WHAT SHOULD 3D-LMMS SEE? SCENE-STATE CURATION FOR 3D LARGE MULTIMODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress in unified 3D Large Multimodal Models (3D-LMMs) has benefited substantially from increasing training-side investment, including large-scale 3D vision-language supervision, stronger cross-modal alignment, and multi-stage training. These advances enable models to acquire rich perceptual evidence from complex 3D scenes. Yet this evidence must still be organized into a shared scene state before it can support downstream reasoning. We identify a systematic mismatch in this process: scene-state construction uses candidate-level perceptual cues to decide which evidence should be prioritized or treated as redundant, but these cues do not necessarily align with the physical structure that the scene state should preserve. In particular, saliency is systematically coupled with physical support, so large-support candidates tend to receive greater representation priority, while semantic similarity alone can cause spatially distinct instances to be treated as redundant, suppressing independent evidence. We therefore propose **P**hysics-**A**ware **3D** Scene **R**epresentation **C**uration (**PARC-3D**), a training-free framework for scene-state curation. PARC-3D calibrates the support-associated saliency advantage and models redundancy jointly through semantic and spatial criteria, yielding scene states that better preserve relevance, complementarity, and physical separability. Evaluations on referring segmentation, 3D question answering, and dense captioning show gains across all three task families while holding model parameters and scene-state size fixed, with the largest improvements on referring segmentation. These results highlight scene-state curation as a distinct representation axis complementary to data scaling, representation learning, and cross-modal optimization.

1 INTRODUCTION

Recent advances in unified 3D Large Multimodal Models (3D-LMMs) have increasingly relied on training-side investment in 3D vision-language supervision, high-quality annotations, cross-modal alignment, and multi-stage training (Hong et al., 2023; Chen et al., 2024a; Huang et al., 2024a; Deng et al., 2025). These advances enable models to acquire rich perceptual evidence from complex 3D environments. For a unified 3D-LMM, however, the resulting perceptual evidence must ultimately be organized into a shared scene state that supports diverse tasks such as referring segmentation, 3D question answering, and dense captioning (Chen et al., 2020; Azuma et al., 2022; Ma et al., 2023; Chen et al., 2021). The evidence available to the language model is thus shaped by two successive processes: **what the upstream model learns from the scene, and how that learned evidence is organized before reasoning begins**. Prior work has focused primarily on the former (Hong et al., 2023; Chen et al., 2024a; Huang et al., 2024a; Deng et al., 2025). How this learned evidence is organized into a high-quality shared scene state therefore constitutes a complementary representation problem that directly shapes downstream reasoning.

This organization problem is non-trivial because 3D evidence is heterogeneous in scale, semantics, and spatial location (Dai et al., 2017; Song et al., 2015; Chang et al., 2017). A useful scene state must therefore preserve fine-grained evidence and physical instance separability while suppressing genuine redundancy, requiring two coupled decisions: **which evidence receives representation priority and which evidence is considered redundant** (Deng et al., 2025; Zhi et al., 2025). In representative constructions, these decisions are instantiated from candidate-level perceptual cues: upstream relevance scores guide representation priority, while feature similarity provides a natural

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.

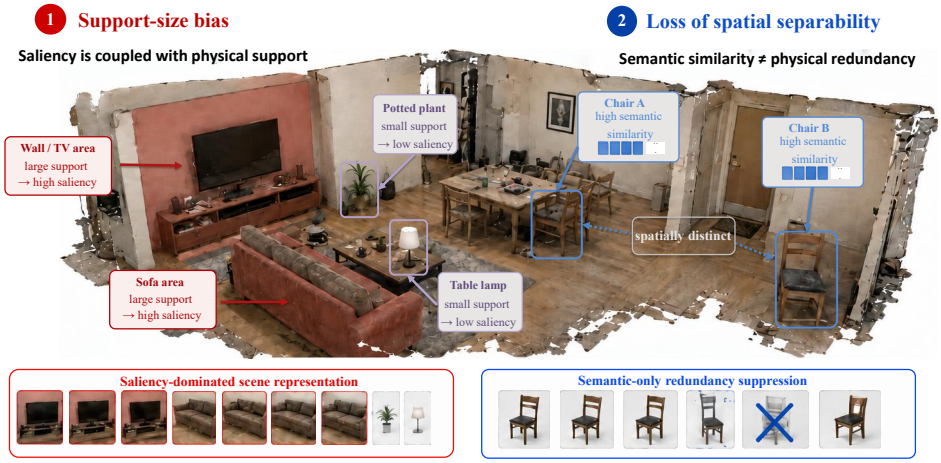


Figure 1: **Motivation: two structural mismatches in 3D scene-state construction. (1) Support-size bias:** saliency is systematically coupled with physical support, so large-support candidates tend to receive higher representation priority while small but informative evidence can be under-represented. **(2) Loss of spatial separability:** semantically similar yet spatially separated instances can be treated as redundant, suppressing independent evidence. Together, they expose a mismatch between candidate-level perceptual cues and scene-level 3D representation requirements.

criterion for redundancy when redundancy suppression is applied. Figure 1 illustrates how directly using these candidate-level cues as scene-level organization criteria can induce two corresponding distortions in 3D scenes. First, saliency is systematically coupled with physical support: candidates covering more geometric primitives tend to receive higher saliency and thus higher priority under saliency-driven ranking. Our candidate-level audit shows that large-support candidates exhibit roughly three times the average saliency of small-support candidates, and large-support evidence is correspondingly over-represented in the shared scene state; we term this **support-size bias**. Second, feature similarity captures semantic proximity, whereas same-category objects at different locations remain physically distinct and can provide independent evidence for localization, counting, and relational reasoning (Achlioptas et al., 2020; Luo et al., 2022; Roh et al., 2022; Chen et al., 2022b; Zhang et al., 2023). As shown on the right of Figure 1, two chairs can exhibit highly similar semantic features despite being spatially separated; treating them as redundant can suppress one instance and discard its independent evidence. We term this **loss of spatial separability**. Together, these effects expose a systematic mismatch between candidate-level cues and scene-level 3D representation requirements: **representation priority can inherit support-associated bias, while redundancy criteria can fail to preserve physical separability**.

These observations motivate a scene-state curation principle that preserves learned relevance while promoting evidence complementarity and physical separability. We therefore view scene-state curation as a distinct representation problem, complementary to data scaling, representation learning, and cross-modal optimization. To this end, we introduce **Physics-Aware 3D Scene Representation Curation (PARC-3D)**, a training-free framework that reorganizes the perceptual evidence from a pretrained 3D-LMM. PARC-3D addresses both mismatches through a unified scene-level curation principle. First, **Volume-Aware Saliency Calibration (VSC)** calibrates the support-associated saliency advantage while preserving upstream learned relevance. Second, Complexity-Aware Spatio-Semantic Routing (CSR) models redundancy jointly in semantic and physical space, suppressing locally repetitive evidence while preserving semantically similar but physically distinct instances. CSR further adapts its routing extent to the spectral complexity of the scene representation. Residual-utility construction then integrates calibrated relevance with structured redundancy to form the final scene state. PARC-3D requires no additional supervision, parameter updates, or task-specific modifications to the pretrained 3D-LMM.

We evaluate PARC-3D across referring segmentation, 3D question answering, and dense captioning (Chen et al., 2020; Zhang et al., 2023; Azuma et al., 2022; Ma et al., 2023; Chen et al., 2021). PARC-3D raises ScanRefer mIoU from 43.3 to 46.2 and Multi3DRefer mIoU from 42.7 to 52.1, while improving ScanQA CIDEr from 92.6 to 96.3 and both exact-match metrics on SQA3D. It also

improves multiple generation metrics on Scan2Cap. Together, these results show that **changing how learned perceptual evidence is organized, while leaving the pretrained model and downstream task design unchanged, can improve performance across all three 3D scene-understanding task families**.

Our contributions are threefold:

- We formulate **scene-state curation as a distinct representation problem** in unified 3D-LMMs and identify two structural mismatches between candidate-level cues and scene-level requirements: support-associated bias in representation priority and insufficient preservation of physical separability.
- We propose **PARC-3D**, a training-free framework for physical-structure-aware scene-state curation that combines Volume-Aware Saliency Calibration with Complexity-Aware Spatio-Semantic Routing to preserve relevance, complementarity, and physical separability.
- We demonstrate **cross-task gains in referring segmentation, 3D question answering, and dense captioning** under a fixed pretrained model and scene-state size.

2 Related Work

3D vision-language scene understanding spans language-guided grounding and segmentation (Chen et al., 2020; Achlioptas et al., 2020; Luo et al., 2022; Wu et al., 2023; Zhang et al., 2023), dense captioning (Chen et al., 2021; 2022a; Cai et al., 2022; Chen et al., 2023b), and question answering and situated reasoning (Azuma et al., 2022; Ma et al., 2023). Although these tasks differ in output form, unified 3D-LMMs must support them from a shared 3D scene state: grounding depends on fine-grained discriminative evidence, captioning on target-centered context, and reasoning on broader relational and geometric information. Most prior work improves these capabilities through stronger task modeling, 3D-text alignment, or unified training (Zhu et al., 2023; Chen et al., 2023a; Hong et al., 2023; Chen et al., 2024a; Deng et al., 2025), while recent scene-centric methods have begun to make the organization of 3D evidence itself an explicit design choice (Halacheva et al., 2025). Several recent unified 3D-LMMs explicitly design how visual evidence is organized before reasoning. 3D-LLaVA (Deng et al., 2025) selects superpoint candidates through learned objectness, LSceneLLM (Zhi et al., 2025) retrieves query-relevant fine-grained regions, PerLA (Mei et al., 2025) combines local detail with global context, and Chat-Scene++ (Huang et al., 2026) enriches object-centric representations with relational and global information. PARC-3D addresses a complementary issue: whether candidate-level cues for representation priority and redundancy remain structurally appropriate when forming a shared 3D scene state, specifically calibrating support-associated priority while incorporating physical separability into redundancy estimation. A related line instead reduces visual tokens for computational efficiency through attention- or importance-based pruning (Huang et al., 2025; Zhang et al., 2025; Lin et al., 2026), token compression or modality routing (Xiong et al., 2025; Mei et al., 2026; Zhang et al., 2026), or geometric diversity and coverage (Li et al., 2026; Pan et al., 2026; Ling et al., 2026; Qu et al., 2026; Bui et al., 2026; Alvar et al., 2025; Shang et al., 2025). In contrast, PARC-3D operates on perception-front-end candidates before the language model and keeps the vanilla scene-state cardinality K fixed; its goal is to improve the structural quality of the retained scene state rather than reduce inference cost.

