Qwen-3D: A Generalist 3D Vision-Language Model for Spatial Understanding

Anonymous CVPR submission

Paper ID 00027

Abstract

Recent Large Multimodal Models (LMMs) achieve impressive performance on images and short videos, but long video reasoning remains computationally expensive and temporally inconsistent due to frame-level tokenization and limited context windows. We observe that 3D geometry provides a natural compression mechanism for multi-view visual streams. Geometric signals such as depth and camera pose allow RGB frames to be fused into persistent world-aligned representations, enabling efficient reasoning over space and time. Motivated by this insight, we introduce Qwen-3D, a geometry-aware LMM that leverages multi-view geometric signals to compress visual tokens within the Qwen backbone, enabling efficient processing of long and highly redundant video sequences. By modifying visual tokens with 3D Rotary Positional Embedding, Qwen-3D performs attention in the world space rather than over independent image frames, enabling efficient cross-view reasoning. Qwen-3D further integrates a query-based segmentation decoder that grounds language tokens directly in visual space, allowing unified reasoning across referential grounding, instance segmentation, and visual question answering for both images and videos. Across a broad suite of benchmarks, Qwen-3D outperforms large-scale proprietary 2D models and existing 3D LMM approaches. Finally, we show that joint training on both 2D and 3D data preserves strong 2D vision–language capabilities while substantially improving 3D reasoning.

1. Introduction

Current Vision–Language Models (VLMs) perform well on images and short video clips, but struggle with long multi-view video streams. Processing long sequences is computationally expensive due to the quadratic cost of attention, and limited context windows prevent consistent spatio-temporal and 3D reasoning across frames. Multi-view 3D geometry, in the form of depth and camera poses, offers a more principled alternative, allowing video frames to be mapped into a shared 3D coordinate system. This enables compression of multi-view streams into persistent scene representations where temporally distant frames correspond to nearby 3D locations.

Existing approaches to integrating 3D spatial awareness into VLMs generally follow one of two distinct paradigms. One line of work introduces 3D point clouds as auxiliary inputs to the model [11, 19, 20], either alongside or in place of multi-view images. While these approaches expose explicit geometric structure, they treat point clouds as a modality separate from the visual tokens, overlooking the fact that point cloud features are inherently aligned with image features with corresponding depth. The second line of work integrates geometry directly into the visual token representation. Methods such as LLaVA-3D[54] and Video-3D-LLM[53] modify positional encodings such that multi-view image tokens are embedded according to their 3D world coordinates rather than their 2D image-plane positions. This approach allows the model to reason over multi-view observations in a shared spatial coordinate system while maintaining the strong visual representations learned by large VLM backbones.

Despite these advances, existing 3D large multimodal models (LMMs) still lag behind specialist 3D vision systems. Dedicated models trained for tasks such as detection, segmentation, or grounding continue to outperform general-purpose 3D LMMs [22, 23, 55]. Moreover, most current 3D LMMs do not attempt 3D object detection on standard benchmarks such as ScanNet [14, 37]. The only exception, Grounded-3D-LLM [12], achieves less than half the performance of state-of-the-art 3D detectors. These results highlight a key tradeoff: while general-purpose 3D LMM's offer a broad range of reasoning capabilities, they currently lack the specialized grounding mechanisms of dedicated 3D systems, preventing them from matching—or in many cases, even addressing—the high-level performance required for core 3D perception.

To bridge these performance gaps, we introduce Qwen-3D, a geometry-aware 3D LMM that extends the Qwen family of models [5] with explicit mechanisms for multi-view reasoning and object grounding. Built on the strong 2D foundation of Qwen2.5-VL-3B, Qwen-3D integrates

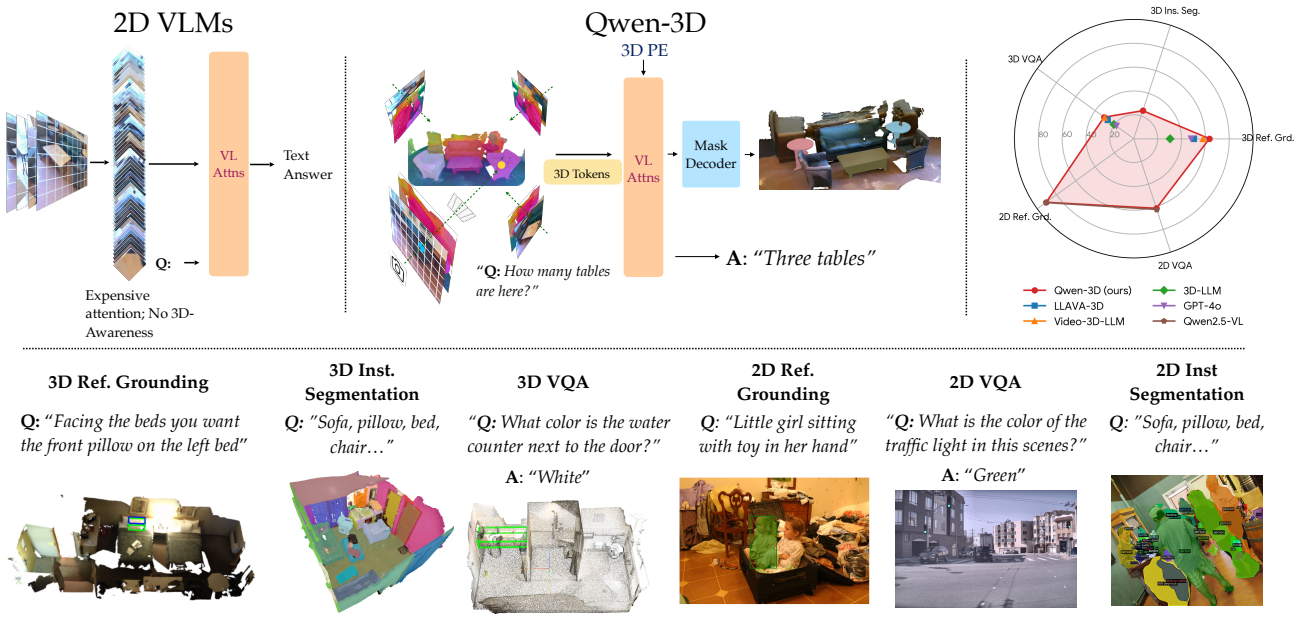


Figure 1. **Qwen-3D replaces frame-level attention with geometry-aligned world-space attention using 3D positional embeddings.** Unlike conventional 2D VLMs that rely on spatio-temporal attention over video frames, Qwen-3D builds a geometry-aware 3D representation from multi-view RGB, depth, and camera poses. This enables attention in the world space via 3D Rotary Positional Embeddings. Qwen-3D supports joint decoding of text responses and 2D/3D object masks. Qwen-3D achieves state-of-the-art performance across a multitude of vision–language tasks from images and multi-view videos.

3D structure directly into the vision–language backbone. First, we leverage geometric information to compress visual tokens within the backbone, merging tokens originating from nearby 3D locations. This enables efficient reasoning over long multi-view videos while preserving spatial consistency across views. Second, we introduce 3D Rotary Positional Embeddings to the Qwen backbone, allowing attention within the backbone to operate in a geometry-aware coordinate system and improving cross-view spatial reasoning.

In addition, Qwen-3D incorporates a 3D query-based segmentation decoder that grounds noun phrases from the language stream directly in world space. Unlike prior approaches that communicate between the VLM backbone and object decoder through a small set of query vectors, our model shares full visual token representations between the backbone and decoder, enabling richer visual–language interactions and stronger grounding performance.

Across a wide range of 3D grounding benchmarks, Qwen-3D outperforms both proprietary 2D VLMs and the strongest existing 3D LMMs while maintaining strong performance on 2D tasks. Compared to the previous 3D LMM state-of-the-art, Qwen-3D improves 3D visual grounding by 4%, surpasses the best single-stage 3D LMMs by 12%, and increases 3D instance segmentation accuracy by 12%. The model also achieves state-of-the-art performance on 3D visual question answering while preserving the strong 2D capabilities of its backbone. Furthermore, Qwen-3D substantially narrows the gap between general-purpose 3D LMMs and specialist 3D grounding models on in-domain benchmarks, while significantly outperforming specialist methods on out-of-distribution 3D scenes and language instructions.

