LENS3D: TARGET-CONDITIONED VISUAL FOVEATION FOR FINE-GRAINED 3D UNDERSTANDING

Anonymous authors
Paper under double-blind review

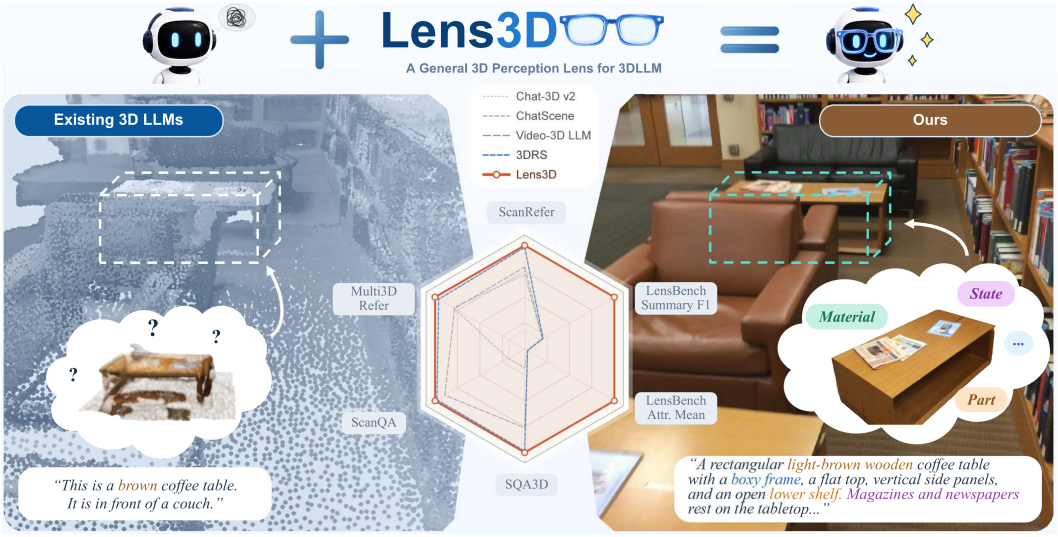


Figure 1: **Lens3D brings fine-grained object understanding to 3D LLMs.** While existing models capture broad scene structure and salient objects (left), Lens3D further capture instance-level attributes and easily overlooked small objects (right), with the performance radar showing these gains in fine-grained understanding do not come at the cost of the models' original grounding and question-answering abilities (center).

ABSTRACT

Existing 3D large language models often overlook fine-grained attributes and less visually salient objects and parts, even when relevant evidence is present in scene videos. We introduce *Lens3D* to improve fine-grained object understanding through external visual assistance and knowledge transfer. Its *LensUnd* pipeline adopt 3D localization to select informative, complementary views for an external 2D vision-language model, supporting fine-grained object captioning, small-object grounding, and fine-grained object question answering. And its *LensDistill* transfers the resulting fine-grained knowledge to 3D LLMs through detailed caption supervision, enabling captioning from native inputs without external VLM calls. We also construct *LensBench*, a held-out evaluation set of 2,068 objects with three silver-standard reference descriptions per object. Experiments with Video-3D LLM and 3DRS demenstrate that LensDistill substantially improves fine-grained object captioning while preserving existing grounding and scene-level QA performance. These results establish the feasibility of transferring externally acquired fine-grained knowledge into native 3D LLMs.

1 INTRODUCTION

Recent 3D large language models (3D LLMs), such as Video-3D LLM (Zheng et al., 2025) and 3DRS (Huang et al., 2025), have advanced 3D scene understanding. However, they often overlook

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.

fine-grained object attributes, as well as less visually salient objects and parts. For example, a model may locate an overhead cabinet yet struggle to identify its door material or handle orientation. These details matter for embodied interaction: handle orientation can affect a robot's grasping pose and manipulation strategy. Reliable scene understanding therefore requires both locating objects and recognizing their fine-grained properties.

High-resolution scene videos often contain the necessary visual evidence, but two factors limit its utilization. First, existing 3D datasets emphasize object categories, referring expressions, and short descriptions, providing limited supervision for fine-grained attributes (Dai et al., 2017; Chen et al., 2020; 2021). Second, relevant details may not be visually salient, and their visibility varies across viewpoints. Even when captured in the video, these details may be difficult for a model to distinguish from the surrounding content and use for the task. Together, these challenges motivate more effective utilization of the available visual evidence and richer supervision for fine-grained understanding.

Recent work on 2D vision-language models (VLMs) shows the value of focusing on informative image regions for closer inspection (Zhu et al., 2025; Carvalho et al., 2026). Inspired by this principle, we exploit 3D localization to guide evidence selection across video frames. Our framework, *Lens3D*, combines externally assisted fine-grained understanding with knowledge transfer to 3D LLMs. Its external understanding pipeline, *LensUnd*, uses a 3D LLM to locate the object or region relevant to a query, such as the cabinet containing the handle of interest. It then selects clear, complementary views from the input video. An external 2D VLM examines these focused evidences to perform fine-grained object captioning, small-object grounding, or fine-grained object question answering (QA). This connects existing 3D grounding capabilities with fine-grained 2D perception.

Beyond external assistance, we ask whether the resulting knowledge can be transferred to the 3D LLM itself. To this end, we introduce *LensDistill*, which uses LensUnd as a teacher through asymmetric privileged-information distillation (Cui et al., 2026). We study this transfer through fine-grained object captioning, which provides supervision for materials, parts, states, and spatial relationships. The teacher accesses the selected visual evidence, while the student learns to generate detailed descriptions from its native inputs. After distillation, the student performs captioning without additional view processing or external VLM calls.

Evaluating this capability requires references that capture instance-level details, which existing captioning benchmarks cover sparsely. We therefore construct *LensBench*, a held-out evaluation set of 2,068 objects with three independently generated VLM descriptions per object. A questionnaire-based quality audit supports their use as silver-standard references. Experiments with two representative 3D LLMs, Video-3D LLM and 3DRS, show that LensDistill substantially improves fine-grained object captioning, reaching performance comparable to a similarly sized 2D VLM while preserving existing grounding and scene-level QA performance. Qualitative results further illustrate LensUnd's utility for small-object grounding and fine-grained object QA. Our contributions are:

- We introduce *LensUnd*, which uses 3D localization to better exploit visual details already present in scene videos. With an external 2D VLM, it supports fine-grained object captioning, small-object grounding, and fine-grained object QA.
- We propose *LensDistill* and demonstrate that detailed caption supervision can transfer externally acquired object knowledge to 3D LLMs. This improves fine-grained object captioning without external VLM calls at inference, while preserving existing grounding and scene-level QA performance.
- We construct *LensBench*, covering 2,068 objects with three silver-standard reference descriptions per object. These references support evaluation of instance-level attributes that are sparsely covered by existing 3D captioning benchmarks.