3 Method

3.1 From Candidate Evidence to Scene State

To process complex environments, unified 3D-LMMs decompose scene inputs into visual primitives, typically through region-based sampling (Chen et al., 2024a), object-centric abstraction (Huang et al., 2024a), or superpoint clustering (Deng et al., 2025). Abstracting away these architectural variations, we formalize the output of any upstream visual encoder as a set of N candidate visual units, $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^{N}$, each aggregating local geometric and semantic evidence:

$$\mathbf{x}_i = (\mathbf{z}_i, \mathbf{p}_i, n_i, s_i), \tag{1}$$

where $\mathbf{z}_i \in \mathbb{R}^d$ is the semantic feature, $\mathbf{p}_i \in \mathbb{R}^3$ the physical center, $n_i \in \mathbb{N}_+$ the support size (the number of underlying geometric primitives, such as points or voxels, covered by the unit), and $s_i \in \mathbb{R}_{>0}$ the raw upstream saliency or objectness score (scores are positive after the upstream

normalization, so logarithms below are well defined). These candidates constitute the perceptual evidence acquired by the upstream model; scene-state curation determines which part of this learned evidence is exposed jointly to downstream reasoning.

Under a fixed-size visual interface, the downstream LLM consumes a scene state containing K ($K \ll N$) candidate units. Scene-state curation therefore determines the index set

$$\mathcal{I} \subseteq \{1, \dots, N\}, \qquad |\mathcal{I}| = K, \tag{2}$$

which induces the scene state $\mathcal{S}(\mathcal{I}) = \{\mathbf{x}_i \mid i \in \mathcal{I}\}$.

Constructing a fixed-size scene state involves two coupled decisions: **which candidate evidence should receive representation priority, and which evidence should be considered redundant**. Candidate-level perceptual cues provide natural signals for both decisions. Learned relevance, objectness, or saliency scores guide candidate priority in existing 3D-LMM interfaces (Deng et al., 2025; Zhi et al., 2025), while feature similarity or related proximity criteria provide a common basis for identifying repeated evidence in redundancy-aware visual selection (Bolya et al., 2023; Shang et al., 2025; Alvar et al., 2025; Li et al., 2026). These cues are informative at the candidate level, but using them directly as scene-level organization criteria can conflict with the physical structure of 3D scenes. First, raw relevance can carry a systematic advantage associated with physical support: candidates covering more geometric primitives tend to receive higher saliency scores, so ranking candidates monotonically by s_i can propagate this support-associated advantage into representation priority. Second, semantic similarity does not imply physical redundancy: two candidates of the same category may be spatially distinct entities, and treating one as redundant on semantic similarity alone can discard independent scene evidence.

PARC-3D therefore treats representation priority and redundancy as two coupled aspects of scene-state curation, jointly accounting for physical support and physical separability. We formulate curation through two scene-level quantities: a calibrated retention utility u_i, which preserves learned relevance while accounting for physical support, and a structured redundancy r_{ij}, which measures whether two candidates provide overlapping evidence under both semantic and physical criteria. Over feasible sets satisfying Equation 2, the curated state should favor high-utility evidence while penalizing redundancy among retained candidates:

$$\max_{\mathcal{I}} F(\mathcal{I}) = \sum_{i \in \mathcal{I}} u_i - \eta \sum_{\substack{i,j \in \mathcal{I} \\ i<j}} r_{ij}, \tag{3}$$

where $\eta > 0$ governs the trade-off.

As illustrated in Figure 2, PARC-3D implements this curation principle through two structural corrections followed by one integration procedure. *Volume-Aware Saliency Calibration* (VSC) computes u_i and thereby calibrates which evidence receives representation priority. *Complexity-Aware Spatio-Semantic Routing* (CSR) computes r_{ij} by distinguishing semantic similarity from physical redundancy, and thereby determines which high-relevance evidence is genuinely redundant at the scene level. *Residual-utility scene-state construction* then integrates the two corrected signals into the final fixed-size scene state.

3.2 Volume-Aware Saliency Calibration (VSC)

Upstream saliency s_i remains the primary relevance signal. VSC preserves the learned relevance signal while calibrating the component of candidate priority systematically associated with physical support, characterized by the pool-normalized support ratio

$$v_i = \frac{n_i}{\sum_{j=1}^{N} n_j}, \tag{4}$$

where $v_i \in (0, 1]$ is the relative support ratio of candidate i, normalized within the candidate pool so that support magnitude is comparable across scenes. Here n_i counts the points, voxels, or proposal points covered by a candidate, so "volume" refers to this relative physical-support proxy rather than to an exact Euclidean volume. VSC then converts the raw saliency s_i into a calibrated retention utility

$$u_i = s_i \exp(-\gamma_{\text{vol}} v_i), \tag{5}$$

where $\gamma_{\text{vol}} > 0$ controls the calibration strength. The multiplicative form preserves s_i explicitly while continuously attenuating priority according to relative support. Candidates with strong upstream

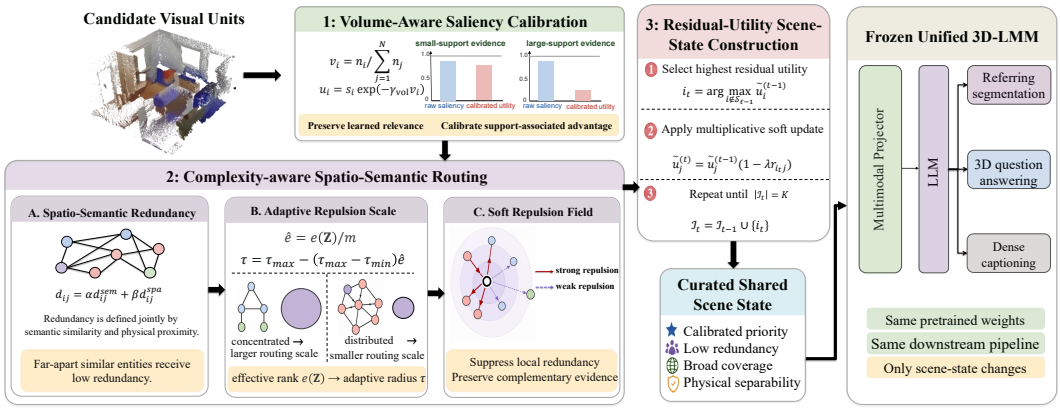


Figure 2: **PARC-3D: two structural corrections for scene-state curation.** PARC-3D reorganizes learned 3D evidence by calibrating support-associated representation priority with VSC and modeling redundancy jointly in semantic and physical space with CSR. Residual-utility construction integrates both signals into the shared scene state. The pretrained model and downstream pipeline remain unchanged; only scene-state composition is modified.

relevance can therefore remain competitive even when they have large support, and candidates with the same support ratio preserve exactly the ordering induced by s_i. VSC thus calibrates the support-associated priority advantage without replacing the upstream relevance signal.

3.3 Complexity-Aware Spatio-Semantic Routing (CSR)

VSC corrects representation priority, but it does not decide whether two high-utility candidates provide overlapping evidence. In a 3D scene, semantic similarity alone is insufficient for this decision: two candidates may encode the same category while corresponding to spatially distinct instances, and treating either as redundant can discard independent evidence. CSR therefore defines redundancy jointly in semantic and physical space. Let d_{ij}^{sem} denote the semantic cosine distance between candidates i and j,

$$d_{ij}^{\mathrm{sem}} = 1 - \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}, \tag{6}$$

and d_{ij}^{spa} the scale-normalized spatial distance,

$$d_{ij}^{\mathrm{spa}} = \frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2}{\rho}, \tag{7}$$

where ρ normalizes the Euclidean distance against the global scene extent for scale invariance,

$$\rho = \|\mathbf{b}_{\max} - \mathbf{b}_{\min}\|_2 + \varepsilon, \tag{8}$$

with $\mathbf{b}_{\max}$ and $\mathbf{b}_{\min}$ the extreme coordinates of the scene bounding box and $\varepsilon > 0$ a stability constant. The two criteria are combined into a joint spatio-semantic distance,

$$d_{ij} = \alpha_{\mathrm{sem}} d_{ij}^{\mathrm{sem}} + \beta_{\mathrm{spa}} d_{ij}^{\mathrm{spa}}, \tag{9}$$

where $\alpha_{\mathrm{sem}}, \beta_{\mathrm{spa}} \geq 0$ balance the two criteria, so that a small joint distance is assigned only to pairs that are both semantically akin and physically adjacent. The structured redundancy follows from a Gaussian kernel,

$$r_{ij} = \exp\left(-\frac{d_{ij}^2}{2\tau^2}\right), \tag{10}$$

where $\tau > 0$ is the routing scale over which redundancy acts. This joint definition yields three behaviors. Semantically similar and spatially close candidates receive a high r_{ij}, which identifies genuine local repetition. Semantically similar but spatially distant candidates keep a large d_{ij} and a reduced r_{ij}, so two physical instances of the same category both remain eligible for retention. Spatially close but semantically different candidates likewise receive a reduced r_{ij}, so spatial density alone does not suppress distinct evidence.

