3D Aware Region Prompted Vision Language Model

An-Chieh Cheng[1], **Yang Fu**[1], **Yukang Chen**[3], **Zhijian Liu**[3], **Xiaolong Li**[3], **Subhashree Radhakrishnan**[3], **Song Han**[2,3], **Yao Lu**[3], **Jan Kautz**[3], **Pavlo Molchanov**[3], **Hongxu Yin**[3,*], **Xiaolong Wang**[1,*], **Sifei Liu**[3,*]

[1]UC San Diego, [2]MIT, [3]NVIDIA

*Equal Advising

We present **S**patial **R**egion **3D** (*SR-3D*) aware vision–language model that connects single-view 2D images and multi-view 3D data through a shared visual token space. SR-3D supports flexible region prompting, allowing users to annotate regions with bounding boxes, segmentation masks on any frame, or directly in 3D, without the need for exhaustive multi-frame labeling. We achieve this by enriching 2D visual features with 3D positional embeddings, which allows the 3D model to draw upon strong 2D priors for more accurate spatial reasoning across frames, even when objects of interest do not co-occur within the same view. Extensive experiments on both general 2D vision language and specialized 3D spatial benchmarks demonstrate that SR-3D achieves state-of-the-art performance, underscoring its effectiveness for unifying 2D and 3D representation space on scene understanding. Moreover, we observe applicability to in-the-wild videos without sensory 3D inputs or ground-truth 3D annotations, where SR-3D accurately infers spatial relationships and metric measurements.

🏠 **Project:** https://www.anjiecheng.me/sr3d

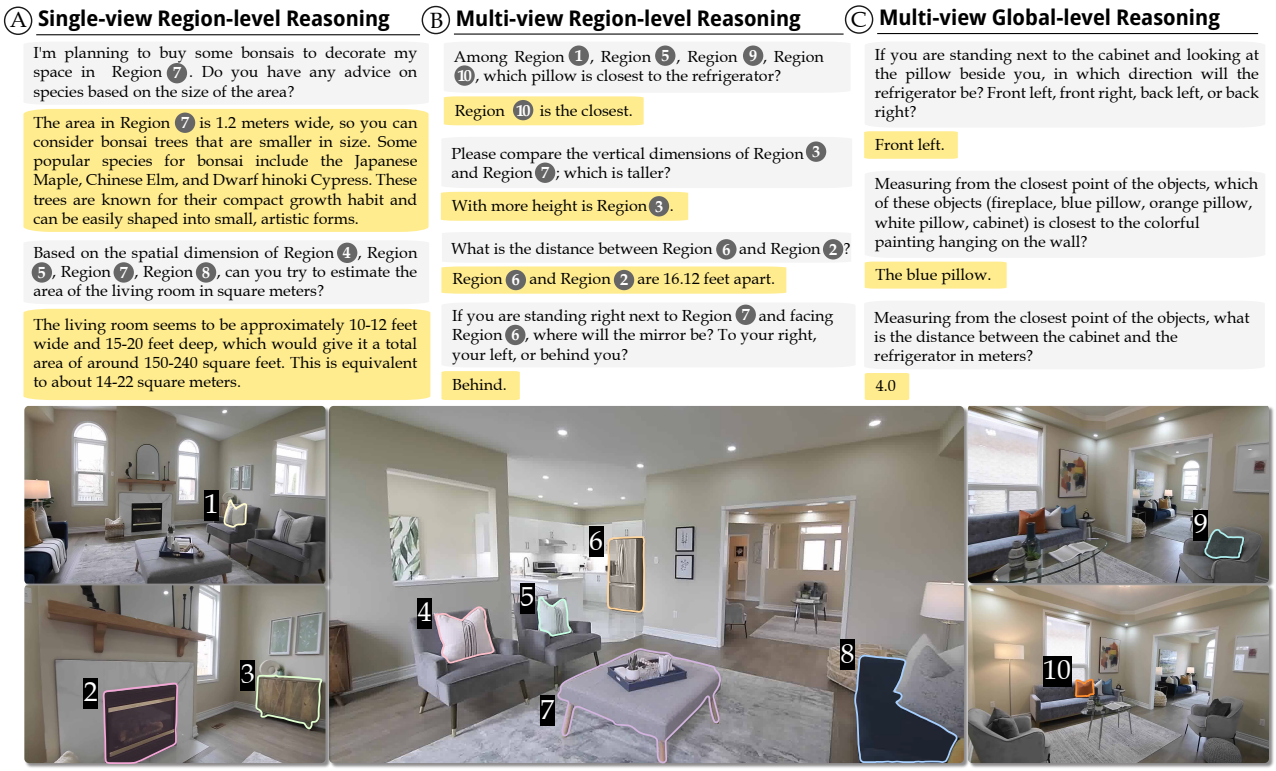

Figure 1 From precise region-based distance estimation (*left*), to intricate multi-view region query (*middle*), and global cross-frame reasoning (*right*), SR-3D delivers flexible and accurate spatial understanding to foundational Vision-Language Models. Notably, this video is obtained in the wild, **without sensory 3D inputs**, showcasing the remarkable generalization capability of our model.

1 Introduction

The rapid advancement of Vision Language Models (VLMs) [1–6] has demonstrated strong capabilities in visual understanding [7, 8] and language grounding [9]. However, extending these strengths to 3D-aware spatial reasoning remains challenging. Foundational 2D VLMs excel at interpreting planar images, but generally lack mechanisms to capture complex 3D structural relationships. In contrast, most 3D VLMs [10–14] operate in a fundamentally different representation space, making it difficult to leverage the prior knowledge from foundational 2D VLMs. Their performance is often hindered by limited 3D training data. Moreover, specifying spatial relationships solely through language can be cumbersome in cluttered scenes, e.g., multiple objects of the same category can coexist. A more direct way of specifying object instances is highly desirable.

To mitigate these challenges, recent efforts adopt multi-view images as a 3D representation that aligns seamlessly with the input space of foundational 2D VLMs [15, 16]. Unlike point clouds [11–13] which require extensive data collection and model alignment, a multi-view approach leverages strong 2D priors for 3D scene understanding. To specify object instances during reasoning, region prompts have proven effective in single-view VLMs [17–20]. However, extending region prompting to multi-view settings remains challenging. Specifically, an object may appear across different views with varying visibility, making comprehensive multi-frame or 3D bounding box annotation tedious and text-based queries imprecise. Ideally, a practical 3D-aware VLM should allow straightforward region annotations, such as marking a bounding box on a single frame, while still accurately reasoning about spatial relationships across the entire multi-view scene.

Thus, we introduce SR-3D, a unified visual representation for 3D spatial understanding that leverages robust 2D foundational priors and supports flexible region prompting. In contrast to previous approaches that incorporate positional information only at 3D finetune stages [16], or in different pathways [15], we directly integrate positional embeddings within the foundational VLM. Specifically, we estimate each input image's depth using an off-the-shelf depth estimator [21] and transform this depth map into normalized 3D positional embeddings. For multi-view inputs representing a coherent scene, we further unify these positional embeddings into a common 3D coordinate space using either provided ground-truth camera poses or a point cloud estimator [22–24] when only video inputs are available. Additionally, we incorporate region tokens directly into user prompts and train these region embeddings consistently at both the foundational single-view stages and the multi-view fine-tuning stage. Since the foundational VLM employs a dynamic tiling-based visual encoder [6, 25], we design a novel branch specifically compatible with this architecture to produce robust region embeddings.

The SR-3D architecture naturally supports flexible region annotation, enabling users to specify regions on any chosen frame. This practical capability arises from two key design choices: first, the consistent 3D positional embeddings in a canonical space enable the model to find coherent correspondences across frames; Second, the aligned embedding space from the foundational single-view stage naturally enables region embeddings to generalize effectively to the multi-frame contexts. As compelling evidence, our 2D-VLM trained exclusively on single-view data exhibits strong zero-shot spatial reasoning in 3D scenes, both with and without region prompts, despite never having been trained on multi-view data.

We conduct extensive evaluations across single-view and 3D multi-view settings, covering both region-level and global question-answering, each with general and spatial-related tasks. Our experiments demonstrate significant improvements in region-level performance. Specifically, our foundational 2D-VLM outperforms prior state-of-the-art methods by a large margin on region-level tasks, excelling in both recognition and spatial understanding. Additionally, we evaluate it on general VQA benchmarks and show that these improvements come without compromising overall VQA performance while also

bringing benefits for general tasks that require spatial knowledge. For the 3D fine-tuned VLM, our model establishes new state-of-the-art results across general 3D question-answering, 3D video spatial understanding, and video region-level spatial tasks.