2 Related Work

2.1 3D scene understanding

3D scene understanding spans object detection, semantic and instance segmentation, and language-conditioned tasks such as grounding, captioning, and question answering. ScanRefer (Chen et al., 2020), ReferIt3D (Achlioptas et al., 2020), and Multi3DRefer (Zhang et al., 2023) study 3D target localization from natural-language expressions, while 3DVG-Transformer (Zhao et al., 2021),

BUTD-DETR (Jain et al., 2022), and EDA (Wu et al., 2023) improve grounding through cross-modal scene modeling. Scan2Cap (Chen et al., 2021) introduced 3D dense captioning; D3Net (Chen et al., 2022a) and Vote2Cap-DETR (Chen et al., 2023) subsequently advanced joint grounding and end-to-end captioning, respectively, and PQ3D (Zhu et al., 2024b) uses promptable queries to unify multiple 3D vision–language tasks. ScanQA (Azuma et al., 2022) and SQA3D (Ma et al., 2023) further evaluate scene-level spatial reasoning. These tasks and datasets, however, primarily emphasize object categories, and spatial relations, with limited coverage of fine-grained object details.

2.2 3D LARGE LANGUAGE MODELS

3D large language models connect 3D scene representations with pretrained language models to support grounding, description, question answering, and even generation within a unified framework. For 3D understanding, 3D-LLM (Hong et al., 2023), Chat-3D (Wang et al., 2023), LL3DA (Chen et al., 2024a), and LEO (Huang et al., 2023b) explore different interfaces between point clouds, object tokens, and language models, while LLaVA-3D (Zhu et al., 2024a) augments multiview 2D features with 3D spatial information. Beyond understanding, ShapeLLM-Omni (Ye et al., 2025) and CG-MLLM (Huang et al., 2026) extend 3D multimodal models toward unified understanding and generation, enabling them to both reason about and synthesize 3D content. However, incorporating 3D perception into pretrained language models typically requires additional 3D encoders or alignment modules, limiting their ability to directly inherit the rich knowledge of large-scale pretrained models. Video-based approaches provide an alternative by representing multiview 3D observations as video, allowing them to directly exploit the strong visual and temporal understanding capabilities of pretrained models.Video-3D LLM (Zheng et al., 2025) treats multiview indoor observations as video and injects 3D positional encodings, while 3DRS (Huang et al., 2025) further introduces explicit 3D representation supervision. Despite these increasingly capable architectures and their access to rich multimodal observations, existing 3D LLMs remain constrained by datasets that emphasize categories, spatial relations, and other coarse supervision (Dai et al., 2017; Chen et al., 2020; 2021). Consequently, their understanding remains coarse, with limited training signal for fine-grained object details and small, easily overlooked objects.

2.3 DISTILLATION WITH PRIVILEGED INFORMATION

Knowledge distillation (Hinton et al., 2015) transfers a teacher's outputs or representations to a student, while Learning Using Privileged Information (Vapnik & Vashist, 2009) allows the teacher to access additional observations that are unavailable to the deployed student. Generalized distillation (Lopez-Paz et al., 2015) combines these ideas under asymmetric teacher–student observations. In 3D understanding, SAT (Yang et al., 2021) uses training-time 2D semantics to improve point-cloud grounding. X-Trans2Cap (Yuan et al., 2022) transfers 2D appearance knowledge from a multimodal teacher to a point-cloud captioner, while ExCap3D (Yeshwanth et al., 2025) uses visibility-ranked multiview images and a VLM to construct fine-grained descriptions for training a dedicated captioning model. Vision-OPD (Yuan et al., 2026) similarly treats target-region crops as privileged visual input and transfers regional perception to the full-image policy through on-policy self-distillation. Lens3D extends this direction with a target-conditioned evidence interface shared across fine-grained object captioning, small-object grounding, and fine-grained object QA. LensUnd selects views by marginal target-surface coverage. LensDistill then uses captions generated from these views to train general 3D LLMs. Supplementary Sections A.3 and A.5 compare view coverage and description quality across different selection strategies.

3 METHOD

Lens3D consists of LensUnd, a general target-conditioned understanding pipeline, and LensDistill, an asymmetric privileged-information distillation method. As illustrated in Figure 2, LensUnd follows a Ground–Focus–Understand scheme. It first uses a 3D LLM to locate the object or region relevant to a language query. The resulting 3D bounding box guides the selection of complementary RGB views, in which the target region is explicitly marked. An external 2D VLM then processes these views according to a task-specific instruction. Through different instructions and output formats, the same evidence collection procedure supports fine-grained object captioning, small-object grounding, and fine-grained object QA. LensDistill then transfers the resulting fine-grained knowl-

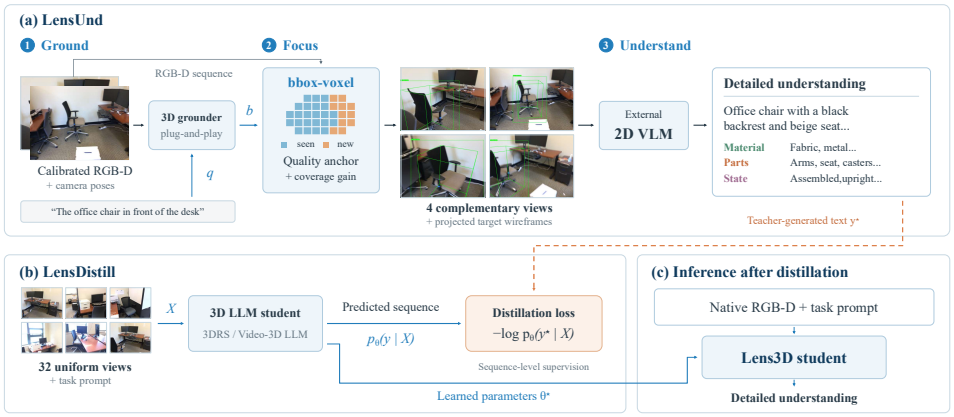


Figure 2: **Overview of Lens3D.** (a) LensUnd follows Ground–Focus–Understand: a 3D grounder localizes an anchor box b from a language query q, bbox-voxel selects complementary views by coverage gain and projects the target wireframe onto them, and an external 2D VLM produces a detailed understanding $y^\star$ of the target. (b) LensDistill trains a 3D LLM student with the sequence-level distillation loss $-\log p_\theta(y^\star \mid X, c(b), p)$, using teacher-generated $y^\star$ as privileged supervision while the student receives uniformly sampled native views, the target center $c(b)$ encoded as <coord>, and a task prompt. (c) After distillation, the external LensUnd path is removed; the student produces the same detailed understanding from native RGB-D, the target-center, and a task prompt.

edge to a 3D LLM operating on its native inputs. We study this transfer through fine-grained object captioning, using structured, detailed captions as supervision. After distillation, the student generates detailed descriptions without additional view processing or external VLM calls at inference.

