

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

Lucy Lin[†], Ayush Jain[†], Yifan Liu, and Katerina Fragkiadaki

[†]Equal contribution
Carnegie Mellon University

Abstract. Large Multimodal Models (LMMs) have achieved remarkable success on images and short videos, yet scaling them to long videos remains challenging due to frame-centric tokenization and limited context windows. 3D geometry provides a natural compression mechanism for visual streams: depth and camera pose enable observations from multiple views and time steps to be fused into a persistent, world-aligned representation. While recent 3D LMMs leverage geometry-aware representations to improve spatial reasoning, they continue to lag behind specialist 3D perception systems on grounding and segmentation tasks. We argue that a key limitation is geometry-aware decoding: existing methods communicate 3D predictions through language tokens, proposal selection, or lightweight grounding queries, creating a bottleneck between language reasoning and dense geometric prediction. Building on these insights, we introduce Qwen-3D, a geometry-aware LMM that compresses visual information within the Qwen backbone using multi-view geometric cues, enabling efficient long-horizon visual reasoning over static scenes. Qwen-3D augments visual tokens with 3D Rotary Positional Embeddings, allowing attention to operate directly in 3D scene space rather than across independent image frames and thereby facilitating scalable cross-view and temporal reasoning. To bridge language and geometry, Qwen-3D incorporates a query-based segmentation decoder that grounds language directly in the underlying 3D scene representation, unifying referential grounding, instance segmentation, and visual question answering across both images and videos. Across a diverse set of benchmarks, Qwen-3D surpasses existing 3D LMMs and outperforms several large proprietary 2D models. Notably, Qwen-3D achieves these improvements while maintaining strong performance on standard 2D vision-language benchmarks by jointly training on 2D and 3D data. Our code and checkpoints can be found at the project website <https://qwen-3d.github.io/>.

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 long-range spatio-temporal reasoning across frames. Multi-view 3D geometry, in the form of depth and camera poses, offers

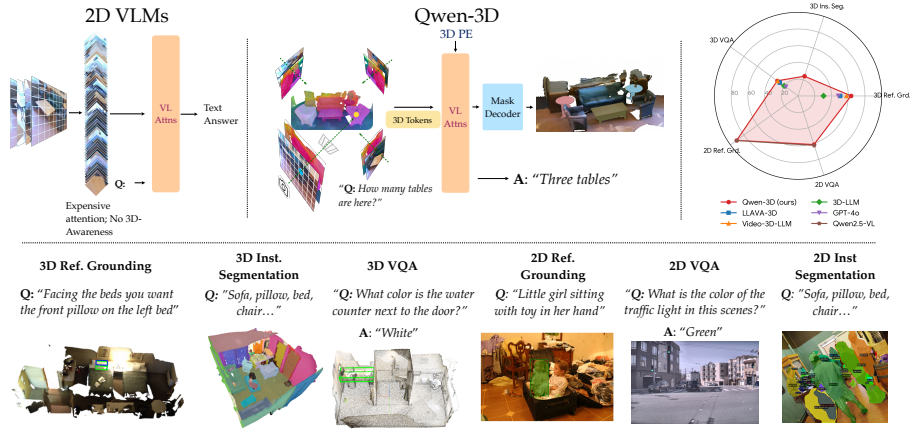


Fig. 1: Qwen-3D performs attention directly in 3D world space rather than over independent image frames. Given multi-view RGB observations, depth, and camera poses, Qwen-3D maps visual tokens into a shared 3D coordinate system and applies geometry-aware attention through 3D Rotary Positional Embeddings. The model jointly supports language reasoning, 2D grounding, and 3D grounding within a unified architecture, achieving state-of-the-art performance across a broad range of vision-language and 3D understanding benchmarks.

a principled alternative, allowing video frames to be mapped into a shared 3D coordinate system. This enables compression of long multi-view streams into compressed persistent scene representations where temporally distant frames may correspond to nearby 3D locations.