Our contributions are as follows:

- We introduce SR-3D, the first 3D-aware vision-language model that unifies representations for both single-view and multi-view tasks.
- We propose a dynamic tiling-based region extractor that handles high-resolution images and produces robust region embeddings. Our unified embedding space enables region representations trained on 2D images to generalize towards multi-view context.
- SR-3D achieves state-of-the-art results in general 3D QA, video spatial reasoning, and region-based video tasks, demonstrating strong generalization and scalability.
- We demonstrate real-world applications where our model effectively handles in-the-wild captured videos without 3D annotations (Figure 1), and can be flexibly prompted with region-level inputs.

2 Related Work

Region-level Vision-Language Models. Region-level VLMs enhance fine-grained visual understanding by focusing on specific regions in images and videos. Early methods [26–29] represent regions as text using bounding box coordinates, making integration easy but relying on the language decoder for spatial reasoning. Others use visual markers like SoM [30], which overlay numbers and masks but alter image appearance and require rule-based placement. Another approach maps region features into LLM tokens using RoI-aligned features [20, 31–36], with RegionGPT [17] and Osprey [19] refining this by pooling pixel-level mask features for flexible region shapes. However, they struggle with resolution and aspect ratio constraints. In the video domain, various representations [37–41] have been explored, but they mainly focus on tracking rather than multi-view spatial reasoning.

Spatial Reasoning in Vision-Language Models. Vision-language models have a strong visual understanding because they integrate the reasoning abilities of LLMs with powerful vision foundation models. Recently, there has been growing interest in equipping VLMs with spatial reasoning capabilities [42–53]. While most previous work has focused on spatial understanding from 2D images, multi-view spatial reasoning remains less explored. Recently, VSI-Bench [54] was introduced as a testbed for evaluating models' 3D video-based spatial understanding. Our work extends this direction by proposing a unified 3D-aware architecture and representation that seamlessly supports both images and videos.

3D Large Multimodal Models. Our work also relates to recent advancements in 3D LMMs [10, 11, 14, 55–59]. Various 3D representations have been explored to integrate position information into LLMs. 3D-LLM [10] and Scene-LLM [58] use multi-view images with object segmentation masks to construct pixel-aligned point representations, while LL3DA [14] directly employs a point cloud encoder to extract 3D scene features. LEO [11] and Chat3D [59] segment objects from the scene's point cloud and extract object features to represent the environment. These methods typically transform 3D scenes into voxel or point representations, but such approaches often limit the effectiveness of LLMs. Aligning these representations with LLMs requires vast amounts of data, which is challenging due to the scarcity of large-scale 3D datasets. Moreover, many of these methods rely on off-the-shelf 3d detection or segmentation models, which inherently constrain performance.

The most closely related works to ours are LLaVA-3D [15] and Video-3D-LLM [16], which also incorporate 3D position-aware features into 2D vision-language models. However, LLaVA-3D processes 3D and

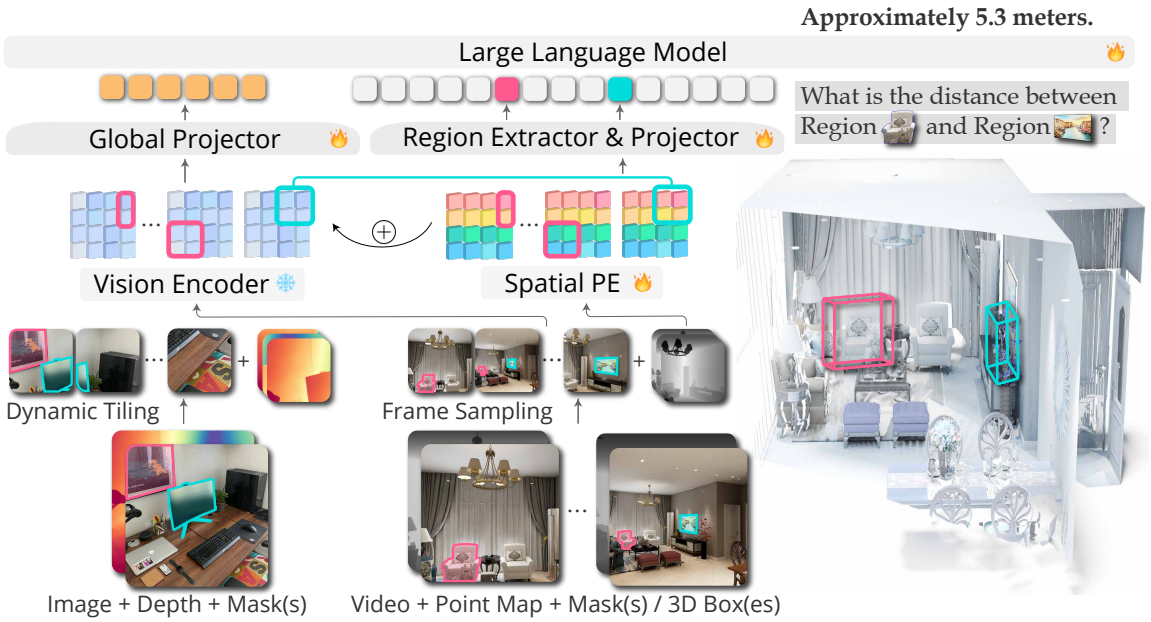

Figure 2 The SR-3D architecture. Given an image or multi-view input with optional region prompts (e.g., bounding boxes or masks), we encode them along with depth-derived positional embeddings using a tiling approach. Region tokens are extracted by stitching masked features, while 3D positional embeddings are mapped to a shared canonical space in the multi-view setting, as shown on the bottom right.

2D data through separate pathways, while Video-3D-LLM fine-tunes 3D video data on a pre-trained video VLM. Both approaches risk overfitting 3D position encodings to specific 3D tasks. In contrast, our method adopts a unified architecture and 3D representation space for both image and video data, enabling better alignment and improving generalization across spatial understanding tasks.

3 Methodology

We introduce a unified 3D-aware VLM architecture designed for both single-view and multi-view spatial understanding. Our approach leverages the strong priors of a foundational 2D model to infer spatial relationships across frames accurately. This is achieved by directly integrating 3D positional embeddings into the foundational 2D visual representations. To further enhance spatial grounding at the region level, we introduce a flexible and efficient module, the Dynamic Tiling-based Region Extractor, which operates seamlessly across both single- and multi-view inputs. As illustrated in Figure 2, our framework consists of a vision encoder, a 3D position encoding module, a region extractor, and an LLM backbone. In this section, we detail three key components: (1) a canonical 3D positional representation for single- and multi-view images (Sec.3.1), (2) the region extractor (Sec.3.2), and (3) the training paradigm (Sec.3.3), along with how our model operates during inference (Sec. 3.4).

3.1 Canonical 3D Positional Representation

A key idea of SR-3D is the introduction of a canonical positional feature that is shared across both single-view and multi-view inputs. This unified representation allows us to leverage large-scale single-view image pretraining, while seamlessly transferring the learned spatial priors to multi-view scenarios.

Single-View Representation. We begin by pretraining our foundational VLM on large-scale 2D images to establish strong visual-language priors. Given a single-view image I, we estimate its relative depth map D using DepthAnythingV2 [21]. We then compute a pixel-wise 3D position map in the

camera coordinate system via back-projection, which is further canonicalized into a normalized world coordinate system. This canonicalization ensures that spatial information is expressed in a consistent and unified space, independent of camera pose.

To inject spatial information into VLM, we encode the corresponding 3D position map into embeddings using a sinusoidal function followed by a learnable point-wise MLP. These embeddings are resized to align with the token dimensions and then added to their respective vision tokens. This fusion enriches visual representations with geometric awareness, enabling the model to better capture object placement and spatial relationships within the scene.

Multi-View Representation. Building on the shared canonical space, we fine-tune the VLM with multi-view inputs to extend spatial reasoning beyond single images. We uniformly sample 32 frames from a video and resize both the images and their point maps to match the vision encoder's input resolution. For multi-view training, we use ground-truth depth rather than estimated depth, performing back-projection and camera transformation to align the frames. The transformed point maps are normalized into the same canonical space as in the single-view setup, ensuring consistency in spatial representation. These processed frames and point maps act as the multi-view analog of the single-view tiles, enabling seamless integration of spatial and visual information across both training stages.