Contributions. Our contributions are as follows:

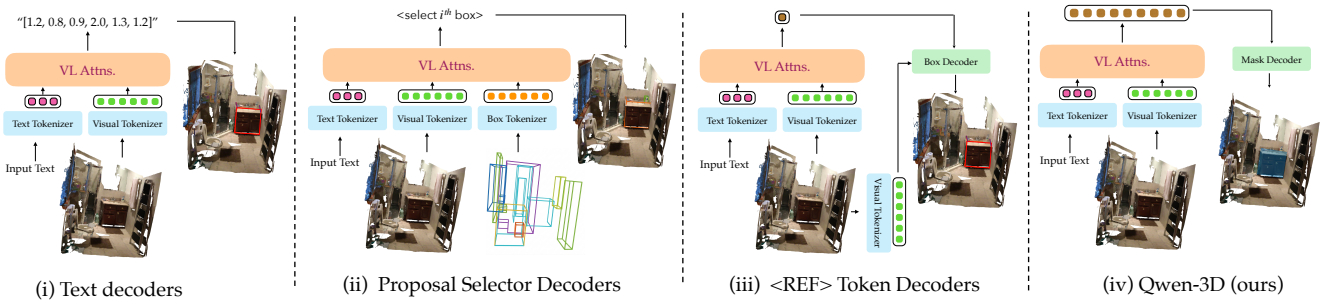
- A geometry-aware VLM backbone that integrates multi-view structure through 3D rotary positional embeddings and geometry-based token compression.
- A unified 3D grounding architecture with a query-based segmentation decoder that grounds language directly in world space and shares full visual token representations with the VLM backbone.
- A general-purpose 3D multimodal training framework that jointly learns from 2D and 3D data, preserving strong vision–language capabilities while improving 3D spatial reasoning and grounding.
- State-of-the-art performance among 3D-LMM methods across 3D benchmarks, improving 3D visual grounding by 4%, instance segmentation by 12%, while maintaining strong 2D multimodal performance.

We will open-source our model upon acceptance.

2. Related Work

3D Large Vision-Language Models Building upon the rapid progress of 2D Large Multimodal Models [5, 35, 42], recent work has focused on extending these models to understand 3D scenes. Existing approaches can be grouped into four main categories:

CVPR
#00027



(i) Text decoders (ii) Proposal Selector Decoders (iii) <REF> Token Decoders (iv) Qwen-3D (ours)

Figure 2. **Decoding object information in 3D LMMs.** From left to right: (i) Direct bounding box decoding in text space, as in 3D-LLM [17]; (ii) Two-stage grounding via proposal selection, as in Video-3D-LLM [19, 20, 53]; (iii) Special token decoding, as in LLaVA-3D [12, 54]; (iv) (Ours) Directly connecting contextualized vision–language features from the VLM backbone to a mask-based segmentation decoder

(a) 3D point cloud encoders trained from scratch. Methods such as LL3DA [11], Scene-LLM [16], and Grounded-3D-LLM [12] augment multi-view image streams with explicit 3D point cloud encoders. The resulting 3D features are projected into an LLM backbone in addition to or as a replacement for 2D image features. These approaches, however, require large-scale point cloud–language datasets for alignment—an acute limitation given the scarcity of 3D data. In contrast, Qwen-3D builds upon powerful 2D pretrained features and augments them with 3D information via positional encodings.

(b) Learnable 3D feature compression. 3D-LLM [17] uses Q-Former layers [28] to compress large numbers of 2D foundation-model features into small sets of latent tokens. In contrast, Qwen-3D performs multi-view feature compression in a parameter-free manner, directly guided by the 3D spatial layout of the tokens.

(c) Object-centric approaches. Another line of work uses object-level features as input to the language model. Approaches such as LEO [20] and ChatScene [19] first detect objects with off-the-shelf 2D or 3D detectors, pool features within detected regions, then feed these pooled features into their VLMs. These methods may yield structured object representations and improve grounding, but performance is fundamentally constrained by the robustness of the detectors themselves, which often struggle due to limited data diversity. In contrast, Qwen-3D is a single-stage model that directly grounds the language in the 3D visual stream.

(d) Positional embedding adaptation. Several methods modify the positional embeddings of multi-view visual tokens to better encode 3D spatial relationships [22, 23, 53, 54]. Our model follows this general paradigm. Similar to these models, we incorporate 3D information via positional embeddings in the vision-language attention. We utilize 3D Rotary Positional Encoding for this purpose.

Beyond spatial encoding, Qwen-3D also differs from prior work in how grounded outputs are decoded. We discuss these grounding architectures in detail in the following section.

3D Visual Grounding. Visual grounding—identifying objects referred to by language—is a fundamental capability for 3D vision–language systems. Early work [3, 21, 33] achieved strong performance by designing specialized architectures tailored for 3D grounding. Subsequent methods [3, 23, 55, 56] unified grounding, question answering, and captioning within a single framework. More recently, 3D LMMs have leveraged large-scale pretrained vision–language features to assist grounding in 3D scenes; our method follows this paradigm. These newer approaches typically adopt one of three designs (Fig. 2):

(a) Direct bounding-box decoding. (Fig. 2a) Models such as 3D-LLM [17] directly decode 3D bounding boxes in the text space. However, they achieve low performance on localization tasks, likely due to the scarcity of 3D-language data and the unstructured nature of 3D scenes.

(b) Two-stage grounding via proposal selection. (Fig. 2b) Approaches such as Video-3D-LLM [53], ChatScene [19], and LEO [20] first run an object detector to generate candidate proposals, then select the object that best matches the query. While this improves grounding robustness, performance is bottle-necked by the quality of the proposals.

(c) Special token decoding. (Fig. 2c) Other methods decode special grounding tokens (e.g. $<$ REF $>$) and localize them using explicit decoder heads, such as in LLAVA-3D [54] and Grounded-3D-LLM [12]. Essentially, the VLM heads and the decoder heads are only connected via the generated $<$ REF $>$ tokens. Although these methods can, in principle, ground multiple instances from a category, only Grounded-3D-LLM has been applied to 3D object detection. We include ablations on the $<$ REF $>$ token design with our architecture in our experiments.

Rather than relying on text-space bounding boxes, proposal selection, or grounding tokens, **Qwen-3D directly connects contextualized vision–language features from**

the VLM backbone to a mask-based segmentation decoder (Fig. 2d). This avoids the bottleneck imposed by the special token decoding and aligns more naturally with contemporary multi-object detection and segmentation architectures. As a result, Qwen-3D achieves state-of-the-art grounding and detection performance among 3D LMMs.

3. Method

We introduce Qwen-3D, a geometry-aware extension of Qwen2.5-VL [5] that enables efficient reasoning over multi-view 3D scenes. Given multi-view RGB-D frames, camera poses, and a language query, Qwen-3D projects visual features into a 3D feature cloud and compresses redundancies via voxel-based token merging. A vision-language backbone processes these geometry-aligned tokens using 3D Rotary Positional Embeddings for spatial reasoning in world coordinates. Finally, a mask-based grounding decoder predicts 3D segmentation masks for queried objects alongside text responses for question-answering. Notably, Qwen-3D preserves the original 2D input–output interface of Qwen2.5-VL while extending it to operate on multi-view 3D scenes. Figure 3 illustrates the architecture, detailed below.

Geometry-Aware Visual Encoding We construct a 3D feature cloud by projecting multi-view RGB features into world coordinates using depth maps and camera poses. This enables geometry-aware token compression based on spatial proximity. We use the Qwen2.5-VL pre-trained ViT encoder to process RGB images into feature maps $\mathcal{F}$, downsampled by a factor of S to a size of $N \times \frac{H}{S} \times \frac{W}{S} \times D$, where D is the feature dimension. We then unproject the pixel-aligned depth maps using camera intrinsics and poses, nearest-downsampling them to match this spatial resolution and yield 3D coordinates $\mathcal{C}$ of size $N \times \frac{H}{S} \times \frac{W}{S} \times 3$.

Geometry-based token compression Multi-view observations inherently produce redundant tokens at overlapping 3D locations. To mitigate memory bottlenecks, we apply voxel-based token merging (with a voxel size of 5cm) to the feature-coordinate pairs, following [22, 23, 54]. This process discretizes the space and mean-pools the features and coordinates within each occupied voxel, yielding a compact, unordered set of geometry-aligned tokens $\mathcal{FC}' = \{(f_j, p_j)\}_{j=1}^{M'}$, where $M' \leq M$.

Vision–Language Attention with 3D RoPE We use the Qwen2.5-VL language tokenizer [5] to embed the input natural language query into a sequence of tokens $\mathcal{T} = \{t_k\}_{k=1}^{L}$, where $t_k \in \mathbb{R}^D$ and L is the number of tokens. The concatenated sequence of voxelized 3D point features and text tokens is then processed by N pre-trained multimodal attention layers from Qwen2.5-VL. To enable spatial reasoning, we adapt Qwen's Multimodal RoPE—originally designed for 1D or 2D grids—to encode 3D world coordinates (x, y, z), resulting in four positional components

$$\mathrm{PE}_{3D}(\mathbf{p}) = [\mathrm{PE}(t); \mathrm{PE}(x); \mathrm{PE}(y); \mathrm{PE}(z)].$$

where t represents a token's temporal position in the sequence and x, y, z denote 3D spatial coordinates. For text, the temporal ID increments sequentially (reducing to standard 1D RoPE), while point cloud tokens share a constant temporal ID and use their world coordinates as spatial IDs.

This embedding defines a rotation matrix $R(\mathbf{p})$ applied to query and key vectors before attention: $\tilde{\mathbf{q}} = R(\mathbf{p}_q)\mathbf{q}$, $\tilde{\mathbf{k}} = R(\mathbf{p}_k)\mathbf{k}$

Because 3D point clouds are permutation-invariant, we replace Qwen2.5-VL's autoregressive causal masking with full attention over visual tokens. This improves 3D performance without degrading the backbone's original capabilities. Further attention masking details and RoPE ablations are provided in the appendix and experiments section, respectively.