3.1 Problem Setting

Let an indoor scene be

$$S = \{(I_t, D_t, \Pi_t)\}_{t=1}^{T}, \tag{1}$$

where I_t, D_t, and Π_t denote the RGB image, depth map, and camera parameters of frame t. Given a language query q, a 3D grounder G predicts an bounding box b for the relevant object or region. LensUnd constructs target-conditioned visual evidence $\mathcal{E}_b$, which an external 2D VLM U processes under task instruction p. LensDistill trains a student f_θ using its native scene input X, the target center $c(b)$ encoded by the <coord> token, and instruction p:

$$b = G(S, q), \qquad y^\star = U(\mathcal{E}_b, p), \qquad y = f_\theta(X, c(b), p). \tag{2}$$

Depth and camera parameters support target visibility estimation, 3D-box projection, and view selection. Following Video-3D LLM and 3DRS, X comprises subsampled RGB views with depth- and camera-derived 3D positional encodings for their visual tokens. We perform distillation through fine-grained object captioning, using the external VLM's descriptions as teacher responses $y^\star$. LensUnd also supports small-object grounding and fine-grained object QA by varying p.

3.2 Target-Conditioned Evidence Acquisition

Given target box b, RGB-D sequence S, and frame budget K, the Focus stage selects complementary views that clearly show the target region and explicitly marks that region in each image.

Frame quality and target-surface voxels. Let $\mathcal{C}$ contain geometrically valid, depth-visible candidate frames. We assign each candidate a quality score s_t based on projected target completeness, image coverage, centrality, and occlusion estimated through depth consistency. We then unproject its depth pixels, retain 3D points inside b, and discretize them into an occupied voxel set $\mathcal{V}_t$. All views use a shared voxel grid with approximately $n_v = 12$ cells along the shortest box axis. Thus, $\mathcal{V}_t$ represents observed surfaces within the target region rather than its projected image area.

Algorithm 1 BBox-voxel frame selection

Input: Scene S, target box b, frame budget K, temporal gap δ, voxel resolution n_v
Output: Selected frame indices $\mathcal{S}$, where $|\mathcal{S}| \leq K$

Construct candidate set $\mathcal{C}$ from S and b
for all $t \in \mathcal{C}$ **do**
    Compute frame quality s_t and visible voxels $\mathcal{V}_t$
end for
$\mathcal{S} \leftarrow \varnothing; \mathcal{U} \leftarrow \varnothing$
while $|\mathcal{S}| < K$ **do**
    $\mathcal{C}' \leftarrow \mathcal{C} \setminus \mathcal{S}$
    Remove $t \in \mathcal{C}'$ if $|t - t'| < \delta$ for any $t' \in \mathcal{S}$
    if $\mathcal{C}' = \varnothing$ **then**
        break
    end if
    for all $t \in \mathcal{C}'$ **do**
        $g_t \leftarrow |\mathcal{V}_t \setminus \mathcal{U}|$
    end for
    if $\mathcal{S} = \varnothing$ **or** $\max_{t \in \mathcal{C}'} g_t = 0$ **then**
        $t^\star \leftarrow \arg\max_{t \in \mathcal{C}'} s_t$
    else
        $t^\star \leftarrow \arg\max_{t \in \mathcal{C}'} g_t$
    end if
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{t^\star\}$
    $\mathcal{U} \leftarrow \mathcal{U} \cup \mathcal{V}_{t^\star}$
end while
return $\mathcal{S}$

Greedy view selection. Selecting frames solely by quality can yield redundant views. We instead initialize the selected set $\mathcal{S}$ with the highest-quality frame, which has the highest s_t, and greedily maximize additional surface coverage:

$$\Delta(t \mid \mathcal{S}) = \left| \mathcal{V}_t \setminus \bigcup_{t' \in \mathcal{S}} \mathcal{V}_{t'} \right|. \tag{3}$$

Selected frames must be separated by at least δ frames. When no feasible candidate adds new voxels, we fill the remaining slots by quality under the same temporal constraint. Selection ends when K frames are chosen or no feasible candidates remain. We use $K = 4$ and $\delta = 12$ by default. We call this bounding-box voxel-selection strategy *bbox-voxel*; Algorithm 1 gives the complete procedure.

Target marking. We project the twelve edges of b onto each selected RGB image using its camera parameters, obtaining the resulting wireframe $\widetilde{I}_t$ and evidence $\mathcal{E}_b$:

$$\widetilde{I}_t = \mathrm{Highlight}(I_t, b, \Pi_t), \qquad \mathcal{E}_b = \{\widetilde{I}_t\}_{t \in \mathcal{S}}. \tag{4}$$

This operation supplies an explicit instance-level spatial condition to the 2D VLM, allowing it to identify the object to describe when multiple same-class objects are visible.

3.3 LensUnd: A plug-and-play understanding pipeline

LensUnd follows Ground–Focus–Understand pipeline as following:

$$b = G(S, q), \qquad \{\widetilde{I}_{t_k}\}_{k=1}^{K} = \mathrm{Focus}(S, b), \qquad r = U(\{\widetilde{I}_{t_k}\}_{k=1}^{K}, p), \tag{5}$$

where G is a 3D grounding model, Focus is the target-conditioned evidence process in Section 3.2, U is a 2D VLM, p is a downstream-task instruction, and r is the task-specific response. In the standard setting, q is a ScanRefer-style referring expression (Chen et al., 2020). LensUnd requires only the predicted box from G and does not modify the grounder's architecture.

The same evidence selection procedure supports fine-grained object captioning, small-object grounding, and fine-grained object QA through different instructions and output formats. For small-object grounding, the initially grounded region can contain the smaller object of interest. LensUnd therefore accommodates compatible 3D grounders and 2D VLMs without changing the evidence selection procedure. We quantitatively evaluate fine-grained object captioning with Video-3D LLM and 3DRS; Figure 3 illustrates the other two tasks.