3.2 Dynamic Tiling-based Region Extractor

Background: Dynamic Tiling-based Encoder. The visual backbone produces a low-resolution feature map, limiting its ability to represent small-scale regions and objects. To address this, we adopt the dynamic tiling mechanism employed in [6] that enables high-resolution processing while maintaining spatial consistency. Instead of resizing entire images, we first determine the optimal aspect ratio by selecting the closest match from a predefined set (e.g., 1:1, 1:2, 2:1, 3:1, ..., 2:6), minimizing distortions. We then resize both the image and any corresponding point map accordingly and divide them into tiles of 448×448, matching the vision encoder's resolution. Each tile is encoded separately before being stitched back together, preserving local details without exceeding memory constraints. This tiling process is applied consistently across single-view and multi-view inputs, forming the basis for both our 3D positional embedding and region feature extraction strategies.

Dynamic Region Extractor. Prior architectures without dynamic tiling rely on feature refinement modules with deconvolution layers to upsample visual tokens [17, 18], attempting to recover lost details. However, this refinement occurs after the vision encoder, meaning the features have already undergone resizing and potential distortion, which may limit its ability to fully recover fine details.

To address this, we introduce a *tile-then-stitch* approach to extract region embeddings from high-resolution features. For single-view input, given a region of interest (RoI) represented by a binary mask, we apply the same dynamic tiling process used in the image pipeline to generate tiles of both the image and the mask. The tiled visual tokens and masks are then stitched back together at a higher resolution, followed by a mask-pooling operation to obtain the final mask feature. This method offers two key advantages: (1) the extracted mask feature is derived from high-resolution features directly, reducing distortion and eliminating the need for post-refinement, and (2) our tile-then-stitch approach extends naturally to multi-view video inputs. In the multi-view setting, each frame is treated as a tile, allowing us to handle one or multiple masks per frame while maintaining spatial consistency across frames for the same RoI.

Methods	Spatial				Math	General Knowledge					OCR-Related		
	BLINK_S	SAT	EmbSpat	RWQA	MathVista	GQA	AI2D	MMMU_p	SEED_I	POPE	Text_{VQA}	Chart_{QA}	Doc_{VQA}
NVILA-Lite-8B	79.7	62.6	68.9	65.6	64.5	**65.3**	**91.0**	**25.1**	76.3	**88.1**	**78.1**	**84.8**	**91.7**
SR-3D-8B	**83.9**$_{+4.2}$	**64.0**$_{+1.4}$	**72.5**$_{+3.6}$	**68.1**$_{+2.5}$	**65.4**	64.2	90.7	24.6	**77.8**	87.6	77.3	83.9	91.0

Table 1 Comparison of SR-3D and base model [6] performance on general image VQA benchmarks.

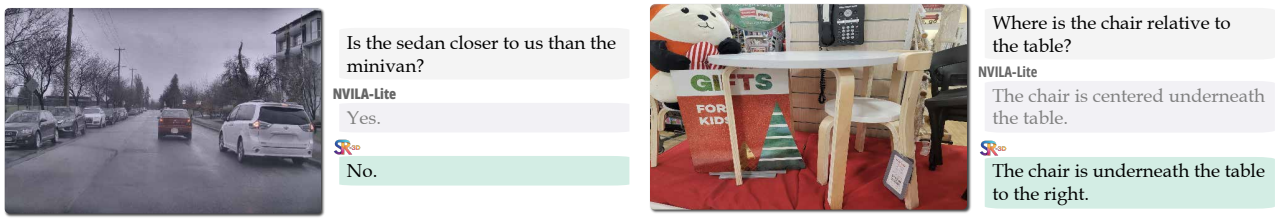

Figure 3 RealWorldQA results. SR-3D shows stronger spatial understanding of physical environments compared to the base model. We omit the answer choices for clarity in visualization.

3.3 Training Paradigm

For the single-view VLM, we initialize the weights from a pre-trained 2D VLM (NVILA-Lite-8B [6]), keeping the vision encoder frozen while fine-tuning the 3D positional encoding module, projectors, and the LLM. We reuse the instruction fine-tuning dataset from the pre-trained VLM and blend it with region-prompted datasets [17, 18] in this stage, resulting in a total data blend of approximately 7 million samples. Full dataset details are provided in the Supplementary Materials.

For the multi-view model, we fine-tune the single-view model using datasets such as ScanQA [60], SQA3D [61], and Scan2Cap [62], as well as a newly curated EmbodiedScan [63] dataset with region- and spatial-focused question-answer pairs. To enhance robustness and generalization, we apply various mask augmentations during multi-view training, including converting segmentation masks into bounding boxes and randomly dropping frames to simulate single-frame annotations. These strategies help the model learn to associate regions across frames while preserving spatial consistency.

We note that, unlike prior work [15] that employs separate pathways for single- and multi-view data, we adopt a unified pipeline where all data flows through the same model architecture. This ensures consistent processing of both single-view and multi-view inputs without distinction between spatial region prompts and global queries, allowing seamless integration of spatial reasoning at different levels.

3.4 Inference

Our tile-and-stitch design enables flexible region-based inference. For single-view inputs, the model accepts bounding boxes or segmentation masks as region annotations. In multi-view scenarios, it supports a range of mask specifications: 3D bounding boxes that project into multi-frame masks, sparse-frame masks, or even a single-frame mask—reflecting our method's ability to handle varying annotation densities while preserving spatial alignment.

For 3D input, although ground-truth depth maps were used during multi-view training, our approach remains highly adaptable due to the canonicalization of 3D positions into a normalized space. This allows us to replace ground-truth depth with point maps estimated from off-the-shelf models such as MAST3R [23] or CUT3R [24]. Our model offers a highly flexible and generalizable solution for spatial reasoning across diverse input modalities by maintaining a unified architecture that normalizes spatial information across different 3D sources.

Methods	Acc. (%)
Human	98.3
Proprietary Models (API)	
Qwen-VL-Max [66]	58.9
Gemini Pro [3]	50.0
Claude 3 OPUS [67]	57.3
GPT-4V-*preview* [1]	58.9
GPT-4V-*Turbo* [1]	66.9
GPT-4o [1]	64.5
Open-source Models	
InstructBLIP-13B [68]	50.0
Yi-VL-34B [69]	53.2
LLaVA-v1.5-13B-xtuner [70]	54.0
LLaVA-v1.6-34B [71]	64.5
MiniGPT-4-v2-7B [28]	49.2
InstructBLIP-7B [68]	50.8
LLaVA-v1.5-7B-xtuner [70]	50.8
CogVLM-7B [29]	50.8
LLaVA-v1.5-7B [72]	51.6
LLaVA-InternLM2-7B [73]	52.4
SpatialRGPT-8B [18]	87.9
SR-3D-8B	**90.3**

Table 2 Results on $\text{BLINK}_{\text{Depth}}$. We follow SpatialRGPT [18]'s protocol to test whether a 3D-injected VLM effectively leverages auxiliary spatial information.

Methods	mAP (↑)	Acc. (%)
CLIP [74]	58.9	-
RegionCLIP [65]	58.3	-
LLaVA-7B [2]	-	40.0
Shikra-7B [27]	-	53.9
GPT4RoI-7B [35]	-	64.0
PVIT-7B [75]	-	64.5
ASM-7B [76]	69.3	-
RegionGPT-7B [17]	70.0	80.6
DynRefer [36]	-	81.2
SpatialRGPT-8B [18]	72.9	82.9
SR-3D-8B	**78.0**	**88.6**

Table 3 Region-level classification results on COCO-2017 val set with ground-truth boxes, following RegionCLIP [65] and RegionGPT [17].

4 Experiments

We first evaluate SR-3D on 2D benchmarks (Section 4.1) to verify whether the introduced positional features improve performance while preserving the generalization of the base single-view model. We then evaluate the multi-view model on 3D benchmarks in Section 4.2. We further show ablation studies in Section 4.4 to analyze the role of pretraining and 3D positional encoding. Finally, we demonstrate that our method can be seamlessly integrated with off-the-shelf 3D geometry foundation models as an application (Section 4.5).

4.1 Evaluation on 2D Benchmarks

Region-level Question Answering. We evaluate our model's object classification performance on the COCO-2017 [64] dataset using mean Average Precision (mAP) and classification accuracy as metrics. Following prior work on region-level recognition [17, 18, 65], we rely on ground-truth boxes for positional information and augment the general prompt with task-specific instructions. As reported in Table 3, SR-3D attains an mAP of 78.0 and an accuracy of 88.6%, demonstrating strong region-level recognition and validating the effectiveness of our region extractor. Compared with SpatialRGPT [18], which is trained on the same region-level data, our model achieves significant gains, largely attributable to the dynamic tiling extractor that provides higher-fidelity regional masks. For reference, we also include DynRefer's RoIAlign (448 variant) [36] as a baseline at the same resolution. Importantly, their proposed strategies are complementary to our approach.