U

Grounding Decoder Unlike prior VLM grounding architectures that communicate with object decoders via a small set of query tokens [12, 54], our approach directly attends to the full set of contextualized visual tokens from the backbone, enabling richer language-visual interactions. We adapt a mask decoder design similar to UniVLG [23] [23]. We instantiate N learnable object queries that cross-attend to the language and visual tokens from the VLM backbone, followed by self-attention among the queries. For all attention involving visual tokens, we 2D or 3D positional embeddings corresponding to the input modality. Finally, we decode segmentation masks via a dot product between object queries and the updated visual tokens and identify grounded object text spans via dot products with the language tokens. For question answering, we use the Qwen2.5-VL text-generation head to perform next-token prediction over the contextualized scene-language features.

Training Objectives We supervise Qwen-3D on three losses: **(a) Mask loss:** We assign predictions to ground-truth instances via Hungarian matching [8] and supervise matched masks with Binary Cross-Entropy (BCE) and Dice loss, following Mask2Former [13].
(b) Span alignment loss: As in [21, 24, 29], we supervise each predicted text span with the corresponding matched ground-truth text span using Binary Cross Entropy loss. The queries that remain unmatched are supervised to predict low probability over all text tokens.
(c) Text generation loss: For question-answering and captioning tasks, we use a token-level cross-entropy on the generated answer.
Our complete loss is formulated as:

$$\mathcal{L} = \alpha_{\text{mask}}\,\mathcal{L}_{\text{mask}} + \alpha_{\text{span}}\,\mathcal{L}_{\text{span}} + \alpha_{\text{gen}}\,\mathcal{L}_{\text{gen}} \qquad (1)$$

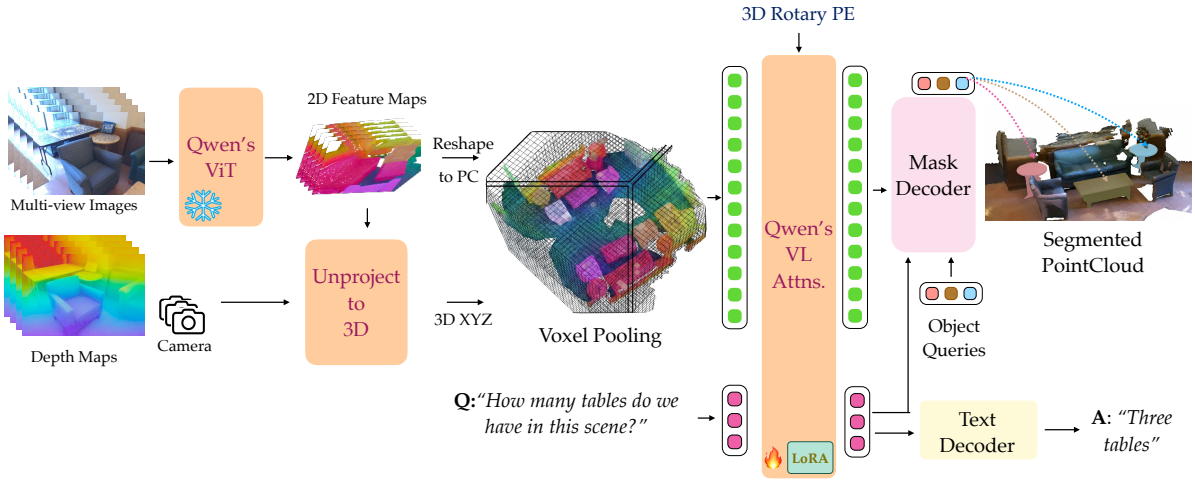


Figure 3. **Qwen-3D architecture.** Given a natural language query and multi-view RGB-D inputs, the Qwen2.5-VL vision encoder extracts multi-view 2D features, unprojects them into world-space XYZ coordinates, and voxel-pools them to reduce spatial redundancy. The resulting tokens are fused with text via Qwen vision–language attention layers augmented with 3D Rotary Positional Embeddings and LoRA adaptation. Two task heads operate on the shared tokens: a 3D mask decoder for referential grounding and instance segmentation and a text decoder for open-ended question answering.

where $\mathcal{L}_{\text{mask}}$ is the mask loss, $\mathcal{L}_{\text{span}}$ is the span alignment loss, $\mathcal{L}_{\text{gen}}$ is the text generation loss, and the α's are the loss weight terms.

Joint 2D-3D Training Qwen-3D shares parameters across 2D and 3D modalities, natively processing single or multiple RGB images as well as multi-view posed RGB-D frames. Within both the VLM backbone and the grounding decoder, 2D visual tokens are structured on a regular grid with 2D positional embeddings, whereas 3D inputs are represented as unordered point tokens with 3D embeddings. To improve alignment between 2D and 3D representations during training, we lift 2D data to 3D with probability p using reconstruction models such as MoGE [43].

Implementation Details Qwen-3D introduces only ∼50M trainable parameters. We freeze the Qwen2.5-VL ViT backbone, fine-tune the vision–language attention layers via LoRA [18], and train the mask decoder from scratch. We train jointly on on 2D and 3D datasets runs for 200k iterations (learning rate 10^{-4}) on eight 48GB L40S GPUs with an effective batch size of 8, which takes approximately three days. Text-generation loss is applied exclusively to captioning and question-answering tasks; for detection, we construct prompts [21] by concatenating object class names (e.g., "find chair. table. sofa.").

Following [22, 23], we subsample 15 frames per scene during training. At inference time, we feed all posed RGB-D frames to our model (∼90 on average for ScanNet), which takes about 2 seconds per scene end-to-end. Our voxel-pooling strategy is critical for enabling this scalability - without it, the model runs out of memory even with substantially fewer input frames. Finally, due to computational constraints, we adopt the 3B VLM backbone variant rather than the 7B models common in prior 3D-LMMs [53, 54]. Additional hyperparameters are detailed in the appendix.

4. Experiments

We evaluate Qwen-3D against existing LMMs and specialized 3D vision models on visual grounding in both in-domain (Sec. 4.1) and out-of-domain settings (Sec. 4.2), along with 3D instance segmentation (Sec. 4.3) and 3D VQA (Sec. 4.4). We also assess how well Qwen-3D retains its 2D multimodal capabilities (Sec. 4.5) and analyze which design choices most significantly impact performance (Sec. 4.6). Qualitative results, failure mode analysis, robustness to depth and camera pose noise, and performance on text-only tasks are provided in the appendix.

Training Datasets We train jointly on a mixture of 3D and 2D datasets to enable 3D comprehension while preserving the base model's pre-trained capabilities. The 3D datasets include referential grounding (SR3D, NR3D [1], ScanRefer [10]), instance segmentation (ScanNet200 [37], Matterport [9]), and question answering (ScanQA [4], SQA3D [34]). To mitigate catastrophic forgetting of Qwen's original capabilities, we co-train on 2D datasets: referential grounding (RefCOCO, RefCOCO+, RefCOCOg [25]), instance segmentation (COCO [30]), captioning and QA (LLaVA-Instruct-150k [31]), and instruction fine-tuning (Alpaca [41]).

Table 1. **Results on 3D Visual Grounding and QA for both experts and LMMs.** *2D VLM numbers obtained from prior works [3, 40]

	Methods	ScanRefer Acc @25 (Det)	ScanRefer Acc @50 (Det)	ScanQA (Val) EM@1	ScanQA (Val) C	ScanQA (Val) M	ScanQA (Val) R	SQA3D (Test) EM@1
Expert Models	BUTD-DETR [21]	52.2	39.8	-	-	-	-	-
	ScanQA [4]	-	-	23.5	67.3	13.6	34.3	-
	PQ3D [56]	56.7	51.8	20.0	65.2	13.9	-	47.1
	3D-VisTA [55]	51.0	46.2	22.4	69.6	13.9	35.7	48.5
	ODIN [22]	43.1	33.4	-	-	-	-	-
	Locate-3D [3]	61.1	50.9	-	-	-	-	-
	UniVLG [23]	63.5	56.4	25.7	78.5	15.2	40.0	50.2
LMMs	**2D VLMs***							
	LLaMA [2]	37.3	24.2	-	-	-	-	-
	GPT-4o [35]	48.2	32.6	18.0	58.3	14.2	33.4	-
	Two-Stage							
	ChatScene [19]	55.5	50.2	21.6	87.7	18.0	41.6	54.6
	LEO [20]	-	-	24.5	101.4	20.0	49.2	50.0
	Video-3D LLM - 7B [53]	58.1	51.7	30.1	102.1	-	-	**58.6**
	Single-Stage							
	NaviLLM [52]	-	-	23.0	75.9	15.4	38.4	-
	LL3DA [11]	-	-	-	76.8	15.9	37.3	-
	SceneLLM [16]	-	-	27.2	80.0	16.6	40.0	54.2
	3D-LLM [17]	30.3	-	20.5	69.4	14.5	35.7	-
	Grounded 3D-LLM [12]	48.6	44.0	-	75.9	-	-	-
	LLaVA-3D - 7B [54]	50.1	42.7	27.0	**103.1**	20.8	**49.6**	55.6
	Qwen-3D - 3B (Ours)	**65.6**	**57.7**	**30.5**	95.2	**38.3**	46.9	56.0

4.1. Evaluation on 3D Referential Grounding

Datasets. We evaluate on the validation sets of three ScanNet-based [14] 3D referential grounding benchmarks: SR3D, NR3D [1], and ScanRefer [10]. While SR3D comprises 88k synthetic utterances, NR3D (41k) and ScanRefer (51k) feature complex, human-annotated queries. Following recent work [3, 7, 22, 23], we operate directly on noisy, raw sensor RGB-D point clouds rather than clean, post-processed meshes. Although this setup introduces sensor-mesh misalignments that can degrade performance [22, 23], it better reflects practical embodied learning scenarios.