The extent over which redundancy should act depends on the representation being curated. We use the normalized effective rank of the candidate feature matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]^\top \in \mathbb{R}^{N \times d}$ as a representation-level complexity signal. With singular values $\{\sigma_t\}_{t=1}^{m}$, $m = \min(N, d)$, the normalized spectral distribution is

$$q_t = \frac{\sigma_t}{\sum_{\ell=1}^{m} \sigma_\ell}, \tag{11}$$

and, following the effective rank of Roy & Vetterli (2007), the normalized effective rank is

$$e(\mathbf{Z}) = \exp\left(-\sum_{t=1}^{m} q_t \log q_t\right), \quad \hat{e} = \frac{e(\mathbf{Z})}{m}, \tag{12}$$

with the convention $0 \log 0 = 0$ and $\hat{e} \in [1/m, 1]$. A low value indicates features concentrated along few dominant spectral directions; a high value indicates a more distributed representation. The repulsion scale is set as

$$\tau = \tau_{\max} - (\tau_{\max} - \tau_{\min})\hat{e}, \tag{13}$$

where $\tau_{\max} > \tau_{\min} > 0$. For a distributed representation, τ shrinks so that redundancy is assigned only to closely overlapping pairs and fine-grained distinctions are preserved; for a concentrated representation, τ expands so that repeated evidence is suppressed over a wider extent.

3.4 Residual-Utility Scene-State Construction

The integration procedure combines calibrated priority and structured redundancy into a shared scene state through residual-utility maximization, a greedy multiplicative surrogate of Equation 3 (Appendix A). Starting from $\mathcal{I}_0 = \varnothing$ and $\tilde{u}_i^{(0)} = u_i$, each iteration t adds the candidate with the highest residual utility,

$$i_t = \arg\max_{i \notin \mathcal{I}_{t-1}} \tilde{u}_i^{(t-1)}, \qquad \mathcal{I}_t = \mathcal{I}_{t-1} \cup \{i_t\}, \tag{14}$$

and attenuates the residual utilities of the unselected candidates j according to their redundancy with the newly selected candidate i_t,

$$\tilde{u}_j^{(t)} = \tilde{u}_j^{(t-1)}\big(1 - \lambda r_{i_t j}\big), \qquad j \notin \mathcal{I}_t, \tag{15}$$

where $\lambda \in (0, 1)$ is the repulsion strength; since $r_{ij} \in (0, 1]$, every factor $1 - \lambda r_{ij}$ lies in $[1 - \lambda, 1)$, so residual utilities remain strictly positive. The unrolled product form and its log-space interpretation are provided in Appendix A. Each selection thus attenuates the remaining candidates only through their redundancy with already-selected evidence: a high-utility candidate complementary to the selected evidence keeps most of its residual utility and remains competitive for subsequent selection; one that repeats selected local evidence decays rapidly; and a semantically similar but physically separated candidate experiences little redundancy-induced attenuation because its r_{ij} is low. The process terminates when $|\mathcal{I}_t| = K$, yielding the final curated state

$$\mathcal{I}^\star = \mathcal{I}_K, \qquad \mathcal{S}^\star = \{\mathbf{x}_i \mid i \in \mathcal{I}^\star\}, \tag{16}$$

which replaces the original scene state. PARC-3D changes only the composition of the scene state; candidate features, pretrained parameters, multimodal interface architecture, prompts, and downstream decoders remain unchanged (full procedure in Appendix A).

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on five standard 3D vision-language benchmarks: ScanRefer (Chen et al., 2020) and Multi3DRefer (Zhang et al., 2023) for referring segmentation, ScanQA (Azuma et al., 2022) and SQA3D (Ma et al., 2023) for 3D question answering, and Scan2Cap (Chen et al., 2021) for dense captioning. Together they test whether a unified 3D-LMM can preserve fine-grained target evidence, parse spatial relations, and describe scenes from a shared 3D input state (details in Appendix B).

Metrics. We report mIoU for referring segmentation; CIDEr, BLEU-4, METEOR, and ROUGE-L on ScanQA; Exact Match (EM) and EM-Refined on SQA3D, following their standard protocols; and

Table 1: **Comparison on five 3D vision-language benchmarks.** 'Mod.': input modality, PC (point clouds) or I (images). Multi3D: Multi3DRefer. C, B-4, M, R, EM, EM-R denote CIDEr, BLEU-4, METEOR, ROUGE-L, Exact Match, and EM-Refined. Baseline results are cited from their publications. Best results are in **bold** and second-best underlined. The ScanQA results of LEO (Huang et al., 2024b) are shown in gray and excluded from the ranking, as LEO uses ground-truth object associations.

Method	Mod.	ScanRefer	Multi3D	ScanQA				SQA3D		Scan2Cap			
		mIoU↑	mIoU↑	C↑	B-4↑	M↑	R↑	EM↑	EM-R↑	C@0.5↑	B-4@0.5↑	M@0.5↑	R@0.5↑
3D-VLP (Jin et al., 2023)	PC	–	–	67.0	11.2	13.5	34.5	–	–	54.9	32.3	24.8	51.5
3D-VisTA (Zhu et al., 2023)	PC	–	–	69.6	10.4	13.9	35.7	48.5	–	61.6	34.1	26.8	55.0
3D-LLM (Hong et al., 2023)	PC+I	–	–	69.4	12.0	14.5	35.7	–	–	–	–	–	–
LEO (Huang et al., 2024b)	PC+I	–	–	101.4	13.2	20.0	49.2	50.0	52.4	72.4	**38.2**	27.9	**58.1**
Scene-LLM (Fu et al., 2025)	PC+I	–	–	80.0	11.7	15.8	35.9	53.6	–	–	–	–	–
Chat-Scene (Huang et al., 2024a)	PC+I	–	–	87.7	14.3	18.0	41.6	54.6	57.5	77.2	36.4	**28.0**	**58.1**
LL3DA (Chen et al., 2024a)	PC	–	–	76.8	13.5	15.9	37.3	–	–	65.2	36.8	26.0	55.1
Grounded 3D-LLM (Chen et al., 2024b)	PC	–	–	72.7	13.4	–	–	–	–	70.6	35.5	–	–
LSceneLLM (Zhi et al., 2025)	PC	–	–	88.2	–	18.0	40.8	–	–	–	–	–	–
3D-LLaVA (Deng et al., 2025)	PC	43.3	42.7	92.6	**17.1**	18.4	43.1	54.5	56.6	**78.8**	36.9	27.1	57.7
PARC-3D (ours)	PC	**46.2**	**52.1**	**96.3**	16.5	**19.0**	**45.2**	**55.7**	**58.7**	**78.8**	37.6	27.4	**58.1**

CIDEr, BLEU-4, METEOR, and ROUGE-L at a 0.5 IoU threshold (denoted @0.5) on Scan2Cap, following 3D-LLaVA (Deng et al., 2025).

Implementation Details and Control Protocol. PARC-3D operates at the visual interface of the 3D-LMM, after the perception front-end and before the LLM. It reads the raw candidate attributes (semantic features, 3D coordinates, support sizes, and saliency scores) and outputs a curated scene state of the same size K as the vanilla model. Our main platform is the official pretrained 3D-LLaVA (Deng et al., 2025). All comparisons between vanilla 3D-LLaVA and PARC-3D use the same pretrained checkpoint, candidate pool, scene-state size, prompts, multimodal projector, LLM, and downstream task decoder; the only changed variable is how candidate evidence is organized into the scene state. PARC-3D is thus a purely inference-time intervention with no learnable parameters and no additional training. The main results use one global hyperparameter setting across all datasets and tasks; only the ablation and sensitivity studies vary individual components or parameters. The values and the instantiation on a second interface (Section 4.5) are given in Appendix C.

4.2 Overall Results

Table 1 reports results on five benchmarks. Since the model is kept fixed, we discuss the results by task family, as the tasks draw on the scene state in different ways.

Referring segmentation requires fine-grained target evidence to be retained and same-category instances to remain separable, and here the changes are largest: mIoU rises from 43.3 to 46.2 on ScanRefer and from 42.7 to 52.1 on Multi3DRefer. Multi3DRefer expressions name zero, one, or several same-category instances, which joint routing keeps as distinct units. 3D question answering draws on scene-level and situated evidence rather than a single target: on ScanQA, CIDEr improves from 92.6 to 96.3, METEOR from 18.4 to 19.0, and ROUGE-L from 43.1 to 45.2, while BLEU-4 decreases from 17.1 to 16.5; on SQA3D, EM improves from 54.5 to 55.7 and EM-R from 56.6 to 58.7. Dense captioning on Scan2Cap is anchored to detected objects and their local context and shows the smallest changes: CIDEr@0.5 remains at 78.8, while BLEU-4, METEOR, and ROUGE-L improve from 36.9 to 37.6, 27.1 to 27.4, and 57.7 to 58.1. Across all three families, PARC-3D improves or matches the principal task metrics, with metric-specific variation in language metrics; the gains exceed the across-subset spread of PARC-3D scores (Appendix E.5).

The remaining rows place these results in context: many systems rely on auxiliary 2D images, larger pretraining pipelines, or architecture-specific designs, whereas PARC-3D, operating on point clouds only and without any parameter update, attains the best or tied-best listed result on nine of the twelve metrics. Scene-state organization thus complements training-side advances even with the pretrained model held fixed.

4.3 Mechanism Analysis: Correcting the Diagnosed Scene-State Distortions

We next ask whether the gains arise from correcting the two distortions diagnosed in Figure 1, following each from the candidate pool to the curated state and then to task performance (full statistics in Appendices D and E).

Table 2: **Scene-state organization.** Baselines under matched controls.

Method	ScanRef. mIoU	ScanQA C	Scan2Cap B-4
Saliency	43.3	92.6	36.9
Random	40.7	92.2	36.4
Spatial FPS	42.0	92.7	36.8
Semantic NMS	44.9	95.0	37.2
MMR	45.4	95.6	37.3
DPP	45.7	95.8	37.3
PARC-3D	**46.2**	**96.3**	**37.6**

Table 3: **Mechanism diagnostics.** ScanRefer target-size mIoU and scene-state structure.