We further evaluate SR-3D on the $\text{BLINK}_{\text{Depth}}$ benchmark [77] using the region-prompts as in SpatialRGPT [18], which tests point-level depth understanding in VLMs. $\text{BLINK}_{\text{Depth}}$ is a challenging task that requires both spatial and regional awareness. We report results in Table 2 showing that SR-3D outperforms current state-of-the-art SpatialRGPT [18], achieving 90% accuracy. These results highlight that our approach excels in region extraction and effectively utilizes the provided 3D-aware input.

General Question Answering. We investigate two key questions: (1) Does incorporating 3D positional information affect general vision-language understanding capabilities? (2) Can it improve performance

Methods	Scan2Cap				ScanQA					SQA3D
	B-4 ↑	Rouge ↑	Cider ↑	Meteor ↑	B-4 ↑	Rouge ↑	Cider ↑	Meteor ↑	EM ↑	EM ↑
Task-specific Specialist										
VoteNet+MCAN [78]	-	-	-	-	6.2	29.8	54.7	11.4	17.3	-
ScanRefer+MCAN [78]	-	-	-	-	7.9	30.0	55.4	11.5	18.6	-
ScanQA [60]	-	-	-	-	10.1	33.3	64.9	13.1	21.0	-
3D-VisTA [79]	34.0	54.3	66.9	27.1	10.4	35.7	69.6	13.9	22.4	-
2D Large Multi-modal Models										
Oryx-34B [80]	-	-	-	-	-	37.3	72.3	15.0	-	-
NaviLLM [81]	-	-	-	-	12.0	38.4	75.9	15.4	23.0	-
LLaVA-Video-7B† [82]	-	-	-	-	3.1	44.6	88.7	17.7	-	-
NaVILA [83]	-	-	-	-	16.9	49.3	102.7	20.1	28.6	-
3D Large Multi-modal Models										
3D-LLM$_{(flamingo)}$ [10]	-	-	-	-	7.2	32.3	59.2	12.2	20.4	-
3D-LLM$_{(BLIP2-flant5)}$ [10]	-	-	-	-	12.0	35.7	69.4	14.5	20.5	-
LL3DA [14]	36.8	55.1	65.2	26.0	13.5	37.3	76.8	15.9	-	-
Chat-3Dv2 [59]	-	-	-	-	14.0	-	87.6	-	-	54.7
LEO [11]	36.9	57.8	68.4	27.7	13.2	49.2	101.4	20.0	24.5	50.0
Scene-LLM [58]	-	-	-	-	12.0	40.0	80.0	16.6	27.2	54.2
ChatScene [12]	36.3	58.1	77.2	28.0	14.3	41.6	87.7	18.0	21.6	54.6
LLaVA-3D [15]	41.1	63.4	79.2	30.2	14.5	50.1	91.7	20.7	27.0	55.6
Video-3D LLM [16]	42.4	62.3	83.8	28.9	16.2	49.0	102.1	19.8	30.1	58.6
SR-3D-8B	**44.7**	**67.3**	**97.9**	**31.5**	**18.1**	**51.2**	**109.3**	**21.2**	**30.4**	**62.2**

Table 4 Evaluation of spatial scene understanding performance on the Scan2Cap, ScanQA, and SQA3D benchmarks. † indicates methods evaluated in a zero-shot setting. SR-3D achieves state-of-the-art results across all metrics.

on spatial-related tasks? To answer these, we evaluate our model on general VLM benchmarks covering Spatial [77, 84–86], Math [87], General Understanding [88–92], and OCR-related [93–95] tasks. As shown in Table 1, compared to the base model NVILA-Lite-8B [6], our model maintains comparable performance in math, general understanding, and OCR-related tasks, confirming that integrating 3D positional information does not degrade overall vision-language capabilities. Additionally, our method improves performance on the spatial understanding benchmark RealWorldQA [84]. We also provide qualitative examples from RealWorldQA in Figure 3, showcasing cases where NVILA-Lite fails while SR-3D succeeds in spatial reasoning tasks. These results demonstrate that our 3D-aware VLM enhances spatial reasoning while preserving general vision-language capabilities.

4.2 Evaluation on 3D Benchmarks

General 3D Question Answering. We report results on three classic 3D vision-language understanding tasks: 3D dense captioning on Scan2Cap [62], ScanQA [60], and SQA3D [61]. Our evaluation metrics include conventional scores (e.g., CIDEr, BLEU, METEOR, ROUGE) as well as exact-match (EM) accuracy. Following prior work, we assume that input scenes may lack 3D object mask annotations during inference and use off-the-shelf models to generate proposals. However, unlike previous approaches, we leverage 2D segmentation models to generate 2D object proposals instead. We compare NaVILA against strong baselines, including task-specific specialist models for each benchmark and leading methods from both 2D and 3D large multimodal models (LMMs). NaVILA significantly outperforms state-of-the-art single-task and task-specific fine-tuned models on 3D dense captioning and 3D QA tasks.

4.3 Video Spatial Intelligence.

Region-level Spatial QA. Currently, no video benchmarks specifically focus on region-level spatial understanding. Without explicit region information, spatial understanding can become ambiguous, especially when multiple identical objects are present or when referring to a specific area in a scene that is difficult to describe precisely using language alone. To address this, we propose SR-3D-Bench,

Methods	Wide/Thin	Tall/Short	Big/Small	Multi. Simple	Multi. Complex	Avg.	Width	Height	Distance	Avg.
	Qualitative						Quantitative			
Blind LLMs w/ Language Referral										
GPT-4o [1]	64.8	64.5	64.0	47.8	41.4	56.5	70.5	70.6	50.4	63.8
VLMs w/ Language Referral										
GPT-4o [1]	52.1	54.1	57.5	62.4	42.4	53.7	72.4	72.8	55.8	67.0
NVILA-Video-8B [6]	48.8	38.9	53.7	52.1	36.0	45.9	59.2	54.3	6.6	40.0
Region VLMs										
GPT-4o [1]+SoM	46.1	39.9	39.3	52.1	43.2	44.1	52.4	47.8	40.0	46.7
NVILA-Video-8B [6]+SoM	49.3	40.0	53.7	52.1	40.4	47.1	59.3	54.1	6.6	40.0
SR-3D-8B	**76.3**	**83.1**	**81.8**	**80.3**	**76.0**	**79.5**	**87.7**	**87.3**	**74.8**	**83.3**

Table 5 Evaluation of region-level spatial scene understanding on the SR-3D-Bench.

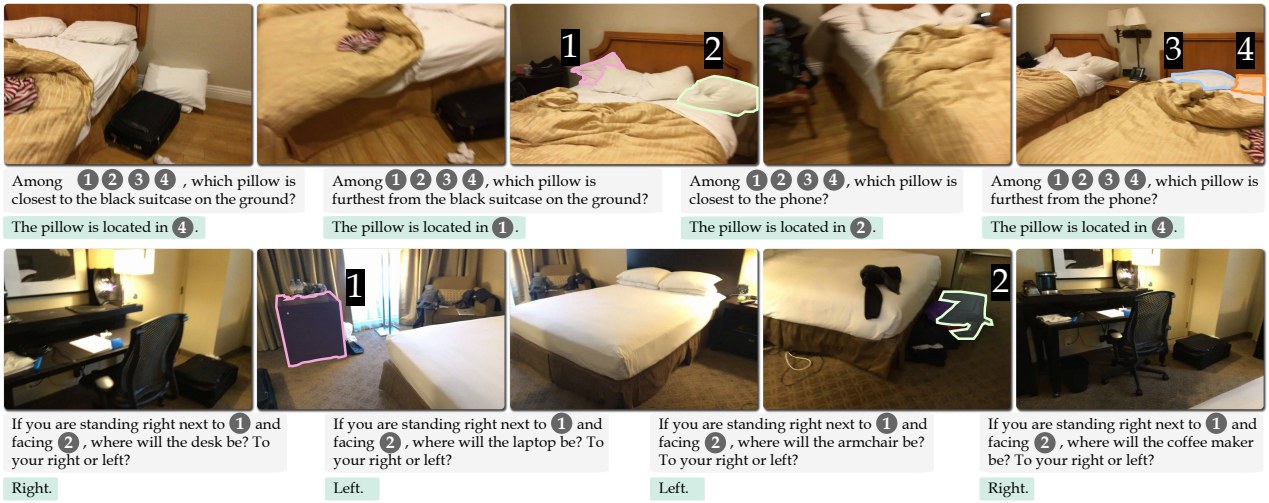

Figure 4 SR-3D results on region-level multi-view spatial understanding. We show extreme cases where the same region prompts are used across samples but with different target objects. SR-3D answers all queries correctly, showing strong evidence that it truly understands 3D spatial relationships.

a region-level spatial benchmark curated from ScanNet [96], ARKitScenes [97], and Matterport[98] video scan datasets with 3D ground truth. Specifically, we utilize preprocessed oriented bounding box annotations from EmbodiedScan [63], where each object is axis-aligned within a canonicalized geodetic coordinate system. This alignment ensures that the bounding box dimensions accurately represent the true width, length, and height. Using these bounding boxes, we construct a conversational benchmark that includes both qualitative and quantitative question-answering tasks. The qualitative QA consists of choice-based, predicate-based, and multiple-choice questions, while the quantitative QA focuses on measuring object width, height, and distance. We generate these QA pairs using template-based conversation generation and allow the VLM to generate free-form language. For qualitative QA evaluation, we use GPT-4o [1] as an evaluator and report the accuracy, while for quantitative QA, we measure the success rate by thresholding the maximum ratio between estimation and the ground truth value.