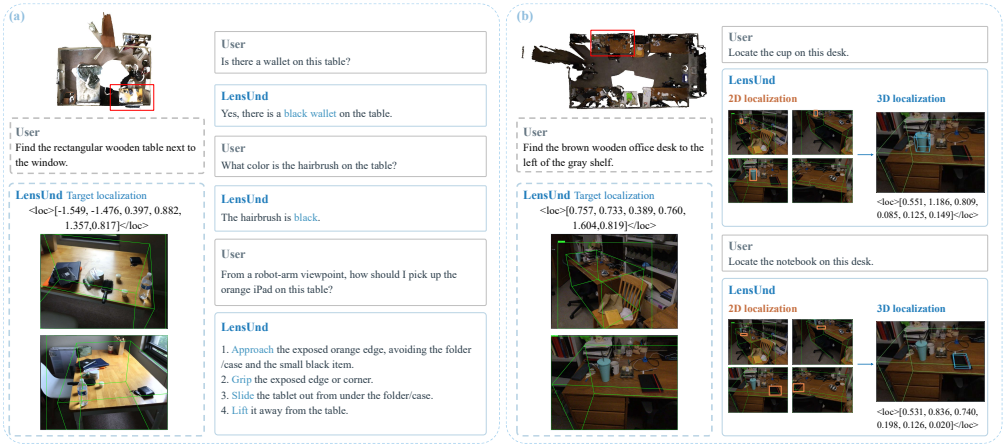


Figure 3: **Fine-grained interactions with LensUnd.** Complementary views selected by LensUnd support (a) object QA on presence, appearance, and manipulation, and (b) small-object grounding by combining per-view 2D predictions into a 3D box using enclosed depth and camera parameters.

3.4 LensDistill: Asymmetric distillation with privileged information

LensUnd relies on the full RGB-D sequence, target-conditioned view selection, and an additional 2D VLM call, introducing inference cost when directly deploying. To avoid this cost, we propose LensDistill, an asymmetric teacher–student framework for transferring externally acquired fine-grained knowledge into native 3D LLMs. During training, the teacher accesses privileged target-level evidence, while the student operates on the same observation space used at test time. At inference, the student produces the corresponding fine-grained understanding without privileged evidence or external VLM calls. We perform and validate this transfer paradigm on fine-grained object captioning.

Teacher and student inputs. For a target region represented by a 3D box b, LensUnd first constructs privileged evidence $\mathcal{E}_b = \{\widetilde{I}_t\}_{t\in\mathcal{S}}$, from which an external VLM teacher generates a fine-grained understanding $y^\star$ under task instruction p. The student retains the base 3D LLM's native scene input X, the target center $c(b)$ encoded by the `<coord>` token, and instruction p. The selected, wireframe-marked views are used only to generate teacher supervision. LensDistill thus transfers knowledge acquired from privileged target-level observations rather than matching teacher and student responses under identical inputs.

Teacher supervision and training. For training target region specified by b, LensUnd constructs target-conditioned evidence and invokes the chosen 2D VLM to generate $y^\star$ as supervision for observation X. This formulation allows supervision to be extended to new scenes, objects, instructions, or teacher models. Given $y^\star = (y_1^\star, \ldots, y_L^\star)$, we train the student with the autoregressive objective

$$\mathcal{L}_{\text{distill}} = -\sum_i \log p_\theta(y_i^\star \mid y_{<i}^\star, X, c(b), p) \tag{6}$$

where p_θ is the student's conditional token distribution. After distillation, the student generates detailed descriptions directly from $(X, c(b), p)$ without additional view processing or 2D VLM calls.

4 Experiments

LensUnd demonstrates small-object grounding and fine-grained object QA over focused multiview evidence. Fine-grained object captioning provides the quantitative testbed for establishing the benefit of target-conditioned evidence and LensDistill's transfer effectiveness across different 3D LLMs.

Table 1: Core LensBench results. "Attr. Mean" averages Parts, Materials, and State F1 and is reported only for models that produce the required structured fields. Human satisfaction (Human Sat.dcfr) is rated on a four-point scale. Best results in each column are highlighted in red, second-best in orange, and third-best in yellow.

Method	Summary F1	Summary Sem.	Attr. Mean	Human Sat.
Qwen3-VL-2B (Bai et al., 2025)	0.380	0.686	0.356	–
Qwen3-VL-2B (LensUnd)	0.445	0.747	0.397	–
Qwen3-VL-8B (LensUnd)	0.485	**0.783**	0.528	–
Chat-3D v2	0.222	0.522	–	–
ChatScene	0.227	0.546	–	–
Video-3D LLM	0.228	0.542	–	1.88
Lens3D (Video-3D LLM)	0.483	0.747	0.533	3.71
3DRS	0.230	0.549	–	1.92
Lens3D (3DRS)	**0.489**	0.752	**0.538**	**3.86**

4.1 Quantitative evaluation

LensBench construction. We construct LensBench to evaluate fine-grained target descriptions against diverse, independent references. It contains 2,068 validation objects from 141 ScanNet validation scenes, with three independent references generated by GPT 5.4 (Singh et al., 2025), Gemini 3 (Team et al., 2023), and Qwen 3.5 (Team, 2026) from views focused using ground-truth boxes. Each reference contains a summary and structured fields for visible parts, materials, and object state; Parts, Materials, and State F1 measure set-token overlap on the corresponding fields. Its scenes do not overlap the 561 ScanNet training scenes used by LensDistill. In a public questionnaire, respondents evaluate 100 randomly sampled objects and separately rate each anonymized reference against the focused views, yielding 300 ratings with a mean of 3.85/4.0 and supporting their use as high-quality silver references.

Benchmarks and metrics. We evaluate fine-grained descriptions on LensBench and retained native capabilities on the standard ScanNet (Dai et al., 2017) benchmarks ScanRefer (Chen et al., 2020), Multi3DRefer (Zhang et al., 2023), ScanQA (Azuma et al., 2022), and SQA3D (Ma et al., 2023). We report Acc@0.25/0.5 for ScanRefer, F1@0.25/0.5 for Multi3DRefer, CIDEr (Vedantam et al., 2015) for ScanQA, and exact match for SQA3D. Description metrics include token F1 and embedding similarity for summaries and set-token F1 for parts, materials, and state.

Implementation details. We instantiate students with Video-3D LLM (Zheng et al., 2025) and 3DRS (Huang et al., 2025). For 7,854 objects, LensUnd provides four target-focused views independently to GPT 5.4, Gemini 3, and Qwen 3.5, chosen for their strong general-purpose visual reasoning and complementary model priors; all successful responses are retained, yielding 23,560 structured captions. We jointly fine-tune each student's main vision–language stack for one epoch, replacing short Scan2Cap descriptions with this supervision while retaining the original grounding and QA data. Students keep their native 32-view inputs. Supplementary Section A.2 gives the complete prompt, data composition, frame-quality definition, and optimization details.

Evaluation of fine-grained object understanding. The quantitative results in Table 1 show consistent fine-grained description gains for both student architectures. LensUnd also benefits from target-conditioned evidence and stronger VLM capacity, while remaining robust to predicted target boxes. In a blind user study on 100 objects using a four-point scale, LensDistill raises the mean rating from 1.92 to 3.86 for 3DRS and from 1.88 to 3.71 for Video-3D LLM, confirming substantial improvements in perceived description quality.