Evaluation Metrics. We report standard Top-1 accuracy, where a prediction is correct if the highest-confidence predicted bounding box achieves an Intersection over Union (IoU) with the ground-truth box above a threshold (0.25, 0.5). Since our model predicts segmentation masks, we convert masks to bounding boxes by thresholding at their extreme corners.

Baselines. Following prior work [53, 54], we compare Qwen-3D against state-of-the-art expert (non-LMM) and LLM-based approaches. Expert baselines include two-stage [55, 56] and single-stage [3, 23] methods. LLM-based baselines include: (i) two-stage models relying on detector proposals (LEO [20], Chat-Scene [19], Video-3D-LLM [53]); (ii) single-stage text-space decoders (3D-LLM [17]); and (iii) single-stage $< REF >$ token decoders (Grounded 3D-LLM [17], LLaVA-3D [54]). Unlike these, our single-stage method directly decodes segmentation masks by routing LMM backbone features to an object mask decoder. We also evaluate against proprietary 2D VLMs (GPT-4o [35], LLaMA [2], Qwen2-VL [32]). Tab. 1 presents quantitative results, and full results on Referit3D are available in the appendix.

Qwen-3D outperforms all prior 3D LMM-based models. It surpasses the text-decoding single-stage model of 3D-LLM [17] by over 30%, the recent single-stage state-of-the-art LLAVA-3D [54] by 12%, and the two-stage Video-3D-LLM [53] by 4%. This establishes Qwen-3D as the new state-of-the-art for 3D referential grounding among LMM-based models. Notably, our strongest baseline, Video-3D-LLM, uses a larger backbone (Qwen2-VL-7B), whereas Qwen-3D uses Qwen2.5-VL-3B (e.g., 70.1 vs. 65.3 on RealWorldQA [45]), indicating that the improvements of Qwen-3D are not solely explained by a stronger VLM backbone.

Qwen-3D closes the gap with expert 3D grounding models. On ScanRefer [10], Qwen-3D outperforms state-of-the-art expert model UniVLG and all other models and substantially narrows the gap between specialist 3D models

Table 2. **Evaluation on Locate3D ScanNet++.**

Model	Acc@25	Acc@50
UniVLG [23]	32.3	24.6
Video-3D LLM [53]	33.2	27.6
Qwen-3D (Ours)	**50.5**	**28.5**

Table 3. **Evaluation on ScanNet200 Instance Segmentation.**

	Model	mAP	mAP25
Closed Vocabulary	Mask3D [39]	27.4	42.3
	PQ3D [56]	27.0	46.3
	MAFT [27]	29.2	43.3
	ODIN [22]	31.5	53.1
Language-Prompted	PQ3D [56]	20.2	32.5
	UniVLG [23]	27.9	46.1
LLM-Based	Grounded-3D-LLM [12]	12.1	16.8
	Qwen-3D (Ours)	**27.7**	**46.2**

and LMM-based approaches on other 3D grounding benchmarks.

4.2. Out-of-Domain 3D Referential Grounding

While LMMs often trail specialists in-domain, they typically excel at out-of-domain (OOD) generalization. We evaluate Qwen-3D on Locate-3D [3], which provides human instructions for ScanNet++ [49] scenes. ScanNet++ introduces a distinct domain shift from our fine-tuning data as it is captured via iPhone LiDAR rather than ScanNet's iPad Structure sensor.

We compare against public checkpoints of UniVLG and Video-3D-LLM (supplied with state-of-the-art ODIN [22] box proposals). We omit LLaVA-3D [54] as its grounding model weights and code are not publicly released. As shown in Tab. 2, Qwen-3D significantly outperforms both baselines on these OOD tasks. We attribute our better generalization over UniVLG to Qwen-3D's stronger pre-training. Furthermore, while Video-3D-LLM uses a large VLM backbone, its reliance on off-the-shelf 3D detectors bottlenecks OOD robustness. Conversely, Qwen-3D directly decodes boxes from VLM features, allowing it to better exploit the underlying representation for superior generalization.

4.3. 3D Instance Segmentation

We evaluate Qwen-3D on the ScanNet200 [37] instance segmentation benchmark. While traditional methods [22, 39] assume a closed vocabulary setup, recent models [23, 56]—like ours—adopt a language-prompted paradigm (e.g., "find chairs. tables. sofa."). Furthermore, because most 3D grounding VLMs predict only a few bounding boxes, they fail as full scene detectors, with Grounded-3D-LLM [12] being the only exception to our knowledge. As shown in Tab. 3, Qwen-3D outperforms Grounded-3D-LLM by 15% AP and 30% AP25, effectively bridging the performance gap between generalist VLMs and matching the performance of language-prompted specialist models.

Table 4. **2D Ref. grounding datasets and RealWorldQA**

	RefCOCO	RefCOCO+	RefCOCOg	RealWorldQA
LAVT [48] (B)	72.7	62.4	61.2	-
ReSTR [26]	67.2	55.7	54.5	-
X-Decoder (L) [57]	-	-	64.6	-
UniVLG [23]	69.2	61.3	64.1	-
Qwen-2.5 3B [5]	89.1	82.4	85.2	**62.6**
Qwen-3D (Ours)	**90.7**	85.3	**86.0**	61.4

4.4. 3D Visual Question Answering

We evaluate Qwen-3D on two 3D question answering benchmarks: ScanQA [4] and SQA3D [34]. Both datasets use visual scenes from ScanNet [14], with ScanQA focusing on spatial-relation questions and SQA3D emphasizing situational reasoning.

Following prior work, we report Exact Match (EM@1), ROUGE (R), CIDEr (C), and METEOR (M). As shown in Tab. 1, Qwen-3D outperforms state-of-the-art expert model UniVLG [23], and achieves comparable performance to the single-stage LMM state-of-the-art LLaVA-3D-7B [54].

4.5. 2D Vision-Language tasks

To prevent degradation of the base model's strong 2D capabilities, we co-train Qwen-3D on 2D datasets, including the RefCOCO family [25] and LLaVA-Instruct-150k [31]. Tab. 4 evaluates Qwen-3D against its pre-trained base model (Qwen2.5-VL 3B [5]) on RefCOCO 2D grounding benchmarks and the held-out RealWorldQA [45] dataset. As in our 3D setup, Qwen-3D predicts segmentation masks which we convert to bounding boxes for evaluation. Results show that Qwen-3D performs comparably to the 2D-only base model across both tasks, successfully maintaining its 2D strengths while extending effectively to 3D.

For additional visualizations of Qwen-3D, ablations on pre-training retention, and results on text-only tasks, please refer to the appendix.

4.6. Additional Analysis and Ablations

We ablate several of our key design choices on ScanRefer (Top-1@0.5).

Full Text Tokens vs. Grounding Token We ablate the full token interface between the vision-language attention module and the object mask decoder against a variant that replaces the full text token injection with a special grounding token $<$ REF $>$ into the mask decoder, similar to LLAVA-3D [54] and Grounded-3D-LLM [12]. As shown in

CVPR #00027

Table 5. **Ablations of Qwen-3D**

(a) **Text vs.** $<$ REF $>$

Model	Top1 Acc.
Grounded 3D-LLM	44.0
LLAVA-3D - 7B	42.7
Qwen-3D $<$ REF $>$	43.0
Qwen-3D (ours)	**57.7**

(b) **Pos. Embed.**

Pos. Embed.	Top1 Acc
2D	53.2
Naive 3D mRoPE	53.5
3D RoPE (ours)	**57.7**

(c) **Attn. Mask**

Mask Type	Top1 Acc.
Causal	36.4
Full	**57.7**

(d) **VLM Tuning**

Finetune	Top1 Acc.
Frozen	37.0
Finetune	**57.7**

Tab. 5a, with the $<$ REF $>$ token, Qwen-3D achieves comparable performance to both LLAVA-3D and Grounded-3D-LLM, however, Qwen-3D with the full text and visual token interface greatly outperforms all $<$ REF $>$ token baselines.

Positional Embedding: We compare the 3D positional embeddings against the 2D MRoPE utilized in Qwen2.5-VL and a naive implementation of 3D RoPE in Qwen's MRoPE on 3D grounding. As shown in Tab. 5b, the 3D RoPE outperforms all other RoPE variants. We hypothesize this is because 3D RoPE improves the model's cross-view spatial reasoning over the 2D variant, while the distribution of the frequencies in Qwen's original MRoPE hinders the model's capability to discern spatial relationships along the point cloud's axes. We provide further details of the alternative RoPE designs in the appendix.