We report three types of baseline models: (1) Blind LLMs, which answer questions using only the provided text without visual input. To improve this, we replace the mask prompt with the object class for each question. This serves as a baseline to measure how much video spatial reasoning can come from general world knowledge alone. We use GPT-4o as the representative, as it is one of the most

Methods	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.
	Quantitative			Qualitative	
Random	-	-	-	25.0	36.1
Human Level†	47.0	60.4	45.9	94.7	95.8
Proprietary Models (API)					
GPT-4o [1]	5.3	43.8	38.2	37.0	41.3
Gemini-1.5 Flash [100]	30.8	53.5	54.4	37.7	41.0
Gemini-1.5 Pro [100]	30.9	64.1	43.6	51.3	46.3
Open-source Models					
InternVL2-2B [101]	24.9	22.0	35.0	33.8	44.2
InternVL2-8B [101]	28.7	48.2	39.8	36.7	30.7
InternVL2-40B [101]	26.9	46.5	31.8	42.1	32.2
LongVILA-8B [102]	9.1	16.7	0.0	29.6	30.7
VILA-1.5-8B [103]	21.8	50.3	18.8	32.1	34.8
VILA-1.5-40B [103]	24.8	48.7	22.7	40.5	25.7
LongVA-7B [104]	16.6	38.9	22.2	33.1	43.3
LLaVA-NeXT-Video-7B[71]	14.0	47.8	24.2	43.5	42.4
LLaVA-NeXT-Video-72B [71]	22.8	57.4	35.3	42.4	36.7
LLaVA-OneVision-0.5B [105]	28.4	15.4	28.3	28.9	36.9
LLaVA-OneVision-7B [105]	20.2	47.4	12.3	42.5	35.2
LLaVA-OneVision-72B [105]	23.9	57.6	37.5	42.5	39.9
SR-3D-8B	**52.8**	**75.5**	**41.9**	**57.3**	**82.3**

Table 6 Results on multi-view global spatial scene understanding evaluated on VSI-Bench [54]. † indicates methods tested on the `Tiny` subset. SR-3D achieves strong performance on the relative direction task, providing clear evidence that the model effectively leverages the 3D positional encoding.

advanced models for general knowledge. (2) VLMs with Language Referral, which have access to visual content, allowing them to potentially perform better than blind LLMs. We use state-of-the-art vision-language models GPT-4o [1] and NVILA-Video [6] as baselines in this category. (3) Region-aware Video VLMs. These models process specific image regions without relying on text descriptions or object class information. We equip GPT-4o and NVILA-Video with Set of Marks (SoM) for region-based reasoning. Note that while [99] and [37] are also region-level video VLMs, they are excluded from comparisons as they cannot handle multi-object input or lack support for multi-frame prompts.

We present results in Table 5. The findings suggest that both Blind LLMs and VLMs with Language Referral perform reasonably well on quantitative tasks, such as estimating object width, due to their general world knowledge. However, region-level VLMs equipped with SoM struggle, likely because the models find it challenging to track the set of marks across frames. Overall, our method outperforms all baselines across all categories.

Global Spatial QA. We also report results on global spatial understanding using VSI-Bench [54], a recently proposed benchmark that quantitatively evaluates the visual-spatial intelligence of VLMs based on egocentric videos. In our evaluation, we exclude categories that are less relevant to spatial reasoning, such as appearance order, which is more about temporal understanding. We use accuracy as the evaluation metric for qualitative questions and Mean Relative Accuracy (MRA) for quantitative questions. As shown in Table 6, SR-3D outperforms all open-source models and performs comparably, if not better, than API-based models.

4.4 Analysis and Ablation Study

Zero-shot Generalization. In this analysis, we aim to answer the question: Can a foundational 2D VLM trained exclusively on single-view image data perform zero-shot spatial reasoning on multi-view 3D scenes? To answer this, we evaluate its zero-shot performance on SR-3D-Bench covering Tall/Short, Big/Small, Height, and Distance categories. We exclude the width-related category because the width

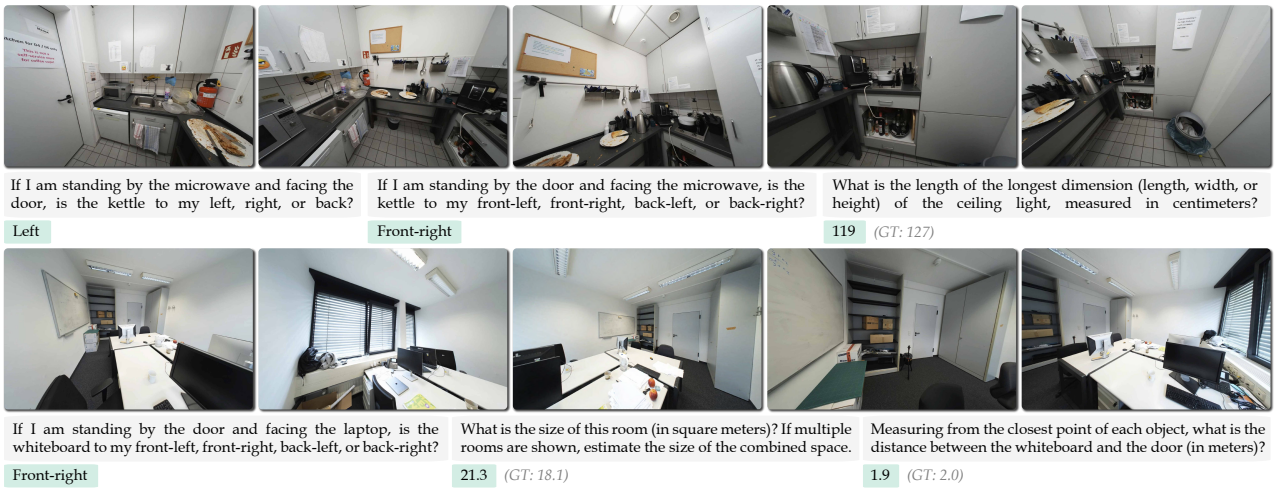

Figure 5 VSI-Bench results. We highlight SR-3D 's outputs and include ground-truth values for numerical answers. The results show that SR-3D answers spatial questions correctly even without region prompts.

	2D Pre-train	3D Tall/Short	3D Big/Small	3D Height	3D Distance
Zero-shot 2D Models					
Base Model		40.0$_{-31.4}$	53.7$_{-26.0}$	54.1$_{-14.4}$	6.6$_{-61.9}$
SR-3D-2D	✓	**71.4**	**79.7**	**68.5**	**68.5**
Finetuned 3D Models					
SR-3D		83.1$_{0.0}$	80.5$_{-1.3}$	85.7$_{-1.6}$	60.3$_{-14.5}$
SR-3D	✓	**83.1**	**81.8**	**87.3**	**74.8**

Table 7 Zero-shot evaluation of our 2D-trained VLM on SR-3D-Bench, testing whether the model's representations are truly aligned. Our 2D model achieves reasonable accuracy without explicit 3D supervision.

is defined differently in single-view and multi-view. In single-view images, width refers to the horizontal extent in the image plane [18], whereas in multi-view settings, it represents an object's maximum length or width. Table 7 shows our results, indicating that the single-view model performs highly competitively. This suggests that our unified representation design effectively transfers knowledge from single-view images, despite the challenge of the model never encountering multi-view data, scene-level position embeddings, or ground-truth spatial annotations.