Ablations study. Supplementary Table 6 reports the complete LensBench results. Wireframe marking contributes to performance, increasing Summary F1 from 0.263 to 0.370 with our selected views. With ground-truth boxes, four views, projected wireframes, and the same VLM, our bbox-voxel frame selection achieves a Summary F1 of 0.445, outperforming ExCap3D's visibility ranking (0.436) and Video-3D LLM's greedy selection used in object-level (0.353). Increasing temporal uniform sampling from 4 to 32 views yields only marginal gains, possibly since the additional views

Table 2: Comparison on standard 3D scene-understanding benchmarks. Asterisks indicate results reproduced in our local environment and evaluation stack. Lens3D uses only the corresponding base model's native input at inference. Best results in each column are highlighted in red, second-best in orange, and third-best in yellow.

Method	ScanRefer		Multi3DRefer		ScanQA	SQA3D
	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	CIDEr	EM
Task-specific models						
ScanRefer (Chen et al., 2020)	37.3	24.3	–	–	–	–
MVT (Huang et al., 2022)	40.8	33.3	–	–	–	–
3DVG-Transformer (Zhao et al., 2021)	45.9	34.5	–	–	–	–
ViL3DRel (Chen et al., 2022b)	47.9	37.7	–	–	–	–
M3DRef-CLIP (Zhang et al., 2023)	51.9	44.7	42.8	–	–	–
ScanQA (Azuma et al., 2022)	–	–	–	–	64.9	47.2
3D-VisTA (Zhu et al., 2023)	50.6	45.8	–	–	69.6	48.5
Generalist models						
3D-LLM (Flamingo) (Hong et al., 2023)	21.2	–	–	–	59.2	–
3D-LLM (BLIP2-FlanT5) (Hong et al., 2023)	30.3	–	–	–	69.4	–
Chat-3D (Wang et al., 2023)	–	–	–	–	53.2	–
Chat-3D v2 (Huang et al., 2023a)	42.5	38.4	45.1	41.6	87.6	54.7
LL3DA (Chen et al., 2024a)	–	–	–	–	76.8	–
Scene-LLM (Fu et al., 2024)	–	–	–	–	80.0	53.6
LEO (Huang et al., 2023b)	–	–	–	–	101.4	50.0
Grounded 3D-LLM (Chen et al., 2024b)	47.9	44.1	45.2	40.6	72.7	–
PQ3D (Zhu et al., 2024b)	57.0	51.2	–	50.1	–	47.1
ChatScene (Huang et al., 2024)	55.5	50.2	57.1	52.4	87.7	54.6
LLaVA-3D (Zhu et al., 2024a)	54.1	42.4	–	–	91.7	55.6
Inst3D-LMM (Yu et al., 2025)	57.8	51.6	58.3	53.5	88.6	–
3D-LLaVA (Deng et al., 2025)	51.2	40.6	–	–	92.6	54.5
Video-3D LLM* (Zheng et al., 2025)	58.2	51.8	57.4	52.1	102.1	58.5
Lens3D (Video-3D LLM)	**60.3**	**53.7**	**58.9**	**53.6**	101.2	58.2
3DRS* (Huang et al., 2025)	62.7	56.0	60.6	55.2	103.9	60.0
Lens3D (3DRS)	**63.1**	**56.4**	**61.3**	**55.9**	**104.3**	**60.1**

Table 3: ScanRefer (Chen et al., 2020) grounding on unique and multiple subsets.

Method	Unique		Multiple	
	@0.25	@0.5	@0.25	@0.5
Video-3D LLM	87.80	78.16	51.06	45.40
Lens3D (Video-3D LLM)	87.26	78.16	53.84	47.87
3DRS	86.72	77.40	56.87	50.79
Lens3D (3DRS)	86.45	77.56	57.44	51.25

provide little target-relevant evidence. Using predicted boxes retains at least 97.5% of the performance obtained with ground-truth boxes across the main metrics. Scaling the VLM from 2B to 8B further improves Summary F1 from 0.445 to 0.485 and Attr. Mean from 0.397 to 0.528. The corresponding gains for parts, materials, and states are 22.2, 9.4, and 7.6 percentage points, respectively.

Retained 3D capabilities. Table 2 shows that, under its detailed-captioning instantiation, Lens-Distill enhances the fine-grained description capability that is limited in the base models while preserving, and in some cases improving, their original grounding and question-answering performance. The gains are most apparent in grounding. Table 3 further shows that both backbones improve at both thresholds when same-class distractors are present, whereas unique-instance scores change by at most 0.54 points. This concentration of gains in challenging multi-instance scenes suggests that detailed-caption supervision improves instance discrimination while retaining standard grounding performance.

4.2 Qualitative evaluation

Beyond fine-grained object captioning, LensUnd supports small-object grounding and fine-grained object QA through task-specific instructions over focused multiview evidence. For small-object