Existing approaches to integrating such 3D information compression into VLMs in order to improve their long range reasoning abilities generally follow one of two distinct paradigms. One line of work introduces 3D point clouds as auxiliary inputs to the model [10, 17, 18], 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 [50] and Video-3D-LLM [48] 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 substantially behind specialist 3D perception systems. Dedicated models trained for detection, segmentation, and grounding continue to outperform general-purpose 3D LMMs by a large margin [21, 22, 51]. Moreover, most current 3D LMMs do not even attempt standard 3D perception tasks such as object detec-

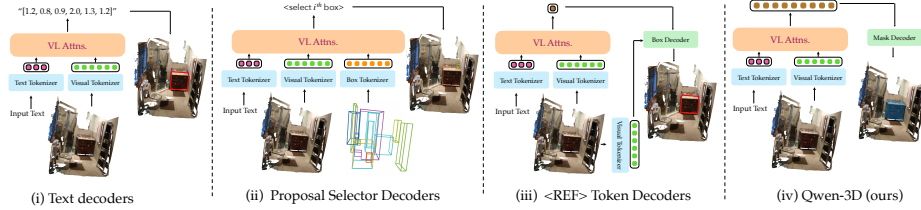


Fig. 2: Decoding object information in 3D LLMs. From left to right: (i) *Text-space decoding*, as in 3D-LLM [15], where object coordinates or bounding boxes are generated autoregressively as language tokens. Unlike images, 3D scenes do not admit a canonical world coordinate frame, making dense spatial prediction through text generation inherently ambiguous. (ii) *Proposal-selection decoding*, as in Video-3D-LLM [17, 18, 48], where the model selects from a predefined set of object proposals. This approach is fundamentally limited by proposal quality and cannot localize objects absent from the proposal set. (iii) *Special-token decoding*, as in LLaVA-3D [11, 50], where grounding information is communicated through a single dedicated token. While effective for coarse language-to-vision communication, this creates a severe information bottleneck between language reasoning and dense geometric prediction. (iv) **Qwen-3D (ours)**, which directly connects contextualized vision-language features from the backbone to a mask-based segmentation decoder. By sharing dense visual representations between language reasoning and geometric prediction, Qwen-3D grounds multiple referential expressions directly in the underlying scene representation, enabling high-quality 2D and 3D grounding and segmentation.

tion on ScanNet [13, 36]. The only exception, Grounded-3D-LLM [11], achieves less than half the performance of state-of-the-art 3D detectors.

We argue that this gap stems from a fundamental challenge in adapting language-centric architectures to 3D perception. While large language models excel at reasoning over discrete tokens, dense 3D grounding requires predicting spatially precise outputs in a continuous world coordinate system. Unlike images, which provide a canonical pixel coordinate frame, 3D scenes admit no universal reference frame: the same object may appear at entirely different coordinates across scans and environments. As a result, autoregressively decoding 3D boxes, coordinates, or masks as language tokens is an unnatural interface for 3D perception.

Existing 3D LLMs typically address this problem either by representing grounding outputs through text generation or by attaching lightweight grounding modules that communicate with the backbone through a small set of query vectors (Figure 2). While these approaches preserve the reasoning capabilities of the underlying language model, they create a severe information bottleneck between high-capacity visual representations and the dense geometric predictions required for grounding. Consequently, current 3D LLMs improve high-level spatial reasoning but remain significantly weaker than specialist systems on core 3D perception tasks. This observation raises an important open question: how should a large multimodal model interface with a 3D grounding system?

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 (Figure 1). Rather than treating grounding as language generation or restricting communication through a small set of latent queries, we propose a unified architecture in which the language model and grounding decoder share dense visual representations. This design enables rich bidirectional interactions between language reasoning and geometric perception, substantially improving grounding accuracy while preserving the general-purpose capabilities of the multimodal backbone. Built on the strong 2D foundation of Qwen2.5-VL-3B, Qwen-3D integrates 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.

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% Acc@25, surpasses the best single-stage 3D LMMs by 12% Acc@25, and increases 3D instance segmentation accuracy by 13% mAP. The model also achieves competitive 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. Following prior 3D LMMs such as LLaVA-3D [50] and Grounded-3D-LLM [11], we characterize Qwen-3D as a vision–language generalist based on its broad multi-modal 2D and 3D capabilities, while further broadening the evaluation scope beyond these baselines to include 3D object detection, text-only baselines, and generalization across unseen distributions.

