3D AWARE REGION PROMPTED VISION LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

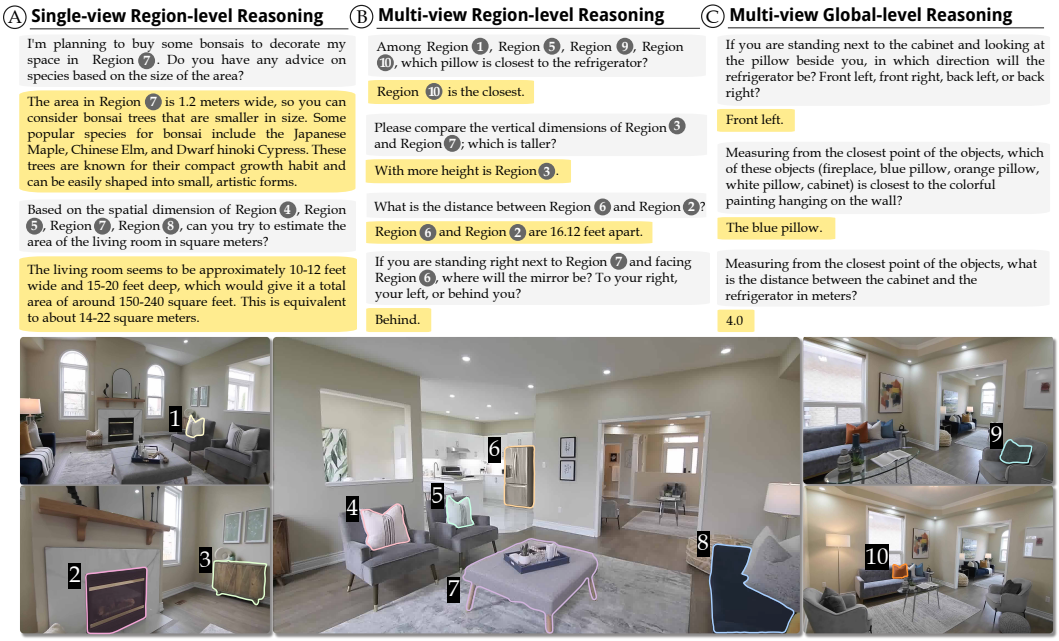

Figure 1: From precise region-based distance estimation (*left*), to intricate multi-view region query (*middle*), and global cross-frame reasoning (*right*), SR-3D delivers flexible and accurate spatial understanding to foundational Vision-Language Models. Notably, this video is obtained in the wild,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

https://sr3d-iclr.github.io/

ABSTRACT

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.

1 INTRODUCTION

The rapid advancement of Vision Language Models (VLMs) (OpenAI, 2024; Liu et al., 2023; Anil et al., 2023; Wang et al., 2024b; Bai et al., 2025; Liu et al., 2025b) has demonstrated strong capabilities in visual understanding (Pratt et al., 2023; Huang et al., 2024a) and language grounding (Lv et al., 2023). 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 (Hong et al., 2023; Huang

et al., 2024c;b; Xu et al., 2024; Chen et al., 2024b) 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 (Zhu et al., 2024; Zheng et al., 2025). Unlike point clouds (Huang et al., 2024c;b; Xu et al., 2024) 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 (Guo et al., 2024; Cheng et al., 2024; Yuan et al., 2024b; Rasheed et al., 2024). 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 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 (Zheng et al., 2025), or in different pathways (Zhu et al., 2024), we directly integrate positional embeddings within the foundational VLM. Specifically, we estimate each input image's depth using an off-the-shelf depth estimator (Yang et al., 2024) 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 (Wang et al., 2024c; Leroy et al., 2024; Wang et al., 2025c) 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 (Chen et al., 2024d; Liu et al., 2025b), we design a novel branch specifically compatible with this architecture to produce robust region embeddings.

SR-3D naturally supports flexible region annotation on any frame. This stems from two design choices: (1) consistent 3D positional embeddings in a canonical space, enabling coherent cross-frame correspondences, and (2) an aligned embedding space from single-view pretraining that unleashes the full potential of region embeddings to generalize in multi-frame contexts. As evidence, our 2D-VLM trained only on single-view data demonstrates strong zero-shot spatial reasoning in 3D scenes, with and without region prompts, despite never seeing multi-view data.

We extensively evaluate across single-view and 3D multi-view settings, covering both region-level and global QA on general and spatial tasks. Our foundational 2D-VLM delivers large gains on region-level performance, surpassing prior state-of-the-art in both recognition and spatial understanding. These gains come without sacrificing general VQA accuracy and even improve tasks requiring spatial knowledge. With 3D fine-tuning, our model sets new state-of-the-art results in general 3D QA, video spatial reasoning, and region-level video 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.