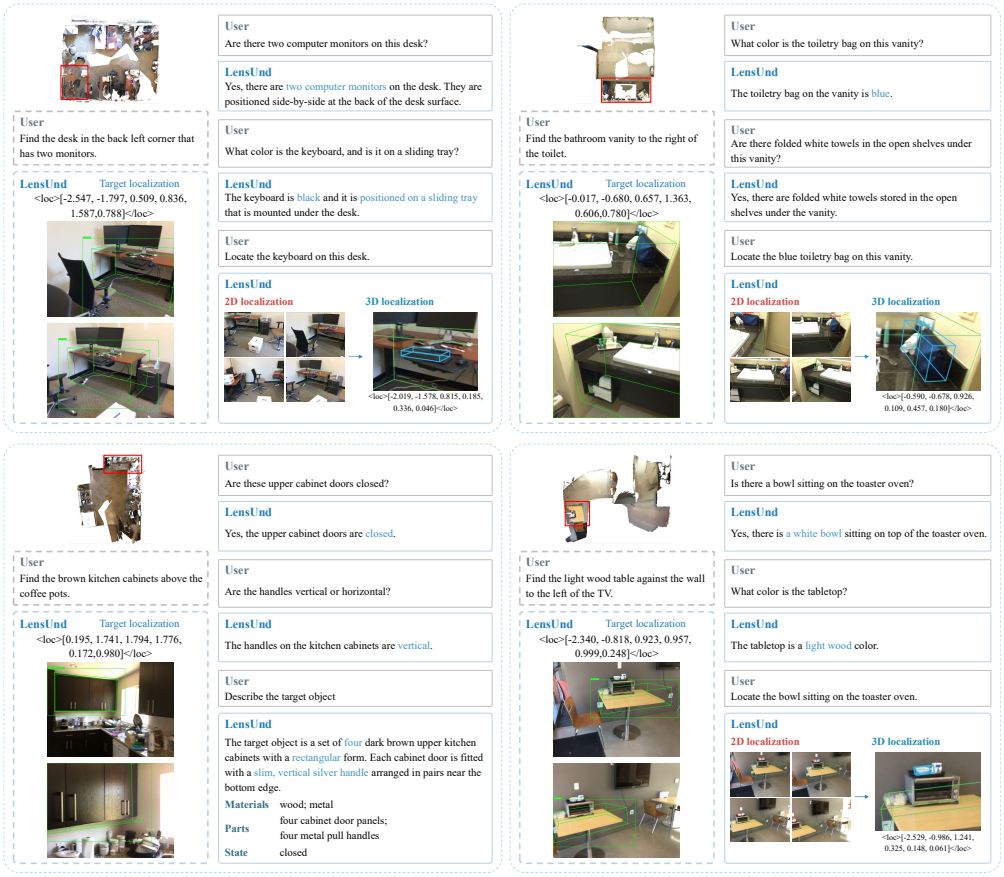


Figure 4: **More qualitative results with LensUnd.** Examples across multiple scenes demonstrate LensUnd's ability to handle varied fine-grained 3D understanding scenarios.

grounding, it focuses on a coarse anchor region, localizes a queried object in the selected views, and aggregates the unprojected observations into a 3D box. For fine-grained object QA, the external VLM reasons over the focused evidence to answer questions about object presence, appearance, and manipulation. Figures 3 and 4 present representative examples of grounding and QA.

4.3 Evidence Analysis

We use two complementary tests to establish that LensDistill uses target-level visual evidence. Masking the target region causes a substantial performance drop: Summary F1 decreases from 0.489 to 0.369, Overall F1 from 0.470 to 0.337, and Summary Sem. from 0.753 to 0.560. In the final four layers, the top 5% of image patches receive 35.0% of answer-token attention on average, seven times the uniform expectation, while attention alignment with the projected target reaches 8.96 times the uniform expectation. Together, these results show that LensDistill retrieves target appearance from native multiview observations rather than relying only on language patterns or coordinate priors. Supplementary Section A.4 provides the masking results, full protocol, and layer-wise analysis.

5 Conclusion

We presented Lens3D, a framework that combines target-conditioned visual foveation with capability transfer for fine-grained 3D understanding. LensUnd constructs focused multiview evidence that enables an external VLM to supplement existing 3D LLMs, while LensDistill transfers the acquired knowledge back into 3D LLMs operating on their native inputs. Experiments with two representative 3D LLMs show that caption-based LensDistill substantially improves fine-grained object descriptions while maintaining their native grounding and question-answering capabilities.

AI USE STATEMENT

Generative AI was used to aid polish the writing of this manuscript. It was also used to generate the synthetic detailed-description supervision for LensDistill and the silver-reference annotations in LensBench, as described in the paper. The authors reviewed the AI-assisted writing and generated data and take responsibility for the final content.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pp. 422–440. Springer, 2020.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19107–19117. IEEE, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Miguel Carvalho, Hélder Dias, and Bruno Martins. Cropvlm: Learning to zoom for fine-grained vision-language perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 11170–11179, 2026.

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. Springer, 2020.

Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D 3 net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *European Conference on Computer Vision*, pp. 487–505. Springer, 2022a.

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in neural information processing systems*, 35:20522–20535, 2022b.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11124–11133. IEEE, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26418–26428. IEEE, 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.

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.

Fangming Cui, Sunan Li, and Jiahong Li. A brief overview: On-policy self-distillation in large language models. *arXiv preprint arXiv:2605.18141*, 2026.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

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

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2(3):6, 2023a.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.

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

Junming Huang, Chi Wang, Letian Li, Guangkai Xu, Donglin Huang, Hao Chen, Qiang Dai, and Weiwei Xu. Cg-mllm: Captioning and generating 3d content via multi-modal large language models, 2026. URL https://arxiv.org/abs/2601.21798.

Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15503–15512. IEEE, 2022.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3DRS: MLLMs Need 3D-Aware Representation Supervision for Scene Understanding, 2025.

Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pp. 417–433. Springer, 2022.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

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.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19231–19242. IEEE, 2023.

Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1836–1846. IEEE, 2021.

Junliang Ye, Zhengyi Wang, Ruowen Zhao, Shenghao Xie, and Jun Zhu. Shapellm-omni: A native multimodal llm for 3d generation and understanding. *arXiv preprint arXiv:2506.01853*, 2025.

Chandan Yeshwanth, David Rozenberszki, and Angela Dai. Excap3d: Expressive 3d scene understanding via object captioning with varying detail. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21699–21709. IEEE, 2025.

Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14147–14157. IEEE, 2025.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8553–8563. IEEE, 2022.

Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15179–15179. IEEE, 2023.

Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2908–2917. IEEE, 2021.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning Position-Aware Video Representation for 3D Scene Understanding, 2025.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024a.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Zheng Huang, Mingyu Liu, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, et al. Active-o3: Empowering multimodal large language models with active perception via grpo. *arXiv preprint arXiv:2505.21457*, 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2899–2909. IEEE, 2023.

Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pp. 188–206. Springer, 2024b.

A.1 EXPERIMENTAL DETAILS

Datasets and tasks. We evaluate existing grounding and question-answering capabilities on four standard 3D scene-understanding benchmarks. ScanRefer (Chen et al., 2020) evaluates single-target 3D referring-expression grounding. Multi3DRefer (Zhang et al., 2023) extends grounding to zero-, single-, and multiple-target expressions. ScanQA (Azuma et al., 2022) and SQA3D (Ma et al., 2023) evaluate scene question answering and situated spatial reasoning, respectively. All are based on ScanNet (Dai et al., 2017) scenes, and we follow the standard splits used by prior work. Our local evaluations contain 9,508 ScanRefer examples, 11,120 Multi3DRefer examples, 4,675 ScanQA questions, and 3,519 SQA3D questions.

For object description, we replace the short-text-centered Scan2Cap (Chen et al., 2021) evaluation with LensBench. LensBench contains 2,068 targets, each with three independently generated detailed references from GPT, Qwen, and Gemini. The references cover an overall summary, materials, parts, state, visible details, and spatial relations. LensBench is used only for evaluation and is strictly separated from LensDistill training samples.