Contributions. Our contributions are as follows:

- A geometry-aware VLM backbone that integrates multi-view structure with 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% and instance segmentation by 13%, while maintaining strong 2D multimodal performance.

We make our code publicly available at <https://qwen-3d.github.io/>.

2 Related Work

Building upon the rapid progress of 2D Large Multimodal Models [5, 35, 40], recent work has focused on extending these models to understand 3D scenes. Existing approaches can be grouped into four main categories:

(a) 3D point cloud encoders trained from scratch. Methods such as LL3DA [10], Scene-LLM [14], and Grounded-3D-LLM [11] 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 pre-trained features and augments them with 3D information via positional encodings.

(b) Learnable 3D feature compression. 3D-LLM [15] 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 [18] and ChatScene [17] 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 [21, 22, 48, 50]. 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, 20, 33] achieved strong performance by designing specialized architectures tailored for 3D grounding. Subsequent methods [3, 22, 51, 52] 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 [15] 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 [48], ChatScene [17], and LEO [18] 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 bottlenecked by the quality of the proposals.

(c) Special token decoding. (Fig. 2c) Following the mask-as-embedding paradigm introduced by LISA [26], several methods decode a special grounding token (e.g. $\langle \text{REF} \rangle$) and localize it using an explicit decoder head, as in LLaVA-3D [50], Grounded-3D-LLM [11], and Reason3D [19]. Essentially, the VLM heads and the decoder heads are only connected via the generated $\langle \text{REF} \rangle$ tokens. Although these methods can, in principle, ground multiple instances from a category, only Grounded-3D-LLM has been applied to 3D object detection.

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 LLMs.

3 Method

Qwen-3D extends Qwen2.5-VL with three key components. First, we construct a geometry-aware scene representation by projecting multi-view visual features into a shared 3D coordinate system and compressing redundant observations through voxel-based token merging. Second, we enable geometry-aware reasoning by replacing image-plane positional encodings with 3D Rotary Positional Embeddings, allowing attention to operate directly in world space. Third, we introduce a geometry-aware decoding mechanism that directly couples contextualized vision-language features with a mask-based grounding decoder, enabling dense language-guided prediction in both 2D and 3D. Figure 3 provides an overview.

3.1 Geometry-Aware Scene Representation

Given a set of posed RGB-D observations, we first lift visual features into a shared world coordinate system. Specifically, we use the Qwen2.5-VL vision encoder to extract image features and unproject their corresponding depth values using camera intrinsics and poses, producing a set of aligned feature-coordinate pairs

$$\mathcal{FC} = \{(f_i, p_i)\}_{i=1}^M,$$

where $f_i \in \mathbb{R}^D$ denotes a visual feature and $p_i \in \mathbb{R}^3$ its corresponding world-space coordinate.

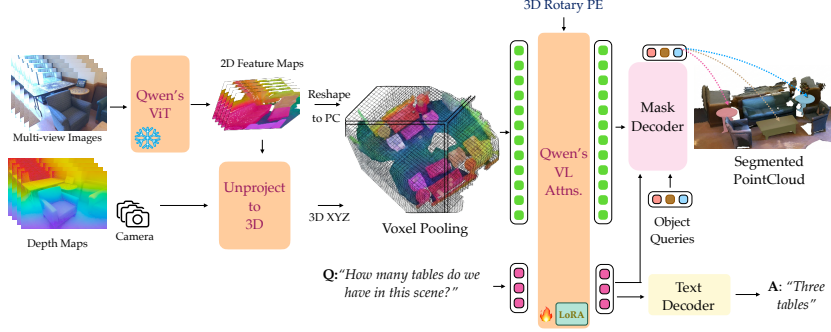


Fig. 3: Qwen-3D architecture. Given a natural language query and multi-view RGB-D inputs, the Qwen2.5-VL vision encoder extracts multi-view 2D features, unprojects them into world-space XYZ coordinates, and voxel-pools them to reduce spatial redundancy. The resulting tokens are fused with text via Qwen 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.

Multi-view observations produce substantial redundancy because pixels from different views often correspond to the same physical location. Rather than reasoning over frame-level tokens, we aggregate observations directly in world space. Specifically, we apply voxel-based token merging (with a voxel size of 5cm) to the feature-coordinate pairs, following [21, 22, 50]. 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 $M' \leq M$.