Diagnostic	Vanilla	PARC-3D	Δ
Small-target mIoU	35.0	**38.2**	+3.2
Medium-target mIoU	38.0	**41.1**	+3.1
Large-target mIoU	57.3	**59.2**	+1.9
NN-Spatial	0.0328	**0.0385**	+0.0057
Pair-Spatial	0.3247	**0.3320**	+0.0073
Pair-Semantic	0.3473	**0.3564**	+0.0091
Pair-Joint	0.6720	**0.6884**	+0.0164

Support-size bias. Auditing the raw candidates of the frozen backbone on Scan2Cap, ScanQA, and ScanRefer (Appendix D), large-support candidates receive about 3.1–3.4$\times$ the mean raw saliency of small-support candidates, are about twice as likely to fall in the top saliency decile, and are over-represented by about 1.5$\times$ after top-K construction relative to their share of the pool; saliency-based construction therefore inherits a support-size advantage. The curated PARC-3D state shows the moderate rebalancing VSC targets: at the same scene-state size, the share of small-support units rises from 23.97% to 26.60% (+2.63 pp) while that of large-support units falls from 50.35% to 48.81% (-1.54 pp), so under-represented small-support evidence gains share while high-relevance large-support evidence is preserved (Appendix E.1). The target-size rows of Table 3 stratify ScanRefer targets by size: PARC-3D improves all three groups, with the largest gains on small (+3.2 mIoU) and medium (+3.1) targets and a smaller gain on large targets (+1.9). The larger gains on small and medium targets are consistent with the hypothesis that improved retention of fine-grained evidence particularly benefits targets with smaller spatial support; these target-size tertiles are defined on ScanRefer targets and are distinct from the candidate support tertiles of the audit.

Loss of spatial separability. Table 2 compares organization assumptions under the same candidate pool, backbone, scene-state size, and decoders. The rows form a progression of assumptions rather than an isolation of individual factors. *Random* and *Spatial FPS* replace learned relevance with no structure or with spatial coverage alone, and both fall below *Saliency* on referring segmentation and dense captioning, confirming that learned relevance must be preserved rather than replaced. *Semantic NMS* adds semantic redundancy suppression on top of learned relevance, and *MMR* (Carbonell & Goldstein, 1998) and *DPP* (Kulesza & Taskar, 2012) are generic selectors that already balance relevance against diversity. The same ordering appears on both ScanRefer mIoU and ScanQA CIDEr: mIoU rises from 43.3 (relevance only) to 44.9 (semantic redundancy) to 45.4–45.7 (generic relevance–diversity) to 46.2 (PARC-3D), and CIDEr from 92.6 to 95.0 to 95.6–95.8 to 96.3. Modeling redundancy helps, and MMR and DPP remain below PARC-3D on all three metrics, so the two structural corrections contribute beyond generic relevance–diversity balancing; Section 4.4 isolates each of them. The structural rows of Table 3 show the predicted change in the curated state: relative to vanilla construction, the mean nearest-neighbor spatial distance among retained units increases from 0.0328 to 0.0385 and the mean pairwise joint spatio-semantic distance from 0.6720 to 0.6884, so the retained evidence is less locally concentrated and less redundant, consistent with suppressing nearby repeated evidence while keeping spatially distinct instances.

4.4 Ablation and Parameter Sensitivity

Table 4 compares the frozen vanilla construction with four controlled variants (defined in Appendix C). Removing either VSC or CSR lowers both ScanQA metrics relative to the full formulation. Single-criterion routing exposes different trade-offs: among the curated variants, the full formulation is best or tied-best on three of four metrics, while semantic-only routing is 0.1 higher on Scan2Cap CIDEr. The results support complementary roles for priority calibration and joint spatio-semantic routing without implying uniform gains on every language metric.

Appendix Table E.2 shows that no tested fixed τ recovers the adaptive result on ScanRefer or ScanQA; Appendix E.4 relates the effective-rank signal to scene statistics.

Appendix Table E.1 varies the spatial weight β_{spa} and the calibration strength γ_{vol}. Across the tested ranges, ScanQA CIDEr stays within 95.5–96.3 (vanilla: 92.6) and Scan2Cap BLEU-4@0.5 within 37.2–37.6 (vanilla: 36.9), while Scan2Cap CIDEr@0.5 varies within 78.4–79.0 around the vanilla 78.8, so the gains are not confined to a narrow choice of either parameter.

Table 4: **Component ablation.** Vanilla is the frozen 3D-LLaVA construction; best values per metric are in **bold**.

Variant	ScanQA		Scan2Cap	
	C	B-4	C@0.5	B-4@0.5
Vanilla	92.6	**17.1**	78.8	36.9
w/o VSC	95.8	16.1	78.3	37.2
w/o CSR	96.0	16.2	78.6	**37.6**
Semantic-only	96.2	16.4	**78.9**	37.2
Spatial-only	95.8	16.2	78.8	37.2
Full	**96.3**	16.5	78.8	**37.6**

Table 5: **Cross-interface transfer at matched $K = 100$.**

Metric	Chat-Scene		Δ
	Base	+ PARC	
ScanRef. Acc@0.25	52.0	**52.9**	+0.9
ScanRef. Acc@0.5	47.1	**47.8**	+0.7
Multi3D F1@0.25	56.0	**56.3**	+0.3
Multi3D F1@0.5	51.2	**52.0**	+0.8
Scan2Cap C@0.5	71.4	**72.3**	+0.9
Scan2Cap B-4@0.5	34.4	**34.9**	+0.5

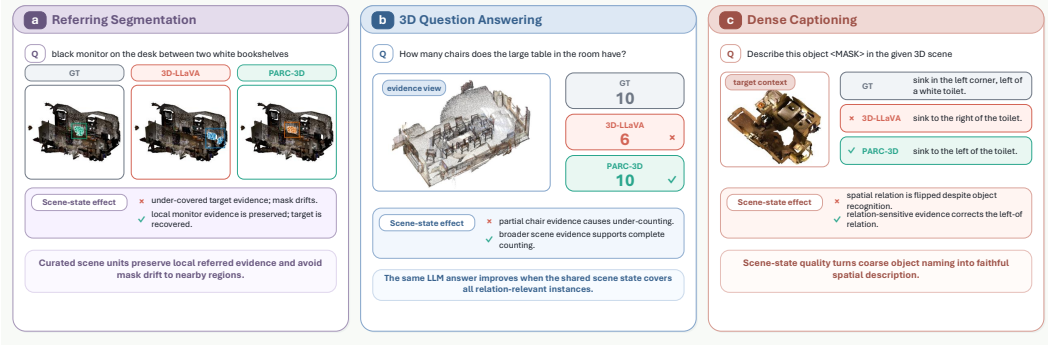


Figure 3: **Qualitative results across downstream tasks.** With the same model and scene-state size, the curated scene state retains fine-grained target evidence (segmentation), multiple relevant instances (question answering), and relation-anchor evidence around the target (dense captioning).

4.5 Transfer Across Scene-State Interfaces

We also test the same formulation on the object-centric Chat-Scene interface (Huang et al., 2024a), where candidates are object proposals (Appendix C); at the matched $K = 100$, all hyperparameters are inherited without retraining or interface-specific tuning (Table 5; protocol in Appendix B).

PARC-3D improves all six metrics, so the principle transfers across granularities; smaller margins may reflect Chat-Scene's far fewer, already deduplicated instance proposals per scene.

4.6 Qualitative Results

Figure F.1 in Appendix F visualizes the retained units in representative scenes. Vanilla construction concentrates on large surfaces and high-response clusters, leaving some small-object evidence uncovered; PARC-3D distributes the retained units more evenly, so small objects and their context are represented within the same scene-state size, the visual counterpart of the statistics in Section 4.3. Figure 3 shows the effect on the three tasks: retained fine-grained target evidence allows small or occluded targets to be segmented in clutter; retaining multiple relevant instances supports counting and questions about non-dominant objects; and retained relation-anchor evidence supports captions that reference neighboring objects correctly, all from the same curated state without task-specific changes. Additional examples and failure cases are given in Appendix F.

5 Conclusion

This work studies how learned 3D perceptual evidence is organized into the shared scene state of unified 3D-LMMs. We identify two structural mismatches between candidate-level cues and scene-level representation requirements: support-associated saliency can bias representation priority, while semantic similarity alone can obscure physical separability. PARC-3D addresses these mismatches by calibrating the support-associated saliency advantage and modeling redundancy jointly in semantic and physical space, without retraining the underlying 3D-LMM. Across three task families and the superpoint-centric and object-centric interfaces of 3D-LLaVA and Chat-Scene, reorganizing already learned evidence improves downstream performance while holding the pretrained model and scene-state size fixed within each comparison. Scene-state curation therefore complements representation learning by determining how learned perceptual evidence is exposed to multimodal reasoning.

Reproducibility Statement

PARC-3D is training-free and adds no learnable parameters, so reproducing it requires no training: it needs the released 3D-LLaVA and Chat-Scene checkpoints, the public benchmark data, and the curation procedure specified in Sections 3.1–3.4. Appendix A gives code-faithful algorithmic steps with equation references, Appendix B lists the benchmarks and evaluation protocols, and Appendix C reports the main configuration (α_{sem}, β_{spa}, γ_{vol}, λ, $\tau_{\min}$, $\tau_{\max}$), the candidate instantiation on the Chat-Scene interface, the organization baselines of Table 2, the scene-state size K, the hardware used, and the per-scene curation cost; Appendix E.5 reports the variation of the main results across evaluation-scene subsets. The main benchmark and transfer results use one fixed global PARC-3D configuration across datasets and tasks; only the explicitly labeled ablation, sensitivity, and fixed-τ experiments vary individual components or parameters. Code will be released upon acceptance.

AI Use Statement

In this work, we used generative AI tools for writing assistance, including language polishing and editing of the manuscript text. We have not used generative AI tools for research ideation, experimental design, or the generation of experimental results. All AI-assisted text was reviewed and revised by the authors, and all claims, equations, and reported numbers were verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Ethics Statement

This work uses only publicly available 3D indoor-scene benchmarks derived from ScanNet under their released terms of use, involves no human-subject data collection, and introduces no new data. PARC-3D reorganizes evidence already produced by pretrained models and adds no training. Potential applications, failure modes, and the need for human oversight in embodied or AR deployments are discussed in Appendix G.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In *European Conference on Computer Vision*, pp. 422–440, 2020.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401, 2025.

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.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token Merging: Your ViT But Faster. In *International Conference on Learning Representations*, 2023.

Nhat-Tan Bui, Varshini Elangovan, Arun Reddy Anugu, Sreyas Mohan, Wei Ye, Dilin Wang, JQ Huang, Rakesh Ranjan, Aviral Chharia, and Fernando De la Torre. CoVeR: Coverage-based token pruning for multi-view 3D reasoning in VLMs. *arXiv preprint arXiv:2609.08345*, 2026.

Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3DJCG: A Unified Framework for Joint Dense Captioning and Visual Grounding on 3D Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16464–16473, 2022.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

Angel X. Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *International Conference on 3D Vision*, pp. 667–676, 2017.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language. In *European Conference on Computer Vision*, pp. 202–221, 2020.

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

Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X. Chang. D3Net: A Unified Speaker-Listener Architecture for 3D Dense Captioning and Visual Grounding. In *European Conference on Computer Vision*, pp. 487–505, 2022a.

Dave Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X. Chang. UniT3D: A Unified Transformer for 3D Dense Captioning and Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18109–18119, 2023a.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language Conditioned Spatial Relation Reasoning for 3D Object Grounding. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20522–20535, 2022b.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-End 3D Dense Captioning with Vote2Cap-DETR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11124–11133, 2023b.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. LL3DA: Visual Interactive Instruction Tuning for Omni-3D Understanding, Reasoning, and Planning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26428–26438, 2024a.

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

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839, 2017.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3D-LLaVA: Towards Generalist 3D LMMs with Omni Superpoint Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3772–3782, 2025.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-LLM: Extending Language Model for 3D Visual Reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2195–2206, 2025.

Anna-Maria Halacheva, Jan-Nico Zaech, Xi Wang, Danda Pani Paudel, and Luc Van Gool. GaussianVLM: Scene-Centric 3D Vision-Language Models Using Language-Aligned Gaussian Splats for Embodied Reasoning and Beyond. *IEEE Robotics and Automation Letters*, 10(12):12588–12595, 2025.

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, pp. 20482–20494, 2023.

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, pp. 113991–114017, 2024a.

Haifeng Huang, Yilun Chen, Zehan Wang, Jiangmiao Pang, and Zhou Zhao. Chat-Scene++: Exploiting context-rich object identification for 3D LLM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48 (8):9813–9825, 2026. doi: 10.1109/TPAMI.2026.3679561.

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 *International Conference on Machine Learning*, pp. 20413–20451, 2024b.

Wencan Huang, Daizong Liu, and Wei Hu. Fast3D: Accelerating 3D multi-modal large language models for efficient 3D scene understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia (ACM MM)*, pp. 10935–10944, 2025. doi: 10.1145/3746027.3755119.

Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-Aware Alignment and Mutual Masking for 3D-Language Pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10984–10994, 2023.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

Han Li, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Geometry-guided 3D visual token pruning for video-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9615–9625, 2026.

Yuhui Lin, Siyue Yu, Yuxing Yang, Guangliang Cheng, and Jimin Xiao. Efficient3D: A unified framework for adaptive and debiased token reduction in 3D MLLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8941–8951, 2026.

Peng Ling, Yingda Yin, Lingting Zhu, Weikai Chen, Shengju Qian, Zeyu Hu, Xin Wang, and Wenming Yang. CoverPrune: Coverage-driven token pruning for 3D VLMs via optimal transport. In *European Conference on Computer Vision (ECCV)*, 2026.

Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3D-SPS: Single-Stage 3D Visual Grounding via Referred Point Progressive Selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16454–16463, 2022.

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.

Guofeng Mei, Wei Lin, Luigi Riz, Yujiao Wu, Fabio Poiesi, and Yiming Wang. PerLA: Perceptive 3D Language Assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14369–14379, 2025.

Guofeng Mei, Wei Lin, Luigi Riz, Yujiao Wu, Yiming Wang, and Fabio Poiesi. Efficient encoder-free Fourier-based 3D large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23785–23794, 2026.

Tianbo Pan, Xingyi Yang, and Xinchao Wang. Merge3d: efficient 3d multimodal llms via joint 2d-3d token merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31066–31077, 2026.

Jiuhe Qu, Yingping Liang, and Ying Fu. HiSC: Hierarchical spatial clustering token compression for efficient 3D scene understanding. *arXiv preprint arXiv:2608.04610*, 2026.

Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. LanguageRefer: Spatial-Language Model for 3D Visual Grounding. In *Conference on Robot Learning*, pp. 1046–1056. PMLR, 2022.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867, 2025.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 567–576, 2015.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19231–19242, 2023.

Haomiao Xiong, Yunzhi Zhuge, Jiawen Zhu, Lu Zhang, and Huchuan Lu. 3UR-LLM: An end-to-end multimodal large language model for 3D scene understanding. *IEEE Transactions on Multimedia*, 27, 2025.

Kai Zhang, Xingyu Chen, and Xiaofeng Zhang. AdaToken-3D: Dynamic spatial gating for efficient 3D large multimodal-models reasoning. *arXiv preprint arXiv:2505.12782*, 2025.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3DRefer: Grounding Text Description to Multiple 3D Objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15225–15236, 2023.

Yue Zhang, Yingzhao Jian, Yunqiu Xu, Xiaoxiao Sun, and Hehe Fan. SmartMage: Dynamic modality orchestration for 3D scene understanding. In *Proceedings of the 34th ACM International Conference on Multimedia (ACM MM)*, 2026.

Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yinjie Lei, Mingkui Tan, and Chuang Gan. LSceneLLM: Enhancing Large 3D Scene Understanding Using Adaptive Visual Preferences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3761–3771, 2025.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2911–2921, 2023.

Appendix for *What Should 3D-LMMs See? Scene-State Curation for 3D Large Multimodal Models*

A Algorithm of PARC-3D

Table A.1: **Algorithmic summary of the PARC-3D scene-state curation procedure.**

Input: Candidate visual units $\mathcal{X} = \{(\mathbf{z}_i, \mathbf{p}_i, n_i, s_i)\}_{i=1}^{N}$ and scene-state size K.
Output: Curated scene state $\mathcal{S}^\star$.

1. Compute the support ratio v_i using Equation 4.
2. Calibrate the retention utility u_i using Equation 5.
3. Compute the semantic d_{ij}^{sem} and spatial d_{ij}^{spa} distances using Equations 6–8.
4. Compute the joint spatio-semantic distance d_{ij} using Equation 9.
5. Estimate the normalized effective rank $\hat{e}$ and derive the adaptive routing scale τ using Equations 11–13.
6. Define the structured redundancy r_{ij} by Equation 10; it is evaluated on the fly against the newly selected unit in step 8 rather than stored as an $N \times N$ matrix.
7. Initialize the selected set $\mathcal{I}_0 = \varnothing$ and residual utilities $\tilde{u}_i^{(0)} = u_i$.
8. Construct the scene state via residual-utility maximization using Equations 14–15 until K units are retained.
9. Return the final index set $\mathcal{I}^\star = \mathcal{I}_K$ and the curated scene state $\mathcal{S}^\star = \{\mathbf{x}_i \mid i \in \mathcal{I}^\star\}$.

Residual-utility derivation. Unrolling the recursion in Equation 15 yields

$$\tilde{u}_j^{(t)} = u_j \prod_{k=1}^{t} \big(1 - \lambda r_{i_k j}\big). \tag{17}$$

For candidates with positive residual utility, the multiplicative update has a simple interpretation in log-utility space. For a selected index set $\mathcal{I}_t$, Equation 17 can be written as

$$\log \tilde{u}_j(\mathcal{I}_t) = \log u_j + \sum_{i \in \mathcal{I}_t} \log(1 - \lambda r_{ij}). \tag{18}$$

Using the first-order approximation $\log(1 - \lambda r_{ij}) \approx -\lambda r_{ij}$ gives

$$\log \tilde{u}_j(\mathcal{I}_t) \approx \log u_j - \lambda \sum_{i \in \mathcal{I}_t} r_{ij}. \tag{19}$$

Thus, the multiplicative update admits a simple log-space interpretation: calibrated utility is progressively discounted by accumulated redundancy with the selected scene state, consistent with the utility–redundancy trade-off in Equation 3: to first order, the procedure performs greedy selection on a log-utility surrogate of that objective, with λ playing the role of η.

B Dataset Details

We evaluate PARC-3D on standard 3D vision-language benchmarks built on real indoor 3D scans from ScanNet (Dai et al., 2017). These datasets cover complementary aspects of unified 3D scene understanding, including question answering, situated reasoning, dense captioning, visual grounding, and referring segmentation. We follow the official splits and the evaluation protocols adopted by prior 3D-LMM works.

ScanRefer and Multi3DRefer. ScanRefer (Chen et al., 2020) provides natural language expressions that refer to target objects in 3D indoor scenes. Multi3DRefer (Zhang et al., 2023) further extends the setting to expressions that may correspond to zero, one, or multiple target objects. We use these datasets to evaluate language-guided localization and referring segmentation, where the model must associate textual descriptions with the correct object-level 3D evidence. Following the referring

segmentation setting, we report mean Intersection-over-Union (mIoU). For the object-centric Chat-Scene interface (Section 4.5), we follow its box-based protocol: Acc@0.25/0.5 on ScanRefer is the fraction of expressions whose predicted box reaches an IoU of at least 0.25/0.5 with the ground truth, and F1@0.25/0.5 on Multi3DRefer is the F1 score between the predicted and ground-truth box sets at the same IoU thresholds.

ScanQA and SQA3D. ScanQA (Azuma et al., 2022) evaluates open-ended question answering over 3D scenes and requires models to reason about objects, attributes, spatial relations, and scene context. We report standard language generation metrics, including CIDEr, BLEU-4, METEOR, and ROUGE-L. SQA3D (Ma et al., 2023) introduces situated question answering, where each question is grounded in an agent-centric situation with position and orientation information. We report Exact Match (EM) and EM-Refined following the standard protocol.

Scan2Cap. Scan2Cap (Chen et al., 2021) evaluates dense captioning in 3D scans, requiring models to localize objects and generate natural language descriptions. This benchmark tests whether the learned scene representation supports both object-level grounding and descriptive reasoning. We follow the standard IoU-based matching protocol and report CIDEr, BLEU-4, METEOR, and ROUGE-L. Metrics with the suffix @0.5 follow the official Scan2Cap evaluation: a prediction whose box has an IoU below 0.5 with its ground-truth object is replaced by an empty caption, and the metrics are computed over all annotated objects.