Human reference audit. We publicly distribute a questionnaire that samples 100 LensBench objects uniformly at random without replacement. Each question shows the four target-focused views and the three model identities are hidden as "Data 1–3." Respondents score each reference separately using four anchors: 1, incorrect (wrong target or mostly incorrect key facts); 2, partially correct (some visible facts but clear hallucinations or missing key evidence); 3, mostly correct (the target and main details are correct, with only minor inaccuracies); and 4, fully correct (consistent with the images and no evident hallucination). Each reference receives one independent rating, yielding a mean of 3.85/4.0 across 300 judgments.

Metrics. For ScanRefer (Chen et al., 2020), we report 3D-box grounding accuracy at IoU thresholds 0.25 and 0.5 (Acc@0.25/0.5). Multi3DRefer (Zhang et al., 2023) uses F1 at the same thresholds. ScanQA (Azuma et al., 2022) reports CIDEr (Vedantam et al., 2015), and SQA3D (Ma et al., 2023) reports exact match (EM). On LensBench, each prediction is compared with all three references and the scores are averaged. Our principal metrics are token F1 for the target summary (Summary F1) and semantic similarity (Summary Sem.), where the latter is cosine similarity between all-MiniLM-L6-v2 sentence embeddings (Reimers & Gurevych, 2019). Summary F1 measures lexical agreement, whereas semantic similarity is more robust to synonymous paraphrases.

Base models and training. We use Video-3D LLM (Zheng et al., 2025) and 3DRS (Huang et al., 2025) as two independent base models. Both construct 3D-aware visual representations from calibrated multiview RGB-D observations and report results on the same indoor-scene tasks. The student uniformly samples 32 views per scene; LensUnd selects four target views with bbox-voxel. An external VLM teacher generates LensDistill supervision from the target-conditioned LensUnd evidence under a detailed-description instruction. Training retains each base model's ScanRefer (Chen et al., 2020), Multi3DRefer (Zhang et al., 2023), ScanQA (Azuma et al., 2022), and SQA3D (Ma et al., 2023) data and replaces its short Scan2Cap (Chen et al., 2021) descriptions with detailed supervision.

A.2 LENSDISTILL REPRODUCIBILITY DETAILS

Teacher supervision and split. We start from 7,875 targets in the ScanRefer training split and successfully annotate 7,854 objects from 561 ScanNet training scenes. For every target, bbox-voxel selects four 1296×968 RGB frames using its ground-truth box and overlays the projected 3D wireframe. GPT 5.4 (Singh et al., 2025), Gemini 3 (Team et al., 2023), and Qwen 3.5 (Team, 2026) are queried independently with the same evidence and provider-default decoding settings. We retain each successful response as a separate training example rather than merging teacher outputs, producing 23,560 captions. LensBench contains 6,204 references for 2,068 objects from 141 ScanNet validation scenes; its scene set has zero overlap with the 561 training scenes.

Generation prompt. The teacher is instructed as follows: *You will be shown four screenshots of the same target object from a ScanNet scene. In every image, the target is marked by a bounding box. Describe only its visible local appearance and small objects directly attached to, placed*

on, or supported by it. Report a one–two sentence target summary, visible details, attached or supported objects, parts, materials, state, local relations, and uncertain evidence. Exclude room-level and long-range spatial relations; do not infer invisible properties or guess. Use an empty list when a field is not visible and return strict JSON only. The required keys are `target_summary`, `visible_details`, `attached_or_supported_objects`, `parts`, `materials`, `state`, `local_relations`, and `uncertain`.

Multitask composition and optimization. One epoch contains all 79,445 SQA3D, 26,515 ScanQA, 36,665 ScanRefer, and 43,838 Multi3DRefer examples, together with 23,560 detailed-caption examples; hence task sampling is proportional to these counts, with detailed captions comprising 11.2% of the 210,023-example mixture. We use AdamW with a global batch size of 16, peak learning rate 10^{-5}, vision-tower learning rate 2×10^{-6}, zero weight decay, 3% warmup, and cosine decay. Training uses bfloat16, gradient checkpointing, and ZeRO-3 for 13,125 optimizer steps on eight NVIDIA H20 GPUs.

Updated modules and student input. We perform full-parameter tuning of the main vision–language stack rather than LoRA: the vision tower, multimodal projector, language model, spatial positional embeddings, and grounding head are updated, while the vision resampler remains frozen. The student uniformly samples 32 native scene views and applies the base models' SigLIP patch-14 preprocessing with 384-pixel base tiles and `anyres_max_9`; the four focused teacher views and wireframes are never exposed to the student.

Frame-quality score. Candidates are evaluated every eight frames and must have projected target-area ratio $r \geq 0.002$. Let c be the fraction of front-facing box corners inside the image, i the fraction of the unclipped projected rectangle inside the image, m its normalized minimum edge margin, d the projected-center proximity to the image center, and u indicate whether the projected target-region center lies inside the image. We define

$$a(r) = \begin{cases} \min(r/0.12, 1), & r \leq 0.12, \\ 1, & 0.12 < r \leq 0.35, \\ \max(0, 1 - (r - 0.35)/0.35), & r > 0.35, \end{cases} \tag{7}$$

$$f = 0.50i + 0.35a(r) + 0.15m, \qquad g = 0.20c + 0.25f + 0.30d + 0.25u. \tag{8}$$

We multiply g by 0.35 when $u = 0$ and by 0.50 when the projected 3D box center is outside the image. For the best $40K = 160$ geometric candidates, depth consistency uses tolerance $\tau = \max(0.3\,\mathrm{m}, 0.15 z_{\min})$. Let o be the fraction of ROI depth pixels closer than $z_{\min} - \tau$, h the fraction within $[z_{\min} - \tau, z_{\max} + \tau]$, and $v = h(1 - o)$. A candidate is depth-visible when $o < 0.9$ and $h > 0.15$, and its final quality is

$$s_t = (0.85g + 0.15v)\, p(o), \qquad p(o) = \begin{cases} 1, & o \leq 0.5, \\ \max(0, 1 - o^2), & o > 0.5. \end{cases} \tag{9}$$

If no candidate passes the depth test, geometric candidates are retained with $v = p = 0.5$. Bbox-voxel then uses $K = 4$, a 12-frame minimum gap, 12 voxel cells along the shortest box axis, depth subsampling by four, and a 2% box margin.

A.3 View-coverage comparison with ExCap3D

We compare the joint target-region coverage of LensUnd and the visibility-based view selection used by ExCap3D on all 2,068 LensBench targets. Both methods use the same budget of four views and the same voxelization. For each frame, depth pixels are unprojected into 3D, and only occupied voxels inside the target box are retained. Joint coverage is the number of unique voxels in the union of the four selected views. We also report inter-view redundancy,

$$1 - \frac{\left|\bigcup_{t \in \mathcal{S}} \mathcal{V}_t\right|}{\sum_{t \in \mathcal{S}} |\mathcal{V}_t|}, \tag{10}$$

where a lower value indicates less repeated evidence across views.