3.2 Geometry-Aware Vision–Language Attention

Unlike conventional VLMs that perform attention over image-plane coordinates, Qwen-3D performs attention directly in a shared world coordinate system. 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

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

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.

3.3 Geometry-Aware Decoding

Most existing 3D LMMs communicate grounding information through language tokens, proposal indices, or a small number of dedicated grounding tokens [11, 15, 48, 50]. While these interfaces preserve the reasoning capabilities of the language model, they create a bottleneck between the rich visual representations learned by the backbone and the dense geometric predictions required for grounding. Instead, we directly expose the full contextualized vision–language representation produced by the backbone to a mask-based segmentation decoder.

Let

$$\mathcal{V} = \{v_i, i = 1..M'\}$$

denote the contextualized visual tokens produced by the geometry-aware attention layers and

$$\mathcal{T} = \{t_j, j = 1..L\}$$

the corresponding language tokens. We instantiate (N) learnable object queries

$$\mathcal{Q} = \{q_n, n = 1..N\}$$

and process them through a Mask2Former-style decoder [12]. Each decoder layer alternates between cross-attention to visual and language tokens, and self-attention among object queries. This allows each query to jointly reason about scene geometry and linguistic context while maintaining access to the full visual representation rather than a compressed grounding bottleneck.

For all attention involving visual tokens, we use positional embeddings corresponding to the underlying modality (2D image coordinates or 3D world coordinates). After the final decoder layer, each object query predicts (i) a segmentation mask through a dot product with the updated visual tokens and (ii) a language grounding score through a dot product with the language tokens. This unified decoder naturally supports both 2D and 3D grounding. For referential grounding, object queries learn to associate text spans with the corresponding visual regions. For question answering and captioning, we use the original Qwen2.5-VL language head to perform next-token prediction over the contextualized scene–language representation. This design enables dense geometric prediction while preserving the reasoning capabilities of the underlying language model.

Training Objectives We supervise Qwen-3D on three losses:

(a) **Mask loss:** We assign predictions to ground-truth instances via Hungarian matching [7] and supervise matched masks with Binary Cross-Entropy (BCE) and Dice loss, following Mask2Former [12].

(b) **Text-span grounding loss:** As in [20, 23, 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:

$$\mathcal{L} = \alpha_{\text{mask}} \mathcal{L}_{\text{mask}} + \alpha_{\text{textground}} \mathcal{L}_{\text{textground}} + \alpha_{\text{gen}} \mathcal{L}_{\text{gen}} \quad (1)$$

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

3.4 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 [42] following UniVLG [22]. This enables a single set of parameters to operate seamlessly across images, videos, and reconstructed 3D scenes.

Implementation Details Qwen-3D introduces only $\sim 50\text{M}$ trainable parameters. We freeze the Qwen2.5-VL ViT backbone, fine-tune the vision-language attention layers via LoRA [16], and train the mask decoder from scratch. We train jointly on 2D and 3D datasets 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 [20] by concatenating object class names (e.g., “find chair. table. sofa.”).

Following [21, 22], 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 [48, 50]. We include additional hyperparameters and an ablation of the test-time frame subsampling in the appendix.

Table 1: Results on 3D Visual Grounding and VQA for both experts and LMMs. *2D VLM numbers obtained from prior works [3, 32, 38, 47]