Together, these benchmarks evaluate whether PARC-3D can provide a reusable scene-state interface across tasks with different structural requirements. ScanRefer and Multi3DRefer emphasize fine-grained referring and spatial disambiguation; ScanQA and SQA3D evaluate scene-level and situated reasoning; Scan2Cap assesses dense object-level description.

C Implementation Details

PARC-3D is instantiated at the visual interface of the official pretrained 3D-LLaVA backbone (Deng et al., 2025). It operates after candidate 3D visual units are produced and before the scene state is consumed by the language model. Each candidate unit provides a semantic feature, a 3D position, a support size, and a saliency/objectness score. PARC-3D curates a scene state of the same size K as vanilla 3D-LLaVA. The upstream candidate pool is scene-dependent: over the evaluated ScanNet scenes, N ranges from 218 to 3,909 superpoint candidates per scene (median about 1,000), so $K = 100$ retains roughly 14% of the pool on average.

All main experiments use one global hyperparameter setting across datasets, tasks, and both interfaces, summarized in Table C.1; the sensitivity study (Table E.1) and the fixed-τ ablation (Table E.2) vary individual parameters around this setting. PARC-3D introduces no additional training stage, no learnable parameters, and no modification to the pretrained visual encoder, language model, task prompts, task decoders, or scene-state size. All experiments were conducted on NVIDIA A800 80GB GPUs.

Table C.1: **PARC-3D hyperparameters.** One setting is used for all main, transfer, and analysis experiments; the last column states where a parameter is deliberately varied.

Parameter	Role	Value	Varied in
α_{sem}	semantic weight in d_{ij} (Equation 9)	1.0	Table 4 (Spatial-only)
β_{spa}	spatial weight in d_{ij} (Equation 9)	1.0	Table E.1; Table 4 (Semantic-only)
γ_{vol}	calibration strength in VSC (Equation 5)	2.0	Table E.1
λ	residual-utility decay strength (Equation 15)	0.8	Table 4 (w/o CSR)
$\tau_{\min}$, $\tau_{\max}$	routing-scale range (Equation 13)	0.05, 0.25	Table E.2 (fixed τ)
K	scene-state size, equal to the vanilla model	100	–

Organization baselines. All rows of Table 2 share the same candidate pool, the features $\mathbf{z}_i$, positions $\mathbf{p}_i$, scores s_i, the scene-state size K, and the downstream decoders; only the rule used to construct the scene state differs. *Saliency* retains the top-K candidates according to s_i. *Random* samples K candidates uniformly without replacement. *Spatial FPS* applies farthest-point sampling to

the candidate centers $\mathbf{p}_i$. *Semantic NMS* performs score-ordered suppression according to semantic similarity with already retained candidates. *MMR* greedily balances normalized upstream relevance against semantic redundancy with the selected set. *DPP* applies a standard quality–diversity formulation with greedy MAP inference to select K candidates. The same implementation of each baseline is used across all evaluated benchmarks.

Ablation variants. The variants of Table 4 each change one component of the full formulation and keep everything else fixed. *w/o VSC* uses the raw saliency as the utility ($u_i = s_i$) while keeping CSR and residual-utility construction. *w/o CSR* keeps the calibrated utility of VSC but sets $\lambda = 0$, so the scene state is the top-K set under u_i without redundancy discounting. *Semantic-only* and *Spatial-only* keep VSC, the adaptive routing scale, and λ, but define the joint distance of Equation 9 with only the semantic term ($\beta_{\text{spa}} = 0$) or only the spatial term ($\alpha_{\text{sem}} = 0$), respectively.

Cross-interface instantiation. For the transfer experiment in Section 4.5, we apply the same PARC-3D formulation to the object-centric Chat-Scene interface (Huang et al., 2024a). Each object proposal is treated as one candidate unit: $\mathbf{z}_i$ is the object-level feature, $\mathbf{p}_i$ is the proposal center, n_i is the number of points in its mask, and s_i is the native proposal confidence. Both variants construct a scene state with $K = 100$, the same size as in the main experiments, and share the released checkpoint, candidate pool, and downstream evaluation pipeline; the PARC-3D row uses the hyperparameters of Table C.1. Both rows of Table 5 are obtained by re-running this checkpoint under this configuration, with the object identifiers of the retained proposals passed to the language model as in the original pipeline.

Computational cost. Curation runs once per scene before language-model prefill and adds no forward pass through the visual encoder or the language model; since the scene-state size is unchanged, the prefill cost of the language model is identical to that of the vanilla model. Table C.2 reports wall-clock measurements on one NVIDIA A800 80GB GPU with batch size 1 and $K = 100$ over 141 ScanRefer scenes. The complete curation takes 46.9 ms on average and finishes within 58.8 ms for 95% of the scenes; the effective-rank computation accounts for 11.9 ms, and most of the cost is the redundancy-aware selection. For reference, the generation window of the language model on the same scenes averages 298.0 ms, and peak allocated GPU memory is 13.3 GB. The greedy construction evaluates redundancy only against the candidate added at each step, so it requires $O(KNd)$ time and $O(N)$ additional memory, where d is the feature dimension; the effective rank requires one eigendecomposition of a $\min(N, d) \times \min(N, d)$ covariance matrix, whose eigenvalues are the squared singular values entering Equation 11.

Table C.2: **Per-scene wall-clock cost of PARC-3D curation.** NVIDIA A800 80GB, batch size 1, $K = 100$, 141 ScanRefer scenes.

Stage	Mean (ms)	Median (ms)	P95 (ms)
Support calibration (VSC)	0.8	0.5	0.7
Effective rank and routing scale	11.9	11.7	21.3
Redundancy and residual-utility selection	32.7	31.5	36.4
Pairwise summary	1.5	1.0	1.3
Total curation	46.9	44.7	58.8
Language-model generation window	298.0	289.4	324.7

D SUPPORT-SIZE BIAS AUDIT

A key motivation of PARC-3D is that upstream saliency or objectness scores can be coupled with the physical support size of candidate 3D units. This section audits this coupling on the actual candidate pools produced by the 3D-LLaVA backbone. The analysis quantifies how support size relates to raw saliency and how this relation affects the scene state constructed under a fixed scene-state size.

For each candidate unit i, we use its raw upstream saliency score s_i and support size n_i. We compute the Pearson correlation between s_i and $\log n_i$, where the logarithm reduces the heavy-tailed effect of support size. We then divide candidates into *small*, *medium*, and *large* groups by support-size

tertiles, and compare their raw saliency statistics. Finally, we measure whether vanilla saliency-ranked scene-state construction amplifies large-support units in the top-K scene state.

We define the amplification ratio for a support-size group b as

$$A_b = \frac{\Pr(i \in \mathcal{I}_{\mathrm{Top}K} \mid i \in b)}{\Pr(i \in \mathcal{I}_{\mathrm{Top}K})} = \frac{\mathrm{Share}_{\mathrm{Top}K}(b)}{\mathrm{Share}_{\mathrm{All}}(b)}, \tag{20}$$

where $\mathcal{I}_{\mathrm{Top}K}$ is the vanilla top-K scene state constructed by the original upstream saliency ranking. $A_b > 1$ means that group b is over-represented in the scene state relative to its share in the candidate pool, and $A_b < 1$ means that group b is under-represented.

Table D.1: **Support-size bias audit on upstream candidate units.** We measure the relation between raw saliency s_i and support size n_i on the actual 3D-LLaVA candidate pools. *Pearson*, *Saliency L/S*, *Top-10% L/S*, and A_{large} denote $\mathrm{corr}(s_i, \log n_i)$, the ratio of mean raw saliency between large- and small-support candidates, the ratio of their probabilities of entering the top saliency decile, and top-K amplification of large-support candidates in vanilla scene-state construction, respectively.

Task	Pearson	Saliency L/S	Top-10% L/S	A_{large}
Scan2Cap	0.207	3.38×	2.01×	1.52
ScanQA	0.200	3.06×	2.04×	1.49
ScanRefer	0.203	3.37×	2.00×	1.51

Table D.1 shows a stable positive coupling between raw saliency and support size across all evaluated tasks. The Pearson correlation between s_i and $\log n_i$ is consistently around 0.20. In this setting, raw saliency reflects multiple factors, including semantic category, local geometry, objectness, visual quality, and context. The stable correlation around 0.20 across Scan2Cap, ScanQA, and ScanRefer shows that physical support is consistently associated with the upstream saliency score. The stratified statistics provide a more direct view of this relation: large-support candidates have about 3.1–3.4× higher mean saliency than small-support candidates, and they are about 2× more likely to enter the top saliency decile. Vanilla top-K scene-state construction further amplifies large-support candidates by about 1.5× relative to their share in the full candidate pool. These results provide direct empirical evidence that upstream saliency is positively coupled with support size and can bias the scene state toward large-support evidence.

Table D.2: **Detailed support-size stratification.** Candidates are divided into *small*, *medium*, and *large* groups by support-size tertiles. We report mean raw saliency, the probability of entering the top saliency decile, and the top-K amplification ratio A_b under vanilla scene-state construction.

Task	Mean saliency			Top-10% probability			Amplification A_b		
	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large
Scan2Cap	0.0197	0.0233	0.0666	7.28%	8.19%	14.63%	0.70	0.79	1.52
ScanQA	0.0229	0.0269	0.0700	7.22%	8.17%	14.70%	0.73	0.79	1.49
ScanRefer	0.0195	0.0235	0.0658	7.30%	8.15%	14.63%	0.71	0.79	1.51

Table D.2 gives the full stratified statistics. The same pattern appears consistently across Scan2Cap, ScanQA, and ScanRefer: large-support candidates have higher raw saliency and are over-represented after vanilla top-K construction, while small-support candidates have lower raw saliency and are under-represented. This support-associated advantage helps explain why saliency-dominated construction can yield a scene state biased toward large spatial structures. The audit directly motivates VSC, which calibrates raw saliency by support size, and complements CSR, which further models redundancy using both semantic and physical distances.

E Additional Results Analysis

This section provides supplementary analyses of the scene state constructed by PARC-3D. We examine support-size composition, report hyperparameter sensitivity and the fixed-routing-scale

comparison, provide complete coverage and redundancy statistics, relate the effective-rank signal used by CSR to scene statistics, and report the stability of the main results across evaluation-scene subsets.