LensUnd covers 31.5% more unique voxels on average and obtains higher coverage on 97.1% of targets, while reducing redundancy by 9.8 percentage points. ExCap3D ranks frames independently

Table 4: Target-region view coverage under the same four-view budget. Union-normalized coverage uses, for each target, the union of voxels observed by either method as the reference set; it is not absolute full-surface coverage.

Metric	LensUnd	ExCap3D
Unique covered voxels	**2,057.7**	1,564.6
Union-normalized coverage (%)	**90.3**	63.6
Inter-view redundancy (%) ↓	**38.0**	47.8

by visible-vertex fraction, so high-ranking frames may be temporally close and observe nearly identical target surfaces. In contrast, LensUnd greedily selects views according to their marginal voxel contribution, yielding more complementary evidence. Higher joint coverage exposes the target from a wider range of viewpoints, making otherwise occluded sides, parts, materials, and state cues more likely to be observed. It also uses the limited view budget more efficiently by reducing repeated evidence, thereby providing the VLM with a more complete basis for fine-grained target understanding.

Relation to ExCap3D. ExCap3D focuses on object- and part-level captioning with dedicated captioners, while Lens3D uses a shared target-focused evidence interface across fine-grained object captioning, fine-grained object QA, and small-object grounding. LensDistill further uses the resulting detailed captions to train pretrained generalist 3D LLMs together with their native grounding and scene-level QA data. The two approaches thus differ in task scope, model target, and view-selection strategy.

A.4 Visual-dependence analysis

Target-region masking. The student input contains no projected wireframe, only 32 RGB views and the `<coord>` token encoding the target center. We project the ground-truth oriented 3D box into every RGB view and replace pixels inside its 2D enclosing rectangle with gray. This preserves scene context, the view sequence, and target-coordinate encoding while removing most target appearance. As shown in Table 5, masking substantially reduces lexical, overall, and semantic description scores, confirming that generation depends materially on target pixels.

Table 5: Target-region masking for Lens3D (3DRS).

Input	Summary F1	Overall F1	Summary Sem.
Unmasked	0.489	0.470	0.753
Target masked	0.369	0.337	0.560

Attention concentration and target alignment. We analyze attention from language-model query tokens to visual tokens during target-summary generation. Attention concentration is the mass assigned to the top 5% of image patches. Target alignment is the ratio between attention within the projected target and its expectation under a uniform distribution. In the final four layers, the top 5% of patches receive 35.0% of answer-token attention on average, and every sample exceeds five times the uniform expectation.

Target alignment rises systematically with depth. Answer-token alignment increases from 1.29 times uniform expectation in layers L0–3 to approximately four to five times in L12–15 and 8.96 times in L24–27. The `<coord>` token retains approximately 4.47 times target alignment in deep layers despite an attention-sink effect at non-target frame positions. The masking and attention results jointly show that LensDistill uses `<coord>` to retrieve target appearance from the native multiview sequence.

A.5 Attribute-level LensBench analysis

In addition to summary quality, Table 6 reports lexical set agreement for parts, materials, and state. We use multi-reference mean set-token F1: each predicted field is compared independently with each of the three references, after which results are averaged over references and samples. The original 3DRS, Video-3D LLM, Chat-3D v2, and ChatScene outputs are free-form and do not pro-

Table 6: Full LensBench evaluation. "Mean" averages Parts, Materials, and State F1. Temporal and quality-pool denote uniform sampling from the full sequence and from a quality-filtered pool, respectively. Best and second-best results within each block are bolded and underlined, respectively.

Type	Model	Box source	View selection	Views	Wire.	Sum. F1	Sum. Sem.	Parts	Materials	State	Mean
LensUnd	Qwen3-VL-2B (Bai et al., 2025)	Ground truth	Temporal uniform	4	No	0.263	0.619	0.192	0.607	0.272	0.357
LensUnd	Qwen3-VL-2B	Ground truth	Temporal uniform	4	Yes	0.370	0.694	0.218	0.591	0.275	0.362
LensUnd	Qwen3-VL-2B	Ground truth	Temporal uniform	32	Yes	0.380	0.686	0.201	0.589	0.278	0.356
LensUnd	Qwen3-VL-2B	Ground truth	Quality-pool uniform	4	Yes	0.426	0.732	0.245	0.628	0.300	0.391
LensUnd	Qwen3-VL-2B	Ground truth	Video-3D LLM greedy	4	Yes	0.353	0.690	0.215	0.594	0.277	0.362
LensUnd	Qwen3-VL-2B	Ground truth	ExCap3D visibility	4	Yes	0.436	0.742	0.250	<u>0.634</u>	0.299	0.394
LensUnd	Qwen3-VL-2B	Ground truth	bbox-voxel	4	Yes	<u>0.445</u>	<u>0.747</u>	<u>0.252</u>	0.632	<u>0.308</u>	<u>0.397</u>
LensUnd	Qwen3-VL-2B	3DRS	bbox-voxel	4	Yes	0.435	0.739	0.244	0.625	0.305	0.391
LensUnd	Qwen3-VL-2B	Video-3D LLM	bbox-voxel	4	Yes	0.434	0.736	0.249	0.630	0.302	0.393
LensUnd	Qwen3-VL-8B	Ground truth	bbox-voxel	4	Yes	**0.485**	**0.783**	**0.474**	**0.726**	**0.384**	**0.528**
3D LLM	3DRS	Ground truth	Temporal uniform	32	No	0.230	0.549	–	–	–	–
3D LLM	**Lens3D (3DRS)**	Ground truth	Temporal uniform	32	No	**0.489**	**0.752**	**0.477**	**0.709**	**0.429**	**0.538**
3D LLM	Video-3D LLM	Ground truth	Temporal uniform	32	No	0.228	0.542	–	–	–	–
3D LLM	**Lens3D (Video-3D LLM)**	Ground truth	Temporal uniform	32	No	<u>0.483</u>	<u>0.747</u>	<u>0.473</u>	<u>0.698</u>	<u>0.428</u>	<u>0.533</u>

vide corresponding attribute fields. We therefore omit their attribute scores rather than compare full-response tokens against structured fields under a different extraction procedure.

After LensDistill, the three-field means reach 0.538 and 0.533 for 3DRS and Video-3D LLM, respectively. Under identical LensUnd evidence, Qwen3-VL-8B improves over the 2B model by 22.2, 9.4, and 7.6 percentage points on parts, materials, and state; its Materials F1 of 0.726 is the highest in the table. Scores remain similar across box sources. Increasing temporal sampling from four to 32 views reduces the three-field mean from 0.362 to 0.356, consistent with the summary results.