	Methods	ScanRefer		ScanQA (Val)				SQA3D (Test)
		Acc @25	Acc @50	EM@1	C	M	R	EM@1
Experts	BUTD-DETR [20]	52.2	39.8	-	-	-	-	-
	ScanQA [4]	-	-	23.5	67.3	13.6	34.3	-
	PQ3D [52]	56.7	51.8	20.0	65.2	13.9	-	47.1
	3D-VisTA [51]	51.0	46.2	22.4	69.6	13.9	35.7	48.5
	ODIN [21]	43.1	33.4	-	-	-	-	-
	Locate-3D [3]	61.1	50.9	-	-	-	-	-
	UniVLG [22]	63.5	56.4	25.7	78.5	15.2	40.0	50.2
LMMs	2D VLMs*							
	LLaMA [2]	37.3	24.2	-	-	-	-	-
	Qwen2-VL-2B [41]	35.9	31.9	-	-	-	-	-
	Qwen2-VL-72B [41]	44.8	40.3	-	-	-	-	-
	Qwen2.5-VL-3B [44]	36.4	11.8	-	-	-	-	-
	Qwen2.5-VL-7B [44]	41.6	14.9	-	-	-	-	-
	GPT-4o [35]	48.2	32.5	18.0	58.3	14.2	33.4	-
	Two-Stage							
	ChatScene [17]	55.5	50.2	21.6	87.7	18.0	41.6	54.6
	LEO [18]	-	-	24.5	101.4	20.0	49.2	50.0
	Video-3D LLM - 7B [48]	58.1	51.7	30.1	102.1	-	-	58.6
	Single-Stage							
	NaviLLM [49]	-	-	23.0	75.9	15.4	38.4	-
	LL3DA [10]	-	-	-	76.8	15.9	37.3	-
	SceneLLM [14]	-	-	27.2	80.0	16.6	40.0	54.2
	3D-LLM [15]	30.3	-	20.5	69.4	14.5	35.7	-
	Grounded 3D-LLM [11]	48.6	44.0	-	75.9	-	-	-
	Reason3D [19]	49.6	41.1	-	73.5	15.1	37.4	-
	LLaVA-3D - 7B [50]	50.1	42.7	27.0	103.1	20.8	49.6	55.6
	Qwen-3D - 3B (Ours)	62.2	55.3	28.0	92.0	36.4	46.4	55.4

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], ScanRe-

fer [9]), instance segmentation (ScanNet200 [36], Matterport [8]), 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 [24]), instance segmentation (COCO [30]), captioning and QA (LLaVA-Instruct-150k [31]), and instruction fine-tuning (Alpaca [39]).

4.1 Evaluation on 3D Referential Grounding

Datasets. We evaluate on the validation sets of three ScanNet-based [13] 3D referential grounding benchmarks: SR3D, NR3D [1], and ScanRefer [9]. While SR3D comprises 88k synthetic utterances, NR3D (41k) and ScanRefer (51k) feature complex, human-annotated queries. Following recent work [3, 6, 21, 22], 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 [21, 22], 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 [48, 50], we compare Qwen-3D against state-of-the-art expert (non-LMM) and LLM-based approaches. Expert baselines include two-stage [51, 52] and single-stage [3, 22] methods. LLM-based baselines include: (i) two-stage models relying on detector proposals (LEO [18], Chat-Scene [17], Video-3D-LLM [48]); (ii) single-stage text-space decoders (3D-LLM [15]); and (iii) single-stage $< REF >$ token decoders (Grounded 3D-LLM [11], LLaVA-3D [50], Reason3D [19]). 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 [41]). Tab. 1 presents quantitative results, and full results on ReferIt3D are available in the appendix.

Qwen-3D establishes a new state-of-the-art among 3D LMMs. It surpasses the text-decoding single-stage model of 3D-LLM [15] by over 30%, the recent single-stage state-of-the-art LLaVA-3D [50] by 12%, and the two-stage Video-3D-LLM [48] by 4%. This establishes Qwen-3D as the new state-of-the-art for 3D referential grounding among LMM-based models.

Qwen-3D closes the gap with expert 3D grounding models. On ScanRefer [9], Qwen-3D closely matches the state-of-the-art expert model UniVLG, substantially narrowing the gap between specialist 3D models and LMM-based approaches, and outperforming all other methods.

Table 2: Evaluation on Locate3D ScanNet++ and ScanNet200 Instance Segmentation.

(a) Locate3D ScanNet++			(b) ScanNet200 Instance Segmentation		
Model	Acc@25	Acc@50		Model	mAP mAP25
UniVLG [22]	32.3	24.6	Closed Vocabulary	Mask3D [37]	27.4 42.3
Video-3D LLM [48]	33.2	27.6		PQ3D [52]	27.0 46.3
Qwen-3D (Ours)	55.7	43.4		MAFT [27]	29.2 43.3
				ODIN [21]	31.5 53.1
			Language-Prompted	PQ3D [52]	20.2 32.5
				UniVLG [22]	27.9 46.1
			LLM-Based	Grounded-3D-LLM [11]	12.1 16.8
				Qwen-3D (Ours)	25.3 41.5