Causal Mask vs. Full Attention Mask: Qwen2.5-VL applies causal masking in its vision–language attention layers, even for non-autoregressive tasks such as grounding. We compare this original design—which stays closer to the model's pre-training distribution—against a full attention mask where all vision and language tokens attend to each other except during autoregressive text generation. As shown in Tab. 5c, the full-attention variant significantly outperforms the causal-mask variant.

Freezing vs. Fine-tuning the Base VLM: As shown in Tab. 5d, fine-tuning the underlying Qwen VLM is essential to achieve strong grounding performance. Keeping the VLM frozen leads to a substantial drop in accuracy.

5. Limitations and Future Directions

In this work, we focus on visual grounding from multi-view RGB videos that depict largely *static scenes*. Extending Qwen-3D to handle dynamic environments is an important direction for future work. This will likely require integrating temporal motion representations, dynamic scene reconstruction, or object-centric tracking into the model's geometric reasoning pipeline. Additionally, as with prior 3D vision-language models, the performance of Qwen-3D depends on the quality of the input 3D geometry. Although our experiments already operate on real-world noisy 3D reconstructions and demonstrate strong performance, jointly refining geometry during inference remains an interesting direction for future work.

Another promising direction is enabling *multi-step grounded reasoning*. In the current formulation, Qwen-3D primarily performs single-step grounding by directly localizing language-referred entities in the scene. However, many real-world tasks require sequential reasoning, such as composing multiple grounding operations, performing spatial comparisons, or executing multi-stage instructions. Incorporating multi-step reasoning mechanisms, for example through iterative grounding or reasoning frameworks such as [38], could enable richer interaction between language and spatial representations. We expect that combining geometry-aware perception with structured reasoning will further improve performance on complex visual–language tasks in 3D environments.

6. Conclusion

We introduced Qwen-3D, a geometry-aware 3D vision–language model that integrates explicit multi-view geometric reasoning into a strong multimodal backbone. By leveraging depth and camera pose to guide token compression and introducing scale-aligned 3D Rotary Positional Embeddings, Qwen-3D efficiently processes long multi-view video streams while maintaining consistent cross-view spatial understanding. We further proposed a tightly integrated grounding architecture that directly connects contextualized vision–language features from the backbone to a 3D query-based segmentation decoder, enabling direct language-to-3D alignment. Extensive experiments show that Qwen-3D significantly advances the state of the art among general-purpose 3D LMMs, achieving strong improvements in 3D grounding and instance segmentation while preserving competitive 2D multimodal performance. Our results substantially narrow the gap between generalist multimodal models and specialist 3D grounding systems, and demonstrate superior robustness in out-of-domain settings. Overall, our findings highlight the importance of tightly integrating geometric structure with pretrained vision–language representations, suggesting a promising path toward unified models capable of reasoning across both 2D and 3D visual environments.

References

[1] 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 *Proc. ECCV*, 2020. 5, 6, 12, 13, 14

[2] Meta AI. Llama 3: The llama-3 herd of models. https://ai.meta.com/llama/, 2024. Large language model. 6, 14

[3] Sergio Arnaud, Paul McVay, Ada Martin, Arjun Majumdar, Krishna Murthy Jatavallabhula, Phillip Thomas, Ruslan Partsey, Daniel Dugas, Abha Gejji, Alexander Sax, Vincent-Pierre Berges, Mikael Henaff, Ayush Jain, Ang Cao, Ishita Prasad, Mrinal Kalakrishnan, Michael Rabbat, Nicolas Ballas, Mido Assran, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Locate 3d: Real-world object localization via self-supervised learning in 3d, 2025. 3, 6, 7, 14

[4] 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*, pages 19129–19139, 2022. 5, 6, 7, 12

[5] Shuai Bai, Keqin Chen, Xuejing Liu, and Jialin Wang et. al. Qwen2.5-vl technical report, 2025. 1, 2, 4, 7, 12, 13

[6] 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. 13

[7] Ang Cao, Sergio Arnaud, Oleksandr Maksymets, Jianing Yang, Ayush Jain, Sriram Yenamandra, Ada Martin, Vincent-Pierre Berges, Paul McVay, Ruslan Partsey, Aravind Rajeswaran, Franziska Meier, Justin Johnson, Jeong Joon Park, and Alexander Sax. From thousands to billions: 3d visual language grounding via render-supervised distillation from 2d vlms, 2025. 6

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Proc. ECCV*, 2020. 4

[9] Angel 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. *arXiv preprint arXiv:1709.06158*, 2017. 5

[10] Dave Zhenyu Chen, Angel Chang, and Matthias Nießner. ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In *Proc. ECCV*, 2020. 5, 6, 12, 14

[11] 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*, pages 26428–26438, 2024. 1, 3, 6, 14

[12] 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*, 2024. 1, 3, 4, 6, 7, 14

[13] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 4, 13

[14] 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*, pages 5828–5839, 2017. 1, 6, 7

[15] Angela Dai, Matthias Nießner, Michael Zollöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface re-integration. *ACM Transactions on Graphics 2017 (TOG)*, 2017. 12

[16] 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. 3, 6, 14

[17] 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. 3, 6, 14

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 5

[19] 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. 1, 3, 6, 14

[20] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 1, 3, 6, 14

[21] 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*, pages 417–433. Springer, 2022. 3, 4, 5, 6, 14

[22] Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W Harley, Gabriel Sarch, Kriti Aggarwal, Vishrav Chaudhary, and Katerina Fragkiadaki. Odin: A single model for 2d and 3d segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3564–3574, 2024. 1, 3, 4, 5, 6, 7, 14

[23] Ayush Jain, Alexander Swerdlow, Yuzhou Wang, Sergio Arnaud, Ada Martin, Alexander Sax, Franziska Meier, and Katerina Fragkiadaki. Unifying 2d and 3d vision-language understanding, 2025. 1, 3, 4, 5, 6, 7, 12, 13, 14, 15

[24] Aishwarya Kamath, Mannat Singh, Yann André LeCun, Ishan Misra, Gabriel Synnaeve, and Nicolas Carion. MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In *Proc. ICCV*, 2021. 4

[25] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in pho-

tographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics. 5, 7, 12

[26] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers, 2022. 7

[27] Xin Lai, Yuhui Yuan, Ruihang Chu, Yukang Chen, Han Hu, and Jiaya Jia. Mask-attention-free transformer for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3693–3703, 2023. 7

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[29] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 4

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5, 12

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5, 7, 12

[32] Yuanyuan Liu, Haiyang Mei, Li Zhang, et al. View-on-graph: Zero-shot 3d visual grounding via vision-language reasoning on scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026. 6

[33] 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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 3

[34] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 5, 7, 12

[35] OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and Aidan Clark et al. Gpt-4o system card, 2024. 2, 6, 14

[36] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 14

[37] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141. Springer, 2022. 1, 5, 7, 12

[38] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J. Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning, 2025. 8

[39] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 7, 15

[40] Simranjit Singh, Georgios Pavlakos, and Dimitrios Stamoulis. Evaluating zero-shot gpt-4v performance on 3d visual question answering benchmarks, 2024. 6

[41] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://crfm.stanford.edu/2023/03/13/alpaca.html`, 2023. Dataset and model release. 5, 12

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, and Radu Soricut et al. Gemini: A family of highly capable multimodal models, 2025. 2

[43] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024. 5

[44] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. 12, 14

[45] X.AI. Realworldqa. Blog post, 2024. Accessed: 2025-05-12. 6, 7

[46] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, and Bo Zheng et. al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 14

[47] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866, 2021. 14

[48] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Lavt: Language-aware vision transformer for referring image segmentation, 2022. 7

[49] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 7

[50] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021. 14

[51] 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*, pages 15225–15236, 2023. 13

[52] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13624–13634, 2024. 6, 14

[53] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8995–9006, 2025. 1, 3, 5, 6, 7, 12, 13, 14

[54] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness, 2024. 1, 3, 4, 5, 6, 7, 12, 14

[55] 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*, pages 2911–2921, 2023. 1, 3, 6, 14

[56] 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. *arXiv preprint arXiv:2405.11442*, 2024. 3, 6, 7, 14

[57] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 7

CVPR #00027

7. Appendix

7.1. Sensitivity to Pose and Depth Noise

While Qwen-3D already operates on real-world sensor noise in all experiments reported in the paper, we further stress-test the model under controlled settings with substantial depth and camera pose noise.

Following UniVLG [23], we simulate depth noise by injecting Gaussian noise with increasing variance into the raw depth maps before unprojection. To simulate additional pose noise, we add Gaussian noise with increasing variance to both the translation and rotation components of the provided camera poses.

We evaluate on ScanRefer (Top-1@0.25) against UniVLG, the strongest reported baseline. We choose UniVLG as the primary baseline because many strong 3D LMMs, such as Video-3D-LLM [53], adopt two-stage pipelines that rely on external object detection models and operate on preprocessed detections at inference time. Fairly evaluating such methods under additional noise would require these external detectors to be re-run on the perturbed point clouds, making comparisons difficult to standardize. In contrast, existing single-stage 3D LMMs either substantially underperform our model or do not release the necessary visual grounding code (e.g., LLaVA-3D [54]). Nevertheless, UniVLG remains the strongest reported baseline on ScanRefer, and we therefore report comparisons against it.