E.1 Scene-State Composition by Support Size

Complementing the support-size bias audit in Appendix D, Figure E.1 analyzes how PARC-3D changes the support-size composition of the final scene state under the same scene-state size. The audit in Appendix D measures the coupling between raw saliency and support size in the upstream candidate pool, while this analysis compares the resulting scene-state composition after vanilla and PARC-3D construction. Compared with vanilla 3D-LLaVA, PARC-3D raises the share of small-support units from 23.97% to 26.60% (+2.63 pp), while the shares of medium- and large-support units move from 25.68% to 24.59% (-1.09 pp) and from 50.35% to 48.81% (-1.54 pp). The change is a moderate rebalancing: PARC-3D shifts composition toward under-represented small-support evidence while preserving high-relevance large-support evidence, which still forms about half of the scene state. This trend is consistent with the target-size rows of Table 3, where PARC-3D brings larger gains on small and medium targets.

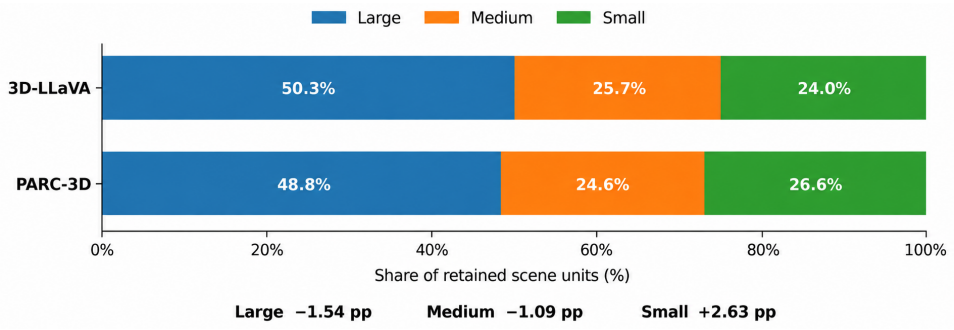


Figure E.1: **Effect on support-size composition under the same scene-state size.** PARC-3D moderately rebalances the composition toward small-support units (+2.63 pp) while largely preserving large-support evidence (-1.54 pp); deltas are in percentage points.

Table E.1: **Hyperparameter sensitivity.** We vary the spatial weight β_{spa} and calibration strength γ_{vol} around the shared default configuration. Best results per metric within each sweep are in **bold**.

Metric	Spatial weight β_{spa}				Calibration strength γ_{vol}			
	0.5	1.0	1.5	2.0	1.0	2.0	3.0	4.0
ScanQA CIDEr	95.5	**96.3**	96.1	96.0	**96.3**	**96.3**	**96.3**	96.0
ScanQA B-4	15.4	**16.5**	**16.5**	15.8	16.2	**16.5**	16.1	**16.5**
Scan2Cap C@0.5	78.5	**78.8**	78.6	78.4	78.5	78.8	**79.0**	78.4
Scan2Cap B-4@0.5	37.2	**37.6**	37.5	37.5	**37.6**	**37.6**	37.4	37.5

E.2 Adaptive Routing Scale

We test whether the effective-rank-driven routing scale in Equation 13 is preferable to a fixed scale. Replacing $\tau(\hat{e})$ with fixed values spanning $[\tau_{\min}, \tau_{\max}]$, while keeping all other components unchanged, yields ScanRefer mIoU of 45.3–45.9 and ScanQA CIDEr of 95.5–95.8. None of the tested fixed scales recovers the adaptive result of 46.2 and 96.3, respectively.

Table E.2: **Adaptive routing-scale ablation.** Fixed τ replaces the effective-rank-driven scale in Equation 13.

Scale	ScanRefer	ScanQA
	mIoU	CIDEr
$\tau = 0.05$	45.5	95.5
$\tau = 0.15$	45.9	95.8
$\tau = 0.25$	45.3	95.5
$\tau(\hat{e})$	**46.2**	**96.3**

E.3 Spatial Coverage and Redundancy

Table E.3 further reports coverage and redundancy statistics of the scene state. Compared with vanilla 3D-LLaVA, PARC-3D increases the average nearest-neighbor spatial distance from 0.0328 to 0.0385 and the average pairwise spatial distance from 0.3247 to 0.3320, indicating broader spatial coverage. The average pairwise semantic distance and the average pairwise joint distance also increase, suggesting lower redundancy in feature space and joint spatio-semantic space. These results show that PARC-3D improves the structural distribution of the scene state, rather than only changing its support-size composition.

Table E.3: **Spatial coverage and redundancy statistics.** Larger values indicate broader coverage or lower redundancy within the scene state. Entries are mean $\pm$ standard deviation over evaluation scenes. NN-Spatial denotes the average nearest-neighbor spatial distance. Pair-Spatial, Pair-Semantic, and Pair-Joint denote average pairwise distances in spatial, semantic, and joint spatio-semantic spaces, respectively.

Method	NN-Spatial↑	Pair-Spatial↑	Pair-Semantic↑	Pair-Joint↑
3D-LLaVA	0.0328 ± 0.0062	0.3247 ± 0.0382	0.3473 ± 0.0640	0.6720 ± 0.0777
PARC-3D	0.0385 ± 0.0063	0.3320 ± 0.0344	0.3564 ± 0.0579	0.6884 ± 0.0660

Together with Figure E.1, these statistics provide complementary evidence for the effect of PARC-3D. The support-size analysis shows that the scene state becomes less dominated by large-support regions, while the coverage and redundancy statistics show that the curated visual units are more spatially dispersed and less redundant in semantic and joint spaces. This supports the design of combining Volume-Aware Saliency Calibration with Complexity-Aware Spatio-Semantic Routing.

E.4 Effective Rank and Scene Statistics

CSR sets the routing scale τ from the normalized effective rank $\hat{e}$ of the candidate feature matrix (Equation 13). To characterize the representation structure captured by this signal, Table E.4 reports its Spearman correlation with several scene statistics computed from the ScanNet point clouds and ground-truth annotations. $\hat{e}$ correlates strongly with the number of scene points and the total volume of annotated object boxes, and also tracks global scene extent and object-instance complexity. These results connect the spectral structure of the candidate representation to the amount and spatial distribution of scene content, providing the basis for scene-adaptive routing: the effective rank rises with the amount and extent of scene content, which is the regime in which CSR shrinks τ so that only closely overlapping pairs are treated as redundant. Computed directly from the candidate features available at inference time, $\hat{e}$ provides a representation-level spectral signal for adapting the routing scale to scene-content complexity.

Table E.4: **Spearman correlation between the normalized effective rank $\hat{e}$ and scene statistics.** Statistics are computed per scene from the point cloud and ground-truth (GT) annotations. GT statistics serve only this post-hoc characterization and are not available to PARC-3D at inference time.

Scene statistic	Spearman ρ with $\hat{e}$
Number of points	0.823
Sum of GT box volumes	0.766
Volume of the enclosing GT box	0.692
Point-cloud AABB volume	0.670
Number of object instances	0.585

E.5 Stability Across Test-Scene Subsets

To check that the main results do not depend on a few particular scenes, we keep the fixed configuration and re-evaluate PARC-3D on 20 independently sampled subsets, each containing 90% of the evaluation scenes of a benchmark. Table E.5 reports the mean and standard deviation across subsets next to the full-set results of Table 1. The standard deviations lie between 0.06 and 0.20 points, an order of magnitude below the main gains, and the 0.6-point ScanQA BLEU-4 decrease shows the same small spread across subsets.

Table E.5: **Stability across evaluation-scene subsets.** Full-set results and the mean $\pm$ standard deviation of PARC-3D over 20 random subsets, each covering 90% of the evaluation scenes, under the fixed configuration.

Benchmark / metric	Vanilla	PARC-3D (full set)	PARC-3D (subsets)
ScanRefer mIoU	43.3	46.2	46.2 ± 0.10
Multi3DRefer mIoU	42.7	52.1	52.1 ± 0.15
ScanQA CIDEr	92.6	96.3	96.3 ± 0.20
ScanQA BLEU-4	17.1	16.5	16.5 ± 0.06
SQA3D EM	54.5	55.7	55.7 ± 0.08
SQA3D EM-R	56.6	58.7	58.7 ± 0.09
Scan2Cap CIDEr@0.5	78.8	78.8	78.8 ± 0.15
Scan2Cap BLEU-4@0.5	36.9	37.6	37.6 ± 0.06

F Additional Qualitative Analysis

This section provides additional qualitative analysis of PARC-3D. Figure F.1 first visualizes the spatial distribution of the retained scene units in two representative scenes. We then show successful examples across referring segmentation, 3D question answering, and dense captioning, and finally discuss representative failure cases. These examples complement the quantitative results by illustrating how the curated scene state affects different downstream tasks under the same scene-state size.

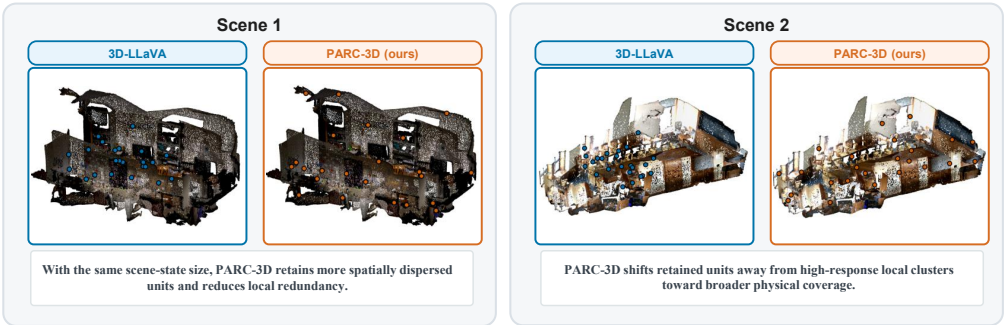


Figure F.1: **Spatial distribution of retained scene units.** Vanilla construction concentrates the retained units on large surfaces and high-response clusters; PARC-3D distributes them more evenly, so small objects and their surrounding context are represented within the same scene-state size.