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++ [46] 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 [21] box proposals). We omit LLaVA-3D [50] as its grounding model weights and code are not publicly released. As shown in Tab. 2a, Qwen-3D significantly outperforms both baselines on these OOD tasks. We attribute the improved 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 [36] instance segmentation benchmark. While traditional methods [21, 37] assume a closed vocabulary setup, recent models [22, 52]—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 [11] being the only exception to our knowledge. As shown in Tab. 2b, Qwen-3D outperforms Grounded-3D-LLM by 13% mAP and 25% mAP25, approaching the performance of language-prompted specialist models.

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 [13], with ScanQA

Table 3: 2D Ref. grounding datasets and RealWorldQA

Model	RefCOCO	RefCOCO+	RefCOCog	RealWorldQA
LAVT [45] (B)	72.7	62.4	61.2	-
ReSTR [25]	67.2	55.7	54.5	-
X-Decoder (L) [53]	-	-	64.6	-
UniVLG [22]	69.2	61.3	64.1	-
Qwen-2.5-VL-3B [5]	89.1	82.4	85.2	62.6
Qwen-2.5-VL-3B (mask decoder)	89.7	86.4	84.4	-
Qwen-3D (Ours)	88.1	82.7	84.0	60.4

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 [22], and achieves comparable performance to the single-stage LMM state-of-the-art LLaVA-3D-7B [50].

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 [24] and LLaVA-Instruct-150k [31]. Tab. 3 evaluates Qwen-3D against its pre-trained base model (Qwen2.5-VL 3B [5]) on RefCOCO 2D grounding benchmarks and the held-out RealWorldQA [43] dataset. As in our 3D setup, Qwen-3D predicts segmentation masks which we convert to bounding boxes for evaluation. Results show that Qwen-3D preserves the strong 2D capabilities of Qwen2.5-VL despite additional 3D fine-tuning, indicating that geometry-aware reasoning can be added without sacrificing the general multimodal capabilities of the base model.

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

Table 4: Ablations of Qwen-3D

(a) Text vs. < REF >		(b) Pos. Embed.		(c) Attn. Mask		(d) VLM Tuning	
Model	Top1 Acc.	Pos. Embed.	Top1 Acc.	Mask Type	Top1 Acc.	Finetune	Top1 Acc.
Grounded 3D-LLM	44.0	2D	53.2	Causal	36.4	Frozen	37.0
LLaVA-3D - 7B	42.7	Vanilla 3D	49.9	Full	53.5	Finetune	53.5
Qwen-3D < REF >	39.7	Aligned 3D (ours)	53.5				
Qwen-3D (ours)	53.5						

We ablate our key design choices on ScanRefer (Top-1@0.5) utterances from 50 randomly sampled scenes from the validation set.

Geometry-Aware Decoding: 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 $\langle \text{REF} \rangle$ into the mask decoder, similar to LLaVA-3D [50] and Grounded-3D-LLM [11]. As shown in Tab. 4a, with the $\langle \text{REF} \rangle$ token, Qwen-3D underperforms both LLaVA-3D and Grounded-3D-LLM, however, Qwen-3D with the full text and visual token interface greatly outperforms all $\langle \text{REF} \rangle$ token baselines.

Geometry-Aware Attention: 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. 4b, our 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. 4c, the full-attention variant significantly outperforms the causal-mask variant.

Freezing vs. Fine-tuning the Base VLM: As shown in Tab. 4d, 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

While Qwen-3D performs attention in a shared 3D coordinate system, we assume that this coordinate system is largely static over time. Dynamic environments introduce non-rigid motion, object interactions, and topology changes that cannot be captured by a single persistent scene representation. Extending geometry-aware attention and grounding to dynamic 4D scene representations 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.

While Qwen-3D grounds language directly into 3D scene representations, most tasks considered in this work involve a single grounding step. Many embodied and interactive applications require compositional reasoning over multiple grounded entities, such as spatial comparisons, relational reasoning, and long-horizon instruction following. Integrating geometry-aware grounding with iterative reasoning frameworks may enable richer forms of scene understanding and planning.