3D Position Embedding and Single-view Pre-training. We conduct an ablation study to evaluate the impact of single-view pre-training and 3D positional embeddings on our model's performance. We compare two model variants: one fine-tuned directly on multi-view data without positional embeddings and another with them. As shown in Table 8, our results indicate that single-view pre-training significantly enhances performance on multi-view data by enabling the model to leverage prior spatial knowledge. In contrast, adding 3D positional embeddings without scaling provides only a marginal improvement. This highlights the necessity of scaling up to fully harness the power of positional representations for spatial reasoning.

4.5 Applications

Our method is flexible in two key ways. First, because SR-3D is trained in a normalized 3D space, it naturally connects with existing 3D foundation models [23, 24, 106, 107] for pointmap estimation. The input is not restricted to 3D scans—SR-3D can also operate on in-the-wild videos such as YouTube footage. To quantitatively validate this, we evaluate SR-3D on both ground-truth point clouds and

3D PE	2D Pre-train	Scan2Cap	ScanQA	SQA3D	3D Region	3D Global
		92.9	101.3	58.6	74.0	51.1
	✔	94.3	108.2	59.5	78.1	52.9
✔		92.7	102.9	59.1	75.3	51.2
✔	✔	**97.9**	**109.3**	**62.2**	**80.9**	**62.0**

Table 8 Ablation study on the impact of incorporating 3D positional embeddings (3D PE) and single-view pre-training. We report CIDEr for Scan2Cap and ScanQA, EM for SQA3D, and the average score for both the 3D region and global benchmarks.

	3D Source	C ↑	B-1 ↑	B-4 ↑	M ↑	R ↑	EM ↑
Video3dLLM [16]	GT	102.1	47.1	16.2	19.8	49.0	30.1
Video3dLLM [16]	Cut3R	100.7	46.6	15.8	19.6	48.6	29.9
SR-3D	GT	109.3	50.9	18.1	21.2	51.2	30.4
SR-3D	**Cut3R**	**109.3**	**50.9**	**18.1**	**21.2**	**51.2**	**30.2**

Table 9 ScanQA results on both ground-truth point clouds and Cut3R-reconstructed point clouds.

Cut3R-reconstructed [24] point clouds, comparing it with the baseline Video3dLLM [16] on ScanQA. As shown in Table 9, SR-3D maintains strong performance with Cut3R outputs, close to its ground-truth results, whereas the baseline exhibits a significant drop.

Second, SR-3D eliminates the need for costly 3D annotations or dense per-frame labeling. Instead, users can provide lightweight region inputs by simply drawing on a single frame, which the model then propagates for spatial reasoning across the video.

Combining these two aspects, SR-3D demonstrates robust spatial understanding from unconstrained video inputs without reliance on 3D scans or exhaustive annotations (Figure 1). These flexibilities open the door to a wide range of real-world applications, such as assisting robots in unstructured environments, analyzing large video collections, and supporting interactive spatial reasoning tasks.

5 Conclusion

We introduce SR-3D, a foundational VLM for 3D-aware spatial reasoning. By unifying single-view and multi-view data in a shared space, our approach leverages 2D priors from pretrained VLMs to tackle complex 3D tasks. Our tile-and-stitch method extracts high-resolution region features, enabling flexible region prompts across both settings. Experiments on 2D vision-language and 3D spatial benchmarks show state-of-the-art performance, validating SR-3D's ability to unify and enhance spatial reasoning.

References

[1] OpenAI. Gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/. 2, 7, 9, 10

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 7

[3] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023. 7

[4] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv:2502.13923*, 2025.

[6] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. In *CVPR*, 2025. 2, 5, 6, 8, 9, 10

[7] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023. 2

[8] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *ECCV*, 2024. 2

[9] Tengchao Lv, Yupan Huang, Jingye Chen, Lei Cui, Shuming Ma, Yaoyao Chang, Shaohan Huang, Wenhui Wang, Li Dong, Weiyao Luo, et al. Kosmos-2.5: A multimodal literate model. *arXiv:2309.11419*, 2023. 2

[10] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *NeurIPS*, 2023. 2, 3, 8

[11] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. In *ICML*, 2024. 2, 3, 8

[12] 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. In *NeurIPS*, 2024. 8, 2

[13] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *ECCV*, 2024. 2

[14] 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 *CVPR*, 2024. 2, 3, 8

[15] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv:2409.18125*, 2024. 2, 3, 6, 8

[16] Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *CVPR*, 2025. 2, 3, 8, 12

[17] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In *CVPR*, 2024. 2, 3, 5, 6, 7

[18] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. In *NeurIPS*, 2024. 5, 6, 7, 11

[19] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In *CVPR*, 2024. 3

[20] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, 2024. 2, 3

[21] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024. 2, 4

[22] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2

[23] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 6, 11

[24] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2, 6, 11, 12

[25] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv:2404.16821*, 2024. 2

[26] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. In *ICLR*, 2024. 3

[27] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023. 7

[28] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv:2310.09478*, 2023. 7

[29] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. In *NeurIPS*, 2024. 3, 7

[30] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv:2310.11441*, 2023. 3

[31] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *ICLR*, 2024. 3

[32] Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. In *ECCV*, 2024.

[33] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *NeurIPS*, 2023.

[34] Qiang Zhou, Chaohui Yu, Shaofeng Zhang, Sitong Wu, Zhibing Wang, and Fan Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. *arXiv:2308.02299*, 2023.

[35] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv:2307.03601*, 2023. 7

[36] Yuzhong Zhao, Feng Liu, Yue Liu, Mingxiang Liao, Chen Gong, Qixiang Ye, and Fang Wan. Dynrefer: Delving into region-level multi-modality tasks via dynamic resolution. In *CVPR*, 2025. 3, 7

[37] Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *ECCV*, 2024. 3, 10

[38] En Yu, Liang Zhao, Yana Wei, Jinrong Yang, Dongming Wu, Lingyu Kong, Haoran Wei, Tiancai Wang, Zheng Ge, Xiangyu Zhang, et al. Merlin: Empowering multimodal llms with foresight minds. In *ECCV*, 2024.

[39] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *ICML*, 2024.

[40] Hanrong Ye, De-An Huang, Yao Lu, Zhiding Yu, Wei Ping, Andrew Tao, Jan Kautz, Song Han, Dan Xu, Pavlo Molchanov, et al. X-vila: Cross-modality alignment for large language model. *arXiv preprint arXiv:2405.19335*, 2024.

[41] Miran Heo, Min-Hung Chen, De-An Huang, Sifei Liu, Subhashree Radhakrishnan, Seon Joo Kim, Yu-Chiang Frank Wang, and Ryo Hachiuma. Omni-rgpt: Unifying image and video region-level understanding via token marks. In *CVPR*, 2025. 3

[42] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *CVPR*, 2024. 3

[43] Chenyang Ma, Kai Lu, Ta-Ying Cheng, Niki Trigoni, and Andrew Markham. Spatialpin: Enhancing spatial reasoning capabilities of vision-language models through prompting and interacting 3d priors. In *NeurIPS*, 2024.

[44] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *ICRA*, 2025.

[45] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. In *CoRL*, 2024.

[46] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv:2412.07825*, 2024.

[47] Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning. *arXiv:2410.16162*, 2024.

[48] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. *arXiv:2411.16537*, 2024.

[49] Mingjie Xu, Mengyang Wu, Yuzhi Zhao, Jason Chun Lok Li, and Weifeng Ou. Llava-spacesgg: Visual instruct tuning for open-vocabulary scene graph generation with enhanced spatial relations. In *WACV*, 2025.

[50] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. *arXiv:2502.06787*, 2025.

[51] Yuecheng Liu, Dafeng Chi, Shiguang Wu, Zhanguang Zhang, Yaochen Hu, Lingfeng Zhang, Yingxue Zhang, Shuang Wu, Tongtong Cao, Guowei Huang, et al. Spatialcot: Advancing spatial reasoning through coordinate alignment and chain-of-thought for embodied task planning. *arXiv:2501.10074*, 2025.

[52] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *CVPR*, 2025.

[53] Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *EMNLP*, 2024. 3

[54] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *CVPR*, 2025. 3, 10, 4

[55] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *CVPR*, 2024. 3

[56] Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Shawn Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. In *NeurIPS*, 2024.

[57] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *ACM MM*, 2024.

