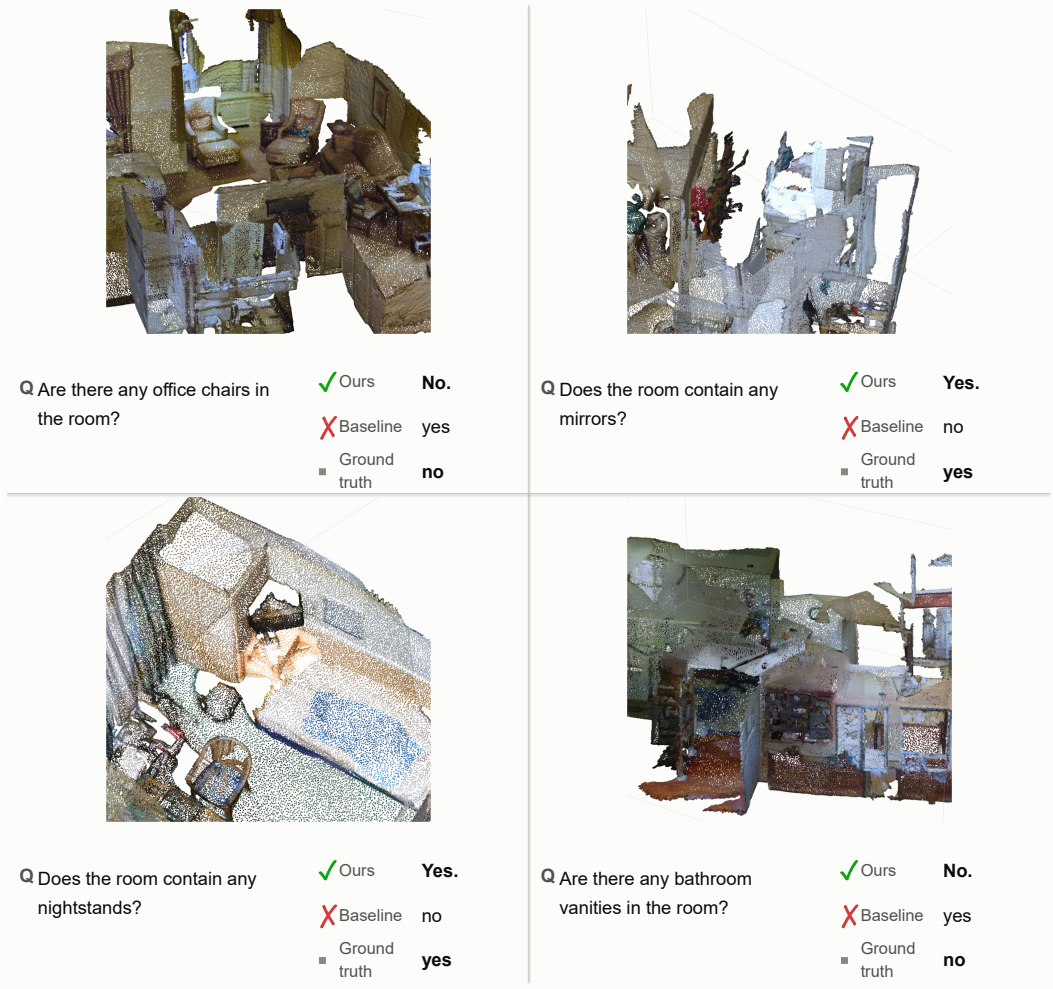


Figure E.2: Qualitative 3D-POPE comparisons. Each card pairs the same question and scene with the OIF and 3DR-FT predictions and the ground truth. The displayed cases were fixed before layout and illustrate the object-existence errors corrected by OIF.