Qwen-3D relies on externally estimated depth maps and camera poses to construct its world-space representation. While our experiments demonstrate

robustness to substantial depth and pose noise, errors in geometric reconstruction ultimately limit downstream grounding performance. An interesting future direction is to jointly optimize geometry estimation and language-guided scene understanding within a single end-to-end framework.

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 3D Rotary Positional Embeddings to perform attention in the world space, 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 demonstrate that Qwen-3D substantially advances the state of the art among general-purpose 3D LMMs, improving grounding, segmentation, and visual question answering while preserving strong 2D capabilities. More broadly, our results suggest that geometry-aware representations alone are insufficient for high-quality 3D perception. Closing the gap between language reasoning and dense geometric prediction requires equally strong geometry-aware decoding mechanisms. We believe that tightly coupling world-space representations with dense grounding modules provides a promising foundation for future multimodal models that reason seamlessly across images, videos, and 3D environments.

7 Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No(s) DGE2140739, an NSF Career award, ONR award N00014-23-1-2415, AFOSR Grant FA9550-23-1-0257. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Ayush Jain is supported in part by the Meta AI Mentorship Fellowship. The authors thank Gabriel Sarch and Akshan Agrawal for proofreading and helpful discussions.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In: Proc. ECCV (2020)
2. AI, M.: Llama 3: The llama-3 herd of models. <https://ai.meta.com/llama/> (2024), large language model
3. Arnaud, S., McVay, P., Martin, A., Majumdar, A., Jatavallabhula, K.M., Thomas, P., Partsey, R., Dugas, D., Gejji, A., Sax, A., Berges, V.P., Henaff, M., Jain, A., Cao, A., Prasad, I., Kalakrishnan, M., Rabbat, M., Ballas, N., Assran, M., Maksymets, O., Rajeswaran, A., Meier, F.: Locate 3d: Real-world object localization via self-supervised learning in 3d (2025), <https://arxiv.org/abs/2504.14151>
4. Azuma, D., Miyanishi, T., Kurita, S., Kawanabe, M.: Scanqa: 3d question answering for spatial scene understanding. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19129–19139 (2022)
5. Bai, S., Chen, K., Liu, X., et. al., J.W.: Qwen2.5-vl technical report (2025), <https://arxiv.org/abs/2502.13923>
6. Cao, A., Arnaud, S., Maksymets, O., Yang, J., Jain, A., Yenamandra, S., Martin, A., Berges, V.P., McVay, P., Partsey, R., Rajeswaran, A., Meier, F., Johnson, J., Park, J.J., Sax, A.: From thousands to billions: 3d visual language grounding via render-supervised distillation from 2d vlms (2025), <https://arxiv.org/abs/2502.20389>
7. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-End Object Detection with Transformers. In: Proc. ECCV (2020)
8. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
9. Chen, D.Z., Chang, A., Nießner, M.: ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In: Proc. ECCV (2020)
10. Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T.: Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26428–26438 (2024)
11. Chen, Y., Yang, S., Huang, H., Wang, T., Xu, R., Lyu, R., Lin, D., Pang, J.: Grounded 3d-llm with referent tokens. arXiv preprint arXiv:2405.10370 (2024)
12. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation (2022)
13. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
14. Fu, R., Liu, J., Chen, X., Nie, Y., Xiong, W.: Scene-llm: Extending language model for 3d visual understanding and reasoning. arXiv preprint arXiv:2403.11401 (2024)
15. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems* **36**, 20482–20494 (2023)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), <https://arxiv.org/abs/2106.09685>
17. Huang, H., Chen, Y., Wang, Z., Huang, R., Xu, R., Wang, T., Liu, L., Cheng, X., Zhao, Y., Pang, J., et al.: Chat-scene: Bridging 3d scene and large language