[58] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint*, 2024. 3, 8, 2

[59] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv:2308.08769*, 2023. 3, 8, 2

[60] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022. 6, 8, 2

[61] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *ICLR*, 2023. 6, 8, 2

[62] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, 2021. 6, 8, 2

[63] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 6, 9

[64] 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 *ECCV*, 2014. 7

[65] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *CVPR*, 2022. 7

[66] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv:2308.12966*, 2023. 7

[67] Anthropic. Claude-3-family, 2024. URL https://www.anthropic.com/news/claude-3-family. 7

[68] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 7

[69] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01. ai. *arXiv:2403.04652*, 2024. 7

[70] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023. 7

[71] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, 2024. 7, 10

[72] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024. 7

[73] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. In *arXiv:2403.17297*, 2024. 7

[74] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 7

[75] Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv:2308.13437*, 2023. 7

[76] Weiyun Wang, Min Shi, Qingyun Li, Wenhai Wang, Zhenhang Huang, Linjie Xing, Zhe Chen, Hao Li, Xizhou Zhu, Zhiguo Cao, et al. The all-seeing project: Towards panoptic visual recognition and understanding of the open world. In *ICLR*, 2024. 7

[77] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *ECCV*, 2024. 7, 8

[78] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 8

[79] 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 *ICCV*, 2023. 8, 2

[80] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv:2409.12961*, 2024. 8, 2

[81] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *CVPR*, 2024. 8, 2

[82] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024. 8, 2

[83] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *RSS*, 2025. 8

[84] xAI. Grok-1.5, 2024. 8

[85] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. In *COLM*, 2025.

[86] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024. 8

[87] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *ICLR*, 2024. 8

[88] Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*, 2019. 8

[89] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In *ECCV*, 2016.

[90] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv:2409.02813*, 2024.

[91] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. In *CVPR*, 2024.

[92] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023. 8

[93] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 8

[94] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *ACL*, 2022.

[95] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A Dataset for VQA on Document Images. In *WACV*, 2021. 8

[96] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 9

[97] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *NeurIPS*, 2021. 9

[98] 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:1709.06158*, 2017. 9

[99] Jihao Qiu, Yuan Zhang, Xi Tang, Lingxi Xie, Tianren Ma, Pengyu Yan, David Doermann, Qixiang Ye, and Yunjie Tian. Artemis: Towards referential understanding in complex videos. In *NeurIPS*, 2024. 10

[100] Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*, 2024. 10

[101] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 10

[102] Fuzhao Xue, Yukang Chen, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. In *ICLR*, 2025. 10

[103] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 10

[104] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv:2406.16852*, 2024. 10

[105] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-onevision: Easy visual task transfer. *TMLR*, 2025. 10

[106] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 11

[107] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025. 11

[108] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *CVPR*, 2022. 2

[109] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *European Conference on Computer Vision*, 2022. 2

[110] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv:2405.10370*, 2024. 2

Appendix: Table of Contents

A More Quantitative Results on 3D General Benchmarks

Following prior work, we report results using additional metrics for a more comprehensive evaluation. Table 10 presents results on Scan2Cap, Table 11 on ScanQA, and Table 12 on SQA3D. Apart from our method, all other results are from Video-3D-LLM [16].

	Cider ↑	Bleu-4 ↑	Meteor ↑	Rouge ↑
Scan2Cap [62]	39.1	23.3	22.0	44.5
3DJCG [108]	49.5	31.0	24.2	50.8
D3Net [109]	62.6	35.7	25.7	53.9
3D-VisTA [79]	66.9	34.0	27.1	54.3
LL3DA [14]	65.2	36.8	26.0	55.1
LEO [11]	68.4	36.9	27.7	57.8
ChatScene [12]	77.2	36.3	28.0	58.1
LLaVA-3D [15]	79.2	41.1	30.2	63.4
Video-3D LLM [16]	83.8	42.4	28.9	62.3
SR-3D	**97.9**	**44.7**	**31.5**	**67.3**

Table 10 Full results on Scan2Cap [62] validation set.

	EM	Bleu-1 ↑	Bleu-2 ↑	Bleu-3 ↑	Bleu-4 ↑	Rouge ↑	Meteor ↑	Cider ↑
ScanQA [60]	21.1	30.2	20.4	15.1	10.1	33.3	13.1	64.9
3D-VisTA [79]	22.4	–	–	–	10.4	35.7	13.9	69.6
Oryx-34B [80]	–	38.0	24.6	–	–	37.3	15.0	72.3
LLaVA-Video-7B [82]	–	39.7	26.6	9.3	3.2	44.6	17.7	88.7
3D-LLM (Flamingo) [10]	20.4	30.3	17.8	12.0	7.2	32.3	12.2	59.2
3D-LLM (BLIP2-flant5) [10]	20.5	39.3	25.2	18.4	12.0	35.7	14.5	69.4
Chat-3D [59]	–	29.1	–	–	6.4	28.5	11.9	53.2
NaviLLM [81]	23.0	–	–	–	12.5	38.4	15.4	75.9
LL3DA [14]	–	–	–	–	13.5	37.3	15.9	76.8
Scene-LLM [58]	27.2	43.6	26.8	19.1	12.0	40.0	16.6	80.0
LEO [11]	–	–	–	–	11.5	39.3	16.2	80.0
Grounded 3D-LLM [110]	–	–	–	–	13.4	–	–	72.7
ChatScene [12]	21.6	43.2	29.1	20.6	14.3	41.6	18.0	87.7
LLaVA-3D [15]	27.0	–	–	–	14.5	50.1	20.7	91.7
Video-3D LLM [82]	30.1	47.1	31.7	22.8	16.2	49.0	19.8	102.1
SR-3D	**30.4**	**50.9**	**34.3**	**25.1**	**18.1**	**51.2**	**21.1**	**109.3**

Table 11 Full results on ScanQA [60] validation set.

	What	Is	How	Can	Which	Others	Avg.
SQA3D [61]	31.6	63.8	46.0	69.5	43.9	45.3	46.6
3D-VisTA [79]	34.8	63.3	45.4	69.8	47.2	48.1	48.5
LLaVA-Video[82]	42.7	56.3	47.5	55.3	50.1	47.2	48.5
Scene-LLM [58]	40.9	69.1	45.0	70.8	47.2	52.3	54.2
LEO [11]	–	–	–	–	–	–	50.0
ChatScene []	45.4	67.0	52.0	69.5	49.9	55.0	54.6
LLaVA-3D [15]	–	–	–	–	–	–	55.6
Video-3D LLM [16]	51.1	72.4	55.5	69.8	51.3	56.0	58.6
SR-3D	**55.0**	**76.4**	**59.8**	**71.6**	**54.7**	**61.1**	**62.2**

Table 12 Full results on SQA3D [61] testing set.

B More Quantitative Results on VSI-Bench

We report additional visual results on VSI-Bench, primarily using scenes from ScanNet^{++}. ScanNet^{++} is not included in EmbodiedScan's annotations, making it a distinct and challenging dataset for evaluation. Compared to ScanNet, ScanNet^{++} offers higher fidelity and greater diversity in indoor environments. Moreover, its 3D annotations are only coarsely aligned to match walls and floors to the axis. Despite these challenges, as shown in Figure 6, our method demonstrates superior capabilities in determining relative direction, highlighting its robustness in real-world tasks.

C More Ablation Study

We present the complete ablation study results on 2D single-view pre-training and 3D positional encoding without pre-training, evaluating their influence on model performance. The detailed results are shown in Table 13 and Table 16, respectively.

Overall, the fully-trained model consistently outperforms baseline models on 3D general QA benchmarks, demonstrating the benefits of leveraging both 2D and 3D spatial information. However, in the 3D spatial-focused dataset, we observe a slight drop in the Wide and Big category, likely due to differences in how width is defined in 2D versus 3D, as discussed in the main paper.

Additionally, we find that removing pre-training leads to a substantial drop in performance for more complex reasoning tasks, particularly in the multi-choice complex category, where the model struggles without prior exposure to large-scale 2D pre-training. These results highlight the importance of both spatial-aware representation learning and strong pre-training strategies in enhancing 3D reasoning capabilities.

		Scan2Cap				ScanQA				SQA3D	
PE	PT	Bleu-4 ↑	Rouge↑	Cider↑	Meteor↑	Bleu-4 ↑	Rouge↑	Cider↑	Meteor↑	EM ↑	EM ↑
		44.2	67.3	92.9	31.1	16.0	48.9	101.3	19.8	28.8	58.6
✔		44.0	67.3	92.7	31.0	17.4	48.8	102.9	20.0	29.1	59.1
✔	✔	**44.7**	**67.3**	**97.9**	**31.5**	**18.1**	**51.2**	**109.3**	**21.2**	**30.4**	**62.2**

Table 13 Ablation study full results on Scan2Cap, ScanQA, and SQA3D benchmarks.