As shown in Fig. 4, Qwen-3D avoids catastrophic failure even under significant noise. Specifically, Qwen-3D demonstrates high robustness to depth noise, showing no significant performance drop even under high variance noise, matching UniVLG's resilience to spurious points. Qwen-3D is similarly robust to pose noise, degrading gracefully and even outperforming the state-of-the-art model under extreme misalignment. We attribute this robustness to the strong integration of 2D pretrained vision-language priors from the Qwen backbone into explicit 3D structures, allowing the model to project reasoning over 2D inputs into accurate 3D masks even when the underlying 3D structures are corrupted.

7.2. Additional 3D Grounding Results

We report detailed 3D grounding evaluations on the SR3D, NR3D [1], and ScanRefer [10] benchmarks in Tab. 6. To our knowledge, no prior 3D LLM-based models report on the SR3D and NR3D benchmarks. Qwen-3D outperforms all prior LLM methods on ScanRefer using only a 3 billion parameter backbone, despite many of the strongest prior methods utilizing more powerful backbones of 7 billion parameters or more.

7.3. Text-only Results

We jointly train our model on visual reasoning tasks alongside the pure text instruction fine-tuning dataset, Alpaca [41], and evaluate Qwen-3D's instruction-following capabilities on the MMLU-Pro [44] benchmark. As shown in Tab. 7, our model performs competitively with Qwen2.5-VL-3B [5] on general instruction-following tasks. This demonstrates no significant degradation of the original backbone's text generation capabilities, despite our adaptations from causal attention masking to full attention over the inputs.

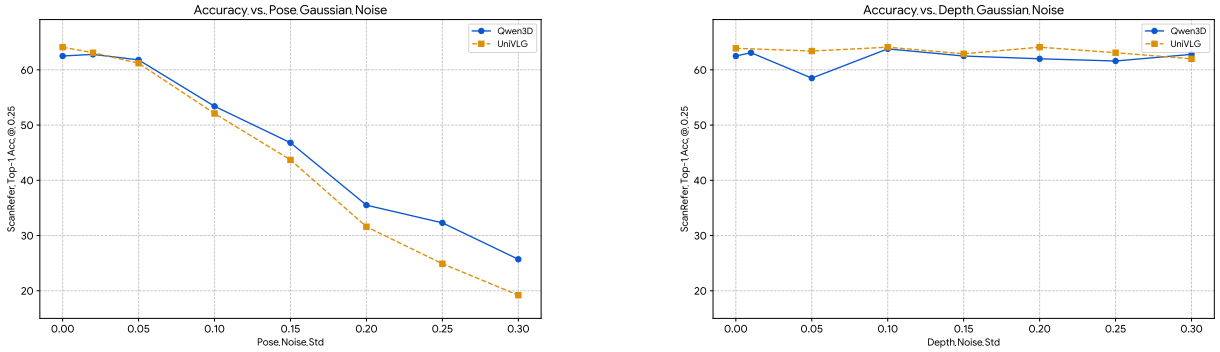
7.4. End-to-end Latency

We estimate the end-to-end latency of Qwen-3D on 3D tasks, tracking from raw scene inputs to the final 3D grounding outputs. As noted in Section 3 of the main text, doing an inference on the full set of multi-view RGB-D frames through Qwen-3D takes approximately 2 seconds. This includes the time required to unproject RGB-D frames (20 ms) and voxelize the 3D point cloud (1 ms). Standard RGB-D SLAM reconstruction methods [15] used for ScanNet scenes add about 1 to 5 seconds, yielding a total end-to-end latency of roughly 3 to 7 seconds. Notably, SLAM reconstruction costs can be amortized per scene as the reconstruction only needs to be run once; subsequent queries to the same environment only require the 2-second model forwarding time. Additionally, we note that all 3D LMMs need to run SLAM to obtain their point clouds; Qwen-3D does not introduce any new dependencies over prior methods.

7.5. Detailed Comparison of Training Datasets

Qwen-3D is trained on a mix of 2D and 3D datasets. The 3D datasets comprises approximately 255k samples spanning 3D question answering [4, 34], referential grounding [1, 10], and instance segmentation [37]. The 2D datastes include referential grounding (RefCOCO, RefCOCO+, RefCOCOg [25]), instance segmentation (COCO [30]), captioning and question answering (LLaVA-Instruct-150k [31]), and instruction fine-tuning (Alpaca [41]). Notably, all of these 2D datasets are included in the pre-training dataset of many modern 2D VLMs, including Qwen2.5VL, the backbone we use. We include them to avoid catastrophic forgetting of 2D capabilities as our model's VLM backbone as it trains on new 3D data.

Our most competent expert model baseline, UniVLG [23], is trained on a nearly identical dataset mixture, excluding only LLaVA-Instruct-150k and the text-only Alpaca dataset. In contrast, the state-of-the-art single-stage 3D LMM, LLaVA-3D [54] utilizes a much larger 3D fine-tuning dataset, compiling over 860K 3D visual-reasoning samples together from various benchmarks alongside the 2D LLaVA-Instruct-150k dataset. The leading two-stage



(a) Pose noise sensitivity.

(b) Depth noise sensitivity.

Figure 4. Sensitivity analysis to camera pose noise (left) and depth noise (right) on ScanRefer (Top-1@0.25). We compare Qwen-3D against state-of-the-art expert model UniVLG [23]

3D LMM, Video-3D-LLM [53], uses a 3D training mixture similar to ours - substituting Referit3D [1] with Multi3DRefer [51]. However, it also relies heavily on external object detectors, which are typically trained on ScanNet200 instance segmentation datasets.

7.6. Rotary Position Embedding Implementation

To process interleaved visual and textual data in the vision-attention layers, Qwen2.5-VL [5] utilizes a Multimodal Rotary Positional Embedding (mRoPE). This method decomposes positional encodings into temporal, height, and width components. Each token i in a sequence of length K is assigned a position ID in each dimension, and the final positional embedding is the concatenation of the sinusoidal encodings of these components:

$$\mathrm{PE}(\mathbf{p}) = \mathrm{PE}(t_i; h_i; w_i)_{i=1}^{K},$$

For text tokens, the temporal ID t increments linearly, while h and w are assigned the same values as t, effectively reducing mRoPE to 1D-RoPE. For image tokens, t is held constant while (h, w) are assigned to corresponding pixel coordinates. To maintain temporal consistency across all modalities and dimensions, the position IDs of each modality are incremented by the maximum position ID of the preceding modality, ensuring visual features and text instructions occupy distinct, non-overlapping positions in the positional embeddings.

2D RoPE Projection: This extension aims to stay as close as possible to the Qwen2.5-VL pre-training distribution. Temporal IDs follow the original scheme, while the height and width IDs for voxelized point-cloud tokens are obtained by sampling corresponding 2D pixel coordinates from one of the views where the voxel is visible.

Naive 3D MRoPE: This extension aims to inject absolute 3D spatial information by replacing the height and width components with embeddings of absolute XYZ coordinates:

$$\mathrm{PE}(\mathbf{p}) = \mathrm{PE}(t_i; x_i; y_i, z_i)_{i=1}^{K},$$

This method is similar to our final 3D RoPE, except it utilizes Qwen2.5-VL's partitioning of the embedding channels among the four dimensions, resulting in an imbalanced frequency spectrum across the four dimensions, with high-frequency channels being assigned to the temporal axis and low-frequency channels being assigned to the z-axis. We find that this degrades the model's 3D grounding capabilities in our ablations.

Our 3D RoPE instead initializes the same range of frequencies for each dimension, resulting in a uniform distribution of high and low frequency channels across all components. We find that this method significantly outperforms all other variants on 3D grounding tasks.

7.7. Additional Hyperparameters

Hungarian Matching cost weights during training have a sizable impact on our model's performance on 3D grounding tasks. Our final configuration utilizes the standard cost coefficients from Mask2Former [13], as we found that alternative weighting schemes proposed in other 3D grounding works [23] can hinder performance.

7.8. Comparison with Token Merging

While Qwen-3D utilizes voxelization pooling to reduce multi-view redundancy, a popular alternative for reducing redundancy in video model architectures is token merging [6], which iteratively fuses visually similar tokens within ViT attention layers. We attempted to compare these methods to our voxelization pooling by replacing the voxelization with a plug-and-play token merging strategy within our vision-language attention layers [6]. This implementation proved infeasible for dense 3D scenes, as

Table 6. **Results on 3D language grounding for both experts and LMMs.** We evaluate top-1 accuracy on the official validation set. * UniVLG numbers reproduced using the official code with a single 40GB 8 GPU node instead of their 32 GPU setup, and verified with the authors.