- models with object identifiers. *Advances in Neural Information Processing Systems* **37**, 113991–114017 (2024)
18. Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.C., Jia, B., Huang, S.: An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871* (2023)
 19. Huang, K.C., Li, X., Qi, L., Yan, S., Yang, M.H.: Reason3d: Searching and reasoning 3d segmentation via large language model. *3DV* (2025)
 20. Jain, A., Gkanatsios, N., Mediratta, I., Fragkiadaki, K.: 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)
 21. Jain, A., Katara, P., Gkanatsios, N., Harley, A.W., Sarch, G., Aggarwal, K., Chaudhary, V., Fragkiadaki, K.: Odin: A single model for 2d and 3d segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3564–3574 (2024)
 22. Jain, A., Swerdlow, A., Wang, Y., Arnaud, S., Martin, A., Sax, A., Meier, F., Fragkiadaki, K.: Unifying 2d and 3d vision-language understanding (2025), <https://arxiv.org/abs/2503.10745>
 23. Kamath, A., Singh, M., LeCun, Y.A., Misra, I., Synnaeve, G., Carion, N.: MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. In: *Proc. ICCV* (2021)
 24. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: Referring to objects in photographs of natural scenes. In: Moschitti, A., Pang, B., Daelemans, W. (eds.) *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 787–798. Association for Computational Linguistics, Doha, Qatar (Oct 2014). <https://doi.org/10.3115/v1/D14-1086>, <https://aclanthology.org/D14-1086/>
 25. Kim, N., Kim, D., Lan, C., Zeng, W., Kwak, S.: Restr: Convolution-free referring image segmentation using transformers (2022), <https://arxiv.org/abs/2203.16768>
 26. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model (2024), <https://arxiv.org/abs/2308.00692>
 27. Lai, X., Yuan, Y., Chu, R., Chen, Y., Hu, H., Jia, J.: Mask-attention-free transformer for 3d instance segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3693–3703 (2023)
 28. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: *International conference on machine learning*. pp. 19730–19742. PMLR (2023)
 29. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10965–10975 (2022)
 30. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. pp. 740–755. Springer (2014)
 31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *Advances in neural information processing systems* **36**, 34892–34916 (2023)
 32. Liu, Y., Mei, H., Zhang, L., 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)

33. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 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 (Jun 2022). <https://doi.org/10.1109/cvpr52688.2022.01596>, <http://dx.doi.org/10.1109/CVPR52688.2022.01596>
34. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. arXiv preprint arXiv:2210.07474 (2022)
35. OpenAI, Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., et al., A.C.: Gpt-4o system card (2024), <https://arxiv.org/abs/2410.21276>
36. Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: European Conference on Computer Vision. pp. 125–141. Springer (2022)
37. Schult, J., Engelmann, F., Hermans, A., Litany, O., Tang, S., Leibe, B.: Mask3d: Mask transformer for 3d semantic instance segmentation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 8216–8223. IEEE (2023)
38. Singh, S., Pavlakos, G., Stamoulis, D.: Evaluating zero-shot gpt-4v performance on 3d visual question answering benchmarks (2024), <https://arxiv.org/abs/2405.18831>
39. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. <https://crfm.stanford.edu/2023/03/13/alpaca.html> (2023), dataset and model release
40. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., et al., R.S.: Gemini: A family of highly capable multimodal models (2025), <https://arxiv.org/abs/2312.11805>
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., et. al., J.B.: Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution (2024), <https://arxiv.org/abs/2409.12191>
42. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. arXiv preprint arXiv:2410.19115 (2024)
43. X.AI: Realworldqa. Blog post (Nov 2024), accessed: 2025-05-12
44. Yang, A., Yang, B., Zhang, B., Hui, B., et. al., B.Z.: Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024)
45. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.S.: Lavt: Language-aware vision transformer for referring image segmentation (2022), <https://arxiv.org/abs/2112.02244>
46. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
47. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. arXiv preprint arXiv:2505.24625 (2025)
48. Zheng, D., Huang, S., Wang, L.: 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). pp. 8995–9006 (June 2025)
49. Zheng, D., Huang, S., Zhao, L., Zhong, Y., Wang, L.: Towards learning a generalist model for embodied navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13624–13634 (2024)

- 50. Zhu, C., Wang, T., Zhang, W., Pang, J., Liu, X.: Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness (2024), <https://arxiv.org/abs/2409.18125>
- 51. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2911–2921 (2023)
- 52. Zhu, Z., Zhang, Z., Ma, X., Niu, X., Chen, Y., Jia, B., Deng, Z., Huang, S., Li, Q.: Unifying 3d vision-language understanding via promptable queries. arXiv preprint arXiv:2405.11442 (2024)
- 53. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)