Category	Thin-Wide	Tall-Short	Big-Small	Multi-Simple	Multi-Complex	Width Data	Distance Data	Height Data	Total Length
Count	219	231	231	117	500	496	242	464	2500

Table 14 Statistical analysis of our SR-3D-Bench, showing the distribution of different spatial attributes.

D Statistics of SR-3D-Bench

Our benchmark follows template designs from prior works on spatial reasoning in vision-language models, including SpatialRGPT and SpatialVLM. To further increase the complexity and diversity of spatial reasoning tasks, we incorporate situated annotations from the EmbodiedScan dataset, ensuring a more realistic and challenging evaluation setting. Specifically, our dataset includes a range of spatial relationships, from basic geometric comparisons such as thin-wide, tall-short, and big-small, to more complex multi-object interactions categorized as multi-simple and multi-complex. Additionally, we introduce explicit width, distance, and height annotations to facilitate fine-grained spatial understanding. With a total of 2,500 samples, our benchmark provides a comprehensive evaluation for assessing the region-level spatial reasoning capabilities of vision-language models in realistic scenarios.

E Implementation Details of SR-3D

We use PaliGemma as our visual backbone with an input size of 448 and a patch size of 14, paired with a Qwen-2-7B LLM backbone. For training the foundational 2D VLM, we follow prior work and set the maximum tiles per image to 12. For the multi-view VLM, we use a frame size of 32 with a uniform sampling strategy to ensure a fair comparison with previous methods. For training the 2D VLM, we adopt a learning rate of 5e-5 with cosine decay and gradient clipping enabled. The same hyperparameters are used for fine-tuning the 3D VLM, except for a reduced batch size due to the

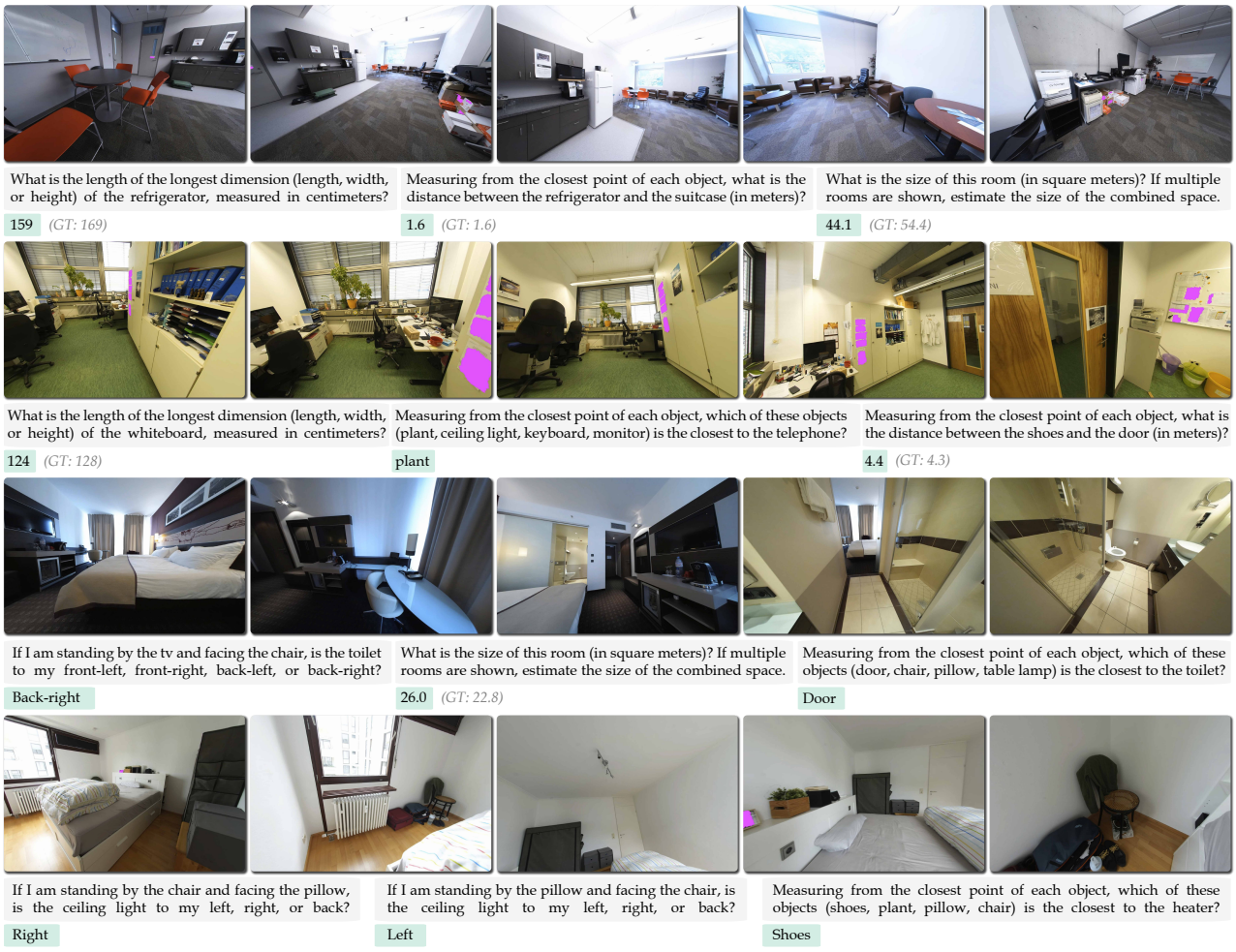

Figure 6 More results on VSI-Bench [54]. We highlight SR-3D 's outputs and include ground-truth values for numerical answers.

increased token length. The data recipe for both training stages is detailed in Table 15. We train on a subset of 2D data, excluding spatial and region-related datasets, to preserve the original vision-language capabilities while incorporating a diverse source.

F Limitations

Orientations Although our method shows promising results, it remains challenging for current vision-language models to accurately perceive and interpret spatial questions related to object orientation. This challenge arises due to the difficulty of scaling up data. We leave this as future work.

Dynamic Videos Our method is designed for multi-view static data, whereas real-world scenarios often involve dynamic environments. Incorporating positional embeddings to handle both static and dynamic inputs is non-trivial. Future work should explore methods to address this limitation.

OCR Tasks In the main paper, Table 1, we report the performance of our 2D foundation model on general benchmarks. While our model maintains comparable performance to the base model, demonstrating improved spatial understanding without significant trade-offs, we observe a consistent slight drop in

2D Data	
Hybrid	ShareGPT4V-SFT, Molmo, The Cauldron, Cambrian, LLaVA-OneVision
Captioning	MSR-VTT, Image Paragraph Captioning, ShareGPT4V-100K
Reasoning	CLEVR, NLVR, VisualMRC
Document	DocVQA, UniChart-SFT, ChartQA
OCR	TextCaps, OCRVQA, ST-VQA, POIE, SORIE, SynthDoG-en, TextOCR-GPT4V, ArxivQA, LLaVAR
General VQA	ScienceQA, VQAv2, ViQuAE, Visual Dialog, GQA, Geo170K, LRV-Instruction, RefCOCO, GeoQA, OK-VQA, TabMVP, EstVQA
Diagram & Dialogue	DVQA, AI2D, Shikra, UniMM-Chat
Instruction	LRV-Instruction, SVIT, MMC-Instruction, MM-Instruction
Text-only	FLAN-1M, MathInstruct, Dolly, GSM8K-ScRel-SFT
Knowledge	WordART, WIT, STEM-QA
Medical	PathVQA, Slake, MedVQA
Region	RegionGPT
Spatial	SpatialRGPT
3D Data	
General	ScanQA, SQA3D, Scan2Cap
Spatial	EmbodiedScan

Table 15 Data recipe for training 2D foundational VLM and 3D fine-tuning.

		3D Region						3D Global			
PE	PT	Wide	Tall	Big	M. Sim.	M. Cpx.	Avg.	Width	Height	Dist.	Avg.
		77.6	80.5	**82.6**	71.7	55.8	73.6	85.8	84.4	53.7	74.4
✔		**77.6**	**83.1**	80.5	70.9	59.0	74.2	85.5	85.7	60.3	77.2
✔	✔	76.3	**83.1**	81.8	**80.3**	**76.0**	**79.5**	**87.7**	**87.3**	**74.8**	**83.3**

Table 16 Ablation study full results.

OCR-related tasks. A potential solution is to incorporate more OCR-related tasks into the training data pipeline.

Unified Checkpoint While our unified architecture and representation provide a foundation for both single- and multi-view 3D-aware VLMs, we leave it to future work to investigate how to effectively combine the two models. This could be achieved either by introducing an agentic flow between single- and multi-view models or by directly training a single model across both settings, which may further improve generalization and efficiency.