	Methods	SR3D				NR3D				ScanRefer		
		Acc @25 (Det)	Acc @50 (Det)	Acc @75 (Det)	Acc (GT)	Acc @25 (Det)	Acc @50 (Det)	Acc @75 (Det)	Acc (GT)	Acc @25 (Det)	Acc @50 (Det)	Acc @75 (Det)
Experts	ReferIt3DNet [1]	27.7	-	-	39.8	24.0	-	-	-	26.4	16.9	-
	ScanRefer [10]	-	-	-	-	-	-	-	-	35.5	22.4	-
	InstanceRefer [50]	31.5	-	-	48.0	29.9	-	-	-	40.2	32.9	-
	LanguageRefer [36]	39.5	-	-	56.0	28.6	-	-	-	-	-	-
	SAT-2D [47]	35.4	-	-	57.9	31.7	-	-	-	44.5	30.1	-
	BUTD-DETR [21]	52.1	-	-	67.0	43.3	-	-	54.6	52.2	39.8	-
	PQ3D [56]	62.0	55.9	46.2	79.7	52.2	45.0	37.6	66.7	56.7	51.8	43.3
	3D-VisTA [55]	56.5	51.5	42.8	76.4	47.7	42.2	35.5	65.1	51.0	46.2	36.7
	ODIN [22]	38.1	29.3	23.1	-	31.6	20.8	15.8	-	43.1	33.4	26.2
	Locate-3D [3]	65.8	52.9	-	-	53.7	40.5	-	-	59.9	49.6	-
	Locate-3D+ [3]	68.2	54.8	-	-	56.1	43.2	-	-	61.1	50.9	-
	UniVLG (1 GPU node) [23]*	71.7	63.6	51.2	-	52.8	42.0	33.0	-	64.1	52.7	43.5
	UniVLG [23]	**73.0**	**64.8**	**51.8**	-	**58.3**	**49.8**	**39.1**	-	**63.5**	**56.4**	**46.0**
LMMs	**2D VLMs***											
	LLaMA [2]	21.3	13.9	-	-	28.0	16.9	-	-	37.3	24.2	-
	Qwen2-VL-2B [46]	-	-	-	-	31.1	-	-	-	35.9	31.9	-
	Qwen2-VL-72B [46]	-	-	-	-	47.6	-	-	-	44.8	40.3	-
	GPT-4o [35]	29.2	18.9	-	-	38.2	25.1	-	-	48.2	32.5	-
	Two-Stage											
	ChatScene [19]	-	-	-	-	-	-	-	-	55.5	50.2	-
	LEO [20]	-	-	-	-	-	-	-	-	-	-	-
	Video-3D LLM - 7B [53]	-	-	-	-	-	-	-	-	58.1	51.7	-
	Single-Stage											
	NaviLLM [52]	-	-	-	-	-	-	-	-	-	-	-
	LL3DA [11]	-	-	-	-	-	-	-	-	-	-	-
	SceneLLM [16]	-	-	-	-	-	-	-	-	-	-	-
	3D-LLM [17]	-	-	-	-	-	-	-	-	30.3	-	-
	Grounded 3D-LLM [12]	-	-	-	-	-	-	-	-	47.9	44.1	-
	LLaVA-3D - 7B [54]	-	-	-	-	-	-	-	-	50.1	42.7	-
	Qwen-3D(Ours)	**61.8**	**55.8**	**45.6**	-	**55.1**	**48.7**	**39.5**	-	**65.6**	**57.7**	**45.0**

Table 7. **Qwen-3D on instruction-following benchmark MMLU-Pro [44]**

Model	EM@1
Qwen2.5-VL-72B	71.2
Qwen2.5-VL-3B	57.1
Qwen-3D (Ours)	55.6

with token merging, the first vision-language attention layer must still process the point cloud at full resolution (averaging around 33K points for ScanNet scenes). This consistently resulted in GPU out-of-memory errors during inference, regardless of how aggressively the token merging would downsample in subsequent layers. Conversely, our voxelization pooling preemptively downsamples the point cloud, reducing the number of points to an average of about 14K points prior to the attention layers. This acts as a natural, 3D-aware compression mechanism that bypasses the memory overhead of full-resolution multi-view attention.

Table 8. **Mask Decoder Text-Query Cross Attention Ablations**

Mask Decoder	Top1 Acc
Q<->V cross-attn	**52.5**
Q->V cross-attn	44.3

7.9. Additional Implementation Details

For our model, we use Qwen2.5-VL as the backbone due to its open-source implementation and its tight integration

of visual and textual features in its visual-language attention layers. Within our codebase, we also introduce several engineering modifications to the original Qwen2.5-VL architecture.

The Qwen2.5-VL vision encoder handles the batching of multiple images by sequentially concatenating images along the token dimension and using attention masks to prevent cross-image interaction, resulting in $O(N^2)$ complexity in the vision encoder with respect to the number of images N. We optimize this by implementing batched forwarding in the attention mechanism to perform per-image batched attention. This modification reduces the complexity from $O(N^2)$ to $O(N)$ while preserving the original behavior of the vision encoder. This also substantially reduces GPU memory usage, which is essential for processing scene videos with large numbers of input RGB images.

Additionally, Qwen2.5-VL utilizes causal masking in its visual-language attention layers, allowing each vision and text token to attend only to itself and to earlier tokens in the sequence. This is problematic for our point cloud inputs because this operation is not permutation-invariant and would cause the model to learn artifacts of the arbitrary ordering of the point cloud sequences rather than the underlying geometry. We rectify this by replacing this causal-masked attention with all-to-all attention for all text and visual inputs and finetuning our backbone with causal masking only over the answer tokens. We ablate these changes in the main text.

7.10. Visualizations of Failure cases

In Fig. 5, we outline three common failure modes for Qwen-3D on 3D referential grounding tasks:

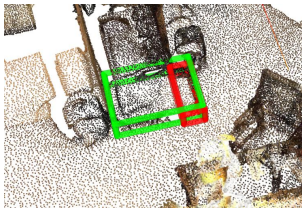
Incomplete object selection (Left): The model occasionally masks only a partial section of the target. This typically occurs when the underlying point cloud contains significant holes or artifacts, skewing the geometric understanding of the complete object shape.

Confusion between instances (Middle): The model correctly identifies the object class but grounds the wrong instance. Confusion between multiple instances is a common failure mode seen in mask-decoding architectures, and similar issues have been noted in previous works, such as Mask3D [39] and UniVLG [23].

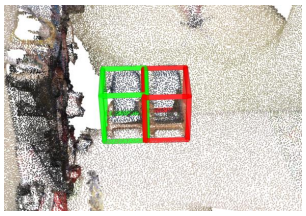
Language ambiguity (Right): The model occasionally confuses the target object with other reference objects mentioned in more ambiguous querying expressions.

7.11. Visualizations of Qwen-3D Results

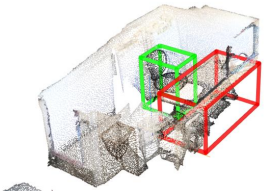
We include visualizations of Qwen-3D predictions in 3D referential grounding tasks in Fig. 6, 2D referential grounding in Fig. 7, instance segmentation in Fig. 8, and visual question-answering tasks in Fig. 9.



select the suitcase that is in the center of the bathtub and the desk



chair with back against the wall directly next to book shelf



there is a black desk chair in the center of the room. it is in front of the desk with a phone sitting on it

Figure 5. Failure cases of Qwen-3D on 3D grounding tasks. The red segmentation masks and boxes refer to Qwen-3D's prediction and the green masks and boxes indicate the ground truth.