F.1 Additional Successful Cases

Figure F.2 shows additional referring segmentation examples. The cases require grounding objects using local spatial cues, including a dishwasher below the counter, a window located between nearby objects, and a backpack on a chair near an orange sofa. Compared with vanilla 3D-LLaVA, PARC-3D produces masks that better align with the queried targets. This suggests that the curated scene state can preserve more useful local evidence for fine-grained grounding.

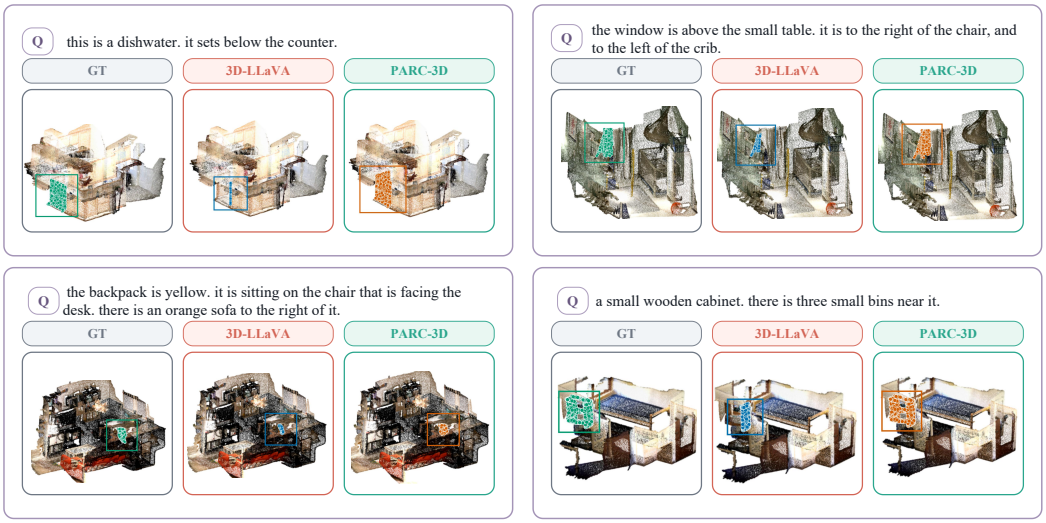


Figure F.2: **Additional qualitative results on referring segmentation.** We compare the ground-truth mask, vanilla 3D-LLaVA, and PARC-3D on four referring expressions. PARC-3D better grounds the queried objects in these examples, especially when the expression depends on local spatial cues and fine-grained object evidence.

Figure F.3 presents additional examples on 3D question answering. The questions involve spatial layout, object category, and object attribute reasoning. PARC-3D correctly answers the side of the room for the bed closest to the entrance door and the color of the couches around the table. In the middle example, PARC-3D predicts “chair”, which recovers the object category of the ground truth “2 chairs” but misses the annotated cardinality, whereas vanilla 3D-LLaVA answers “lamp”. These cases suggest that improving the scene state can provide more reliable local evidence for 3D question answering.

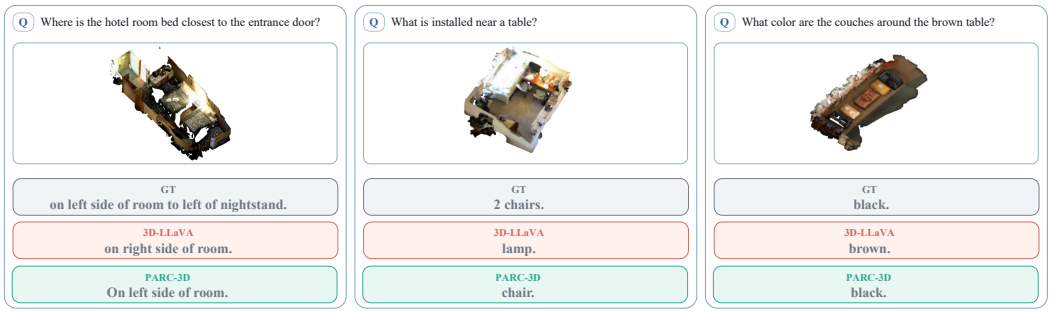


Figure F.3: **Additional qualitative results on 3D question answering.** The examples cover spatial layout, object category, and object attribute questions. PARC-3D provides more accurate or more relevant answers than vanilla 3D-LLaVA in these cases, indicating that scene-state curation can improve the evidence available for 3D question answering.

Figure F.4 shows additional dense captioning examples under the mask-conditioned setting. In the soap-dispenser case, vanilla 3D-LLaVA recognizes the object but grounds its relation to the wrong nearby anchor, while PARC-3D correctly describes it as being to the left of the sink. In the pillow case, PARC-3D produces a more specific spatial description by placing the pillow on the right side of the bed. These examples indicate that PARC-3D can help preserve relation anchors and local context needed for dense object descriptions.

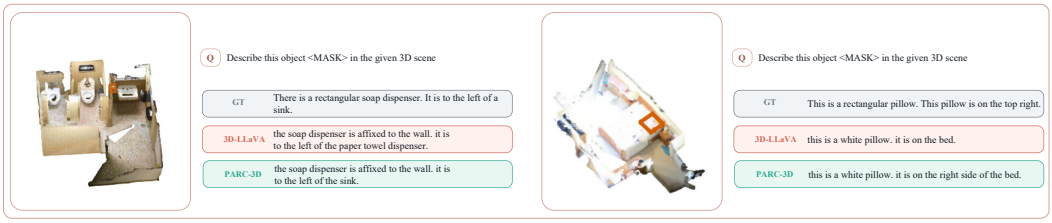


Figure F.4: **Additional qualitative results on dense captioning.** We show two mask-conditioned dense captioning examples. PARC-3D improves relation-anchor grounding in the soap-dispenser example and produces a more specific spatial description in the pillow example.

Across these successful examples, PARC-3D keeps the task prompts, pretrained weights, and downstream decoders unchanged. The qualitative improvements therefore support the main observation that scene-state curation can improve the visual evidence presented to the language model, especially when the task depends on fine-grained object evidence, local spatial relations, or relation-anchor grounding.

F.2 Failure Cases and Discussion

Although PARC-3D improves the scene state in many cases, it remains a task-agnostic curation mechanism and does not explicitly perform query-specific disambiguation or spatial-relation reasoning. Figure F.5 shows representative failure cases that reveal the current limitations of scene-state curation.

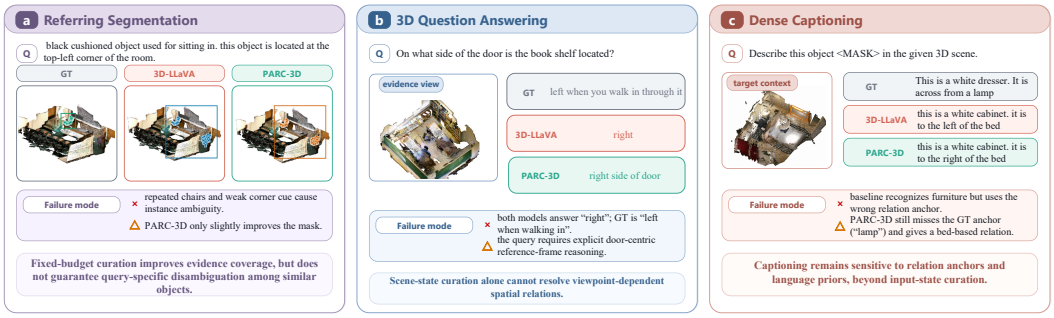


Figure F.5: **Representative failure cases of PARC-3D.** (a) In referring segmentation, repeated chair-like objects and a weak corner cue lead to instance ambiguity. PARC-3D only slightly improves the mask IoU, indicating that better evidence coverage does not guarantee query-specific disambiguation among similar objects. (b) In 3D question answering, both vanilla 3D-LLaVA and PARC-3D answer "right", while the ground-truth answer is "left when you walk in through it". This case requires explicit door-centric reference-frame reasoning, which cannot be fully resolved by scene-state curation alone. (c) In dense captioning, vanilla 3D-LLaVA recognizes the furniture category but predicts an incorrect spatial relation. PARC-3D still misses the ground-truth relation anchor, i.e., the lamp, and produces a bed-based relation instead. This suggests that captioning remains sensitive to relation-anchor grounding and language priors beyond input-state curation.

In Figure F.5(a), the expression refers to a black cushioned object located at the top-left corner of the room. The scene contains multiple visually and semantically similar chair-like objects, and the location cue is not sufficiently discriminative. PARC-3D preserves more local evidence and slightly improves the segmentation mask, but the prediction still drifts toward nearby similar instances. This shows that scene-state curation can improve target evidence coverage, but may still be insufficient when the query requires fine-grained instance-level disambiguation.

Figure F.5(b) illustrates a different limitation in 3D question answering. The question asks which side of the door the bookshelf is located on. Both vanilla 3D-LLaVA and PARC-3D predict the bookshelf to be on the right side, while the annotation defines the answer as left when entering through the door. This error is less about missing local evidence and more about viewpoint-dependent spatial reasoning. It suggests that improving the shared scene state alone does not fully solve reference-frame ambiguity.

Figure F.5(c) further shows that dense captioning can fail when the model grounds the target object but chooses an incorrect relation anchor. PARC-3D improves the input scene state, but the generated caption still describes the object relative to the bed rather than the lamp. This indicates that language generation remains affected by relation-anchor selection and learned language priors. Overall, these cases suggest that PARC-3D should be viewed as a scene-state curation mechanism, complementary to future query-aware curation, explicit spatial-relation modeling, and stronger relation-anchor grounding.

G Limitations and Broader Impact

PARC-3D still relies on hand-designed calibration and routing rules, and it does not explicitly model global scene graphs or hierarchical room-level structures. Future work may develop more principled scene-state curation objectives and extend the method to outdoor, dynamic, and embodied environments. Its potential applications include robotics, AR, and spatial reasoning, where better-organized 3D scene states may improve downstream reasoning. At the same time, incorrect scene understanding in embodied or AR systems may lead to unsafe or misleading decisions, so deployment should follow standard privacy, safety, and human-oversight safeguards.