CVPR
#00027

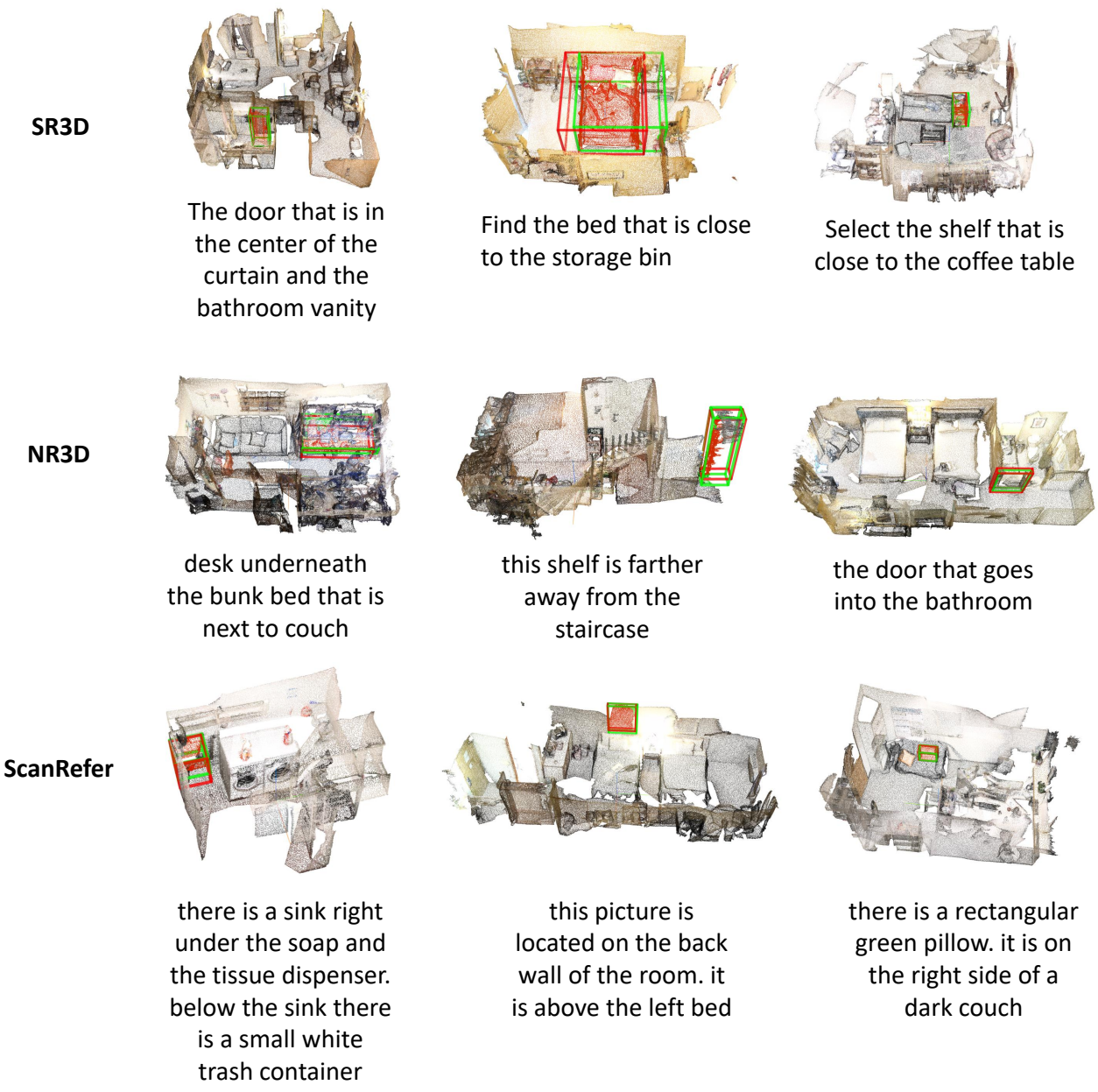


Figure 6. Visualizations of Qwen-3D's predictions on **3D Referential Grounding Datasets SR3D, NR3D, and Scanrefer.** The red segmentation masks and boxes refer to Qwen-3D's prediction and the green masks and boxes indicate the ground truth.

Figure 7. Visualizations of Qwen-3D's predictions on **2D Referential Grounding Datasets RefCOCOg, RefCOCO+, and RefCOCO.** The model's predictions are indicated by the green mask and bounding box.

CVPR
#00027

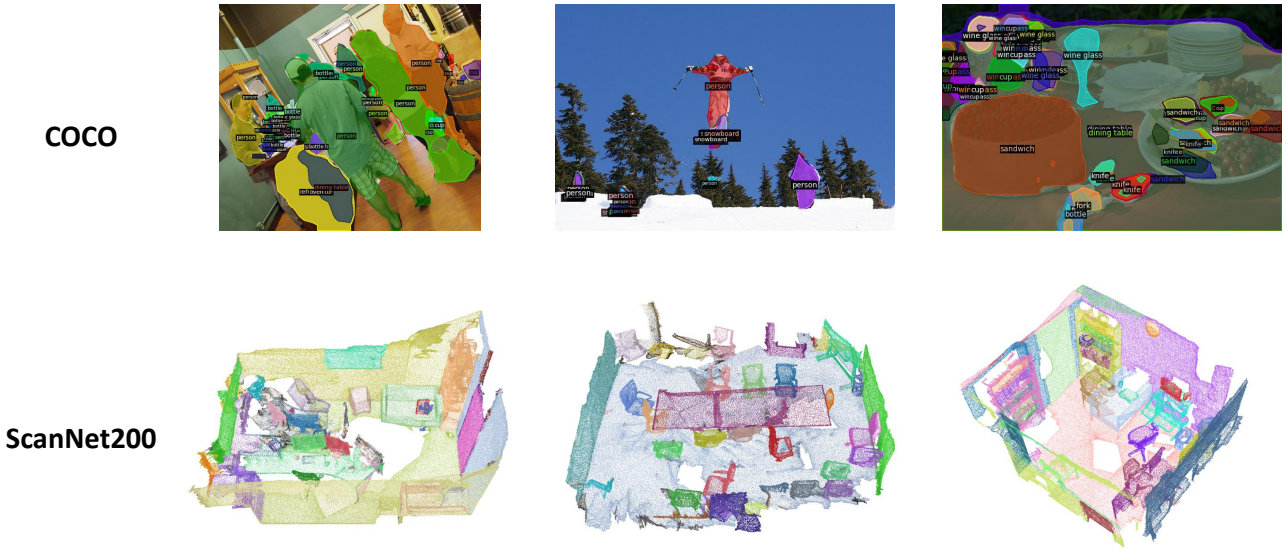


Figure 8. Visualizations of Qwen-3D's instance segmentation predictions on **COCO and ScanNet200**.

CVPR
#00027

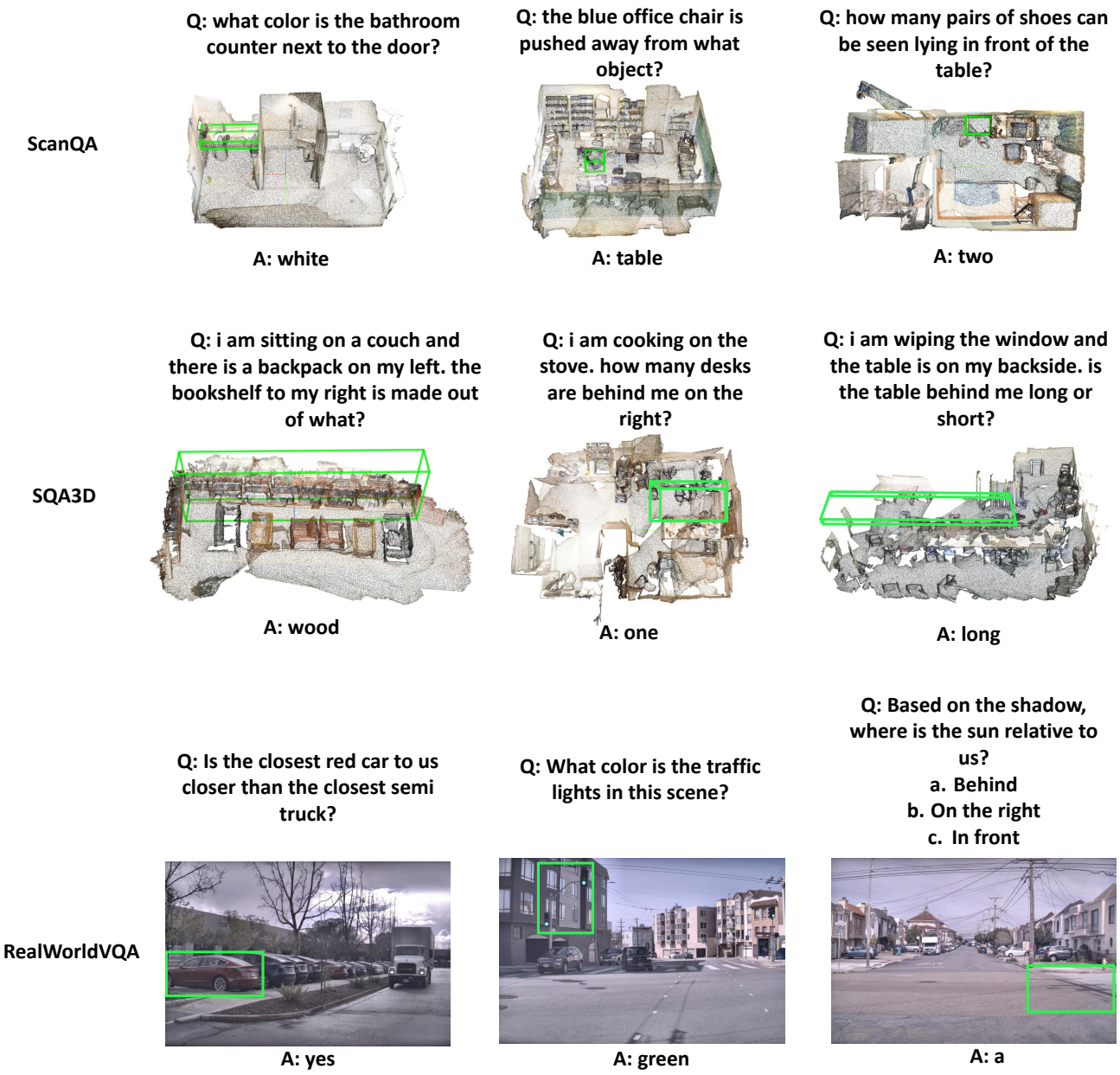


Figure 9. Visualizations of Qwen-3D's responses to visual question-answering tasks on **3D benchmarks SQA3D and ScanRefer and the 2D benchmark RealWorldVQA**. Green boxes denote objects relevant to the posed question.