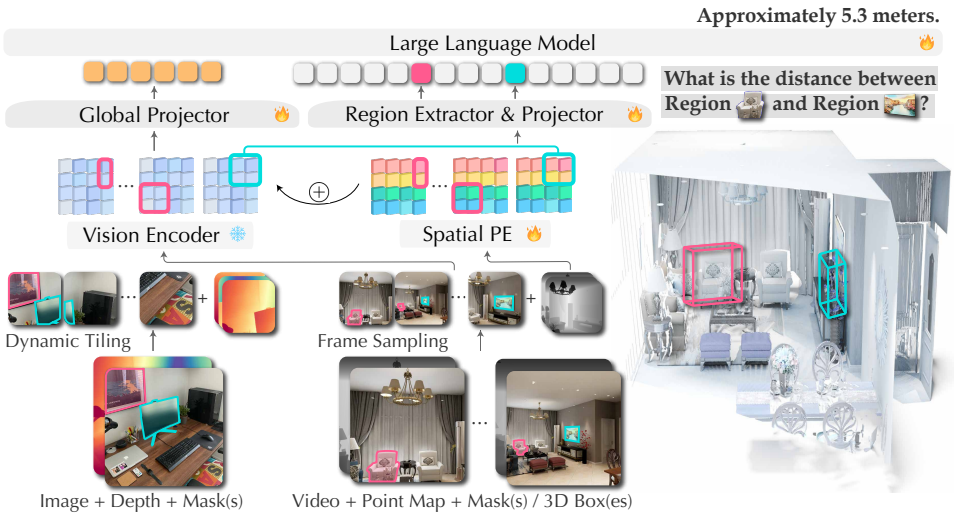

Figure 2: The SR-3D architecture. Given an image or multi-view input with optional region prompts (e.g., bounding boxes or masks), we encode them along with depth-derived positional embeddings using a tiling approach. Region tokens are extracted by stitching masked features, while 3D positional embeddings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

2 METHODOLOGY

We propose a 3D-aware VLM architecture for single- and multi-view spatial understanding. Our approach adapts strong 2D priors by directly integrating 3D positional embeddings into visual representations, enabling accurate cross-frame reasoning. To further improve region-level grounding, we introduce a Dynamic Tiling-based Region Extractor, which works efficiently across both single- and multi-view inputs. As shown in Figure 2, the framework comprises a vision encoder, a 3D position encoding module, a region extractor, and an LLM backbone. In this section, we describe three main components: (1) canonical 3D positional representation (Sec. 2.1), (2) the region extractor (Sec. 2.2), and (3) the training paradigm (Sec. 2.3), along with inference pipeline (Sec. 2.4).

2.1 CANONICAL 3D POSITIONAL REPRESENTATION

The key idea of SR-3D is a canonical positional feature shared across single- and multi-view inputs. This unified representation unleashes the full potential of large-scale single-view pretraining, carrying its spatial priors into 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 (Yang et al., 2024). 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 the point maps to match the vision encoder's 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.

Methods	Spatial				Math
	BLINK$_S$	SAT	EmbSpat	RealWorldQA	MathVista
NVILA-Lite-8B	79.7	62.6	68.9	65.6	64.5
SR-3D-8B	**83.9**$_{+4.2}$	**64.0**$_{+1.4}$	**72.5**$_{+3.6}$	**68.1**$_{+2.5}$	**65.4**

Methods	General Knowledge					OCR-Related		
	GQA	AI2D	MMMU$_p$	SEED$_I$	POPE	TextVQA	ChartQA	DocVQA
NVILA-Lite-8B	**65.3**	**91.0**	**25.1**	76.3	**88.1**	**78.1**	**84.8**	**91.7**
SR-3D-8B	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 NVILA-Lite (Liu et al., 2025b)'s performance on general VQA benchmarks. SR-3D achieves stronger results on spatial-related benchmarks without compromising performance on general and OCR benchmarks, suggesting that 3D-aware pre-training enhances spatial reasoning while preserving the base model's broader knowledge.

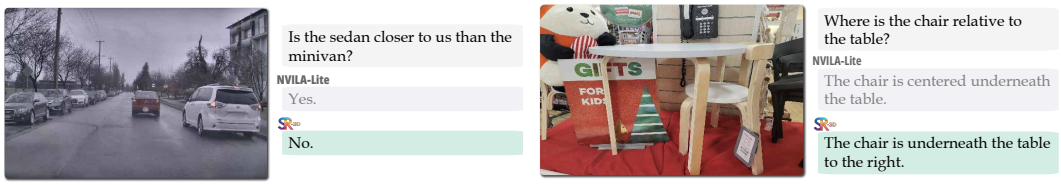

Figure 3: RealWorldQA (xAI, 2024) results. SR-3D shows stronger spatial understanding of physical environments compared to the base model. We omit the answer choices for clarity in visualization.

2.2 Dynamic Tiling-based Region Extractor

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 (Liu et al., 2025b) 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 similarly to point maps and region masks, forming the basis for both our 3D positional embedding and region feature extraction strategies.

Dynamic Region Extractor. Prior architectures without dynamic tiling use feature refinement with deconvolution layers to upsample visual tokens (Cheng et al., 2024; Guo et al., 2024), aiming to recover lost details. But since this occurs after the vision encoder, where features are already resized and potentially distorted, the recovery of fine details is limited.

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.

2.3 Training Paradigm

For the single-view VLM, we initialize the weights from a pre-trained 2D VLM (NVILA-Lite-8B (Liu et al., 2025b)), 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 (Guo et al., 2024; Cheng et al., 2024) 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 (Azuma et al., 2022), SQA3D (Ma et al., 2023), and Scan2Cap (Chen et al., 2021), as

Methods	Acc. (%)
Human	98.3
Proprietary Models (API)	
Gemini Pro (Anil et al., 2023)	50.0
Claude 3 OPUS (Anthropic, 2024)	57.3
GPT-4V-*Turbo* (OpenAI, 2024)	66.9
GPT-4o (OpenAI, 2024)	64.5
Open-source Models	
Yi-VL-34B (Young et al., 2024)	53.2
LLaVA-v1.5-13B-xtuner (Contributors, 2023)	54.0
LLaVA-v1.6-34B (Zhang et al., 2024b)	64.5
InstructBLIP-7B (Dai et al., 2023)	50.8
LLaVA-v1.5-7B-xtuner (Contributors, 2023)	50.8
LLaVA-v1.5-7B (Liu et al., 2024a)	51.6
LLaVA-InternLM2-7B (Cai et al., 2024)	52.4
SpatialRGPT-8B (Cheng et al., 2024)	87.9
SR-3D-8B	**90.3**

Table 2: Results on $\text{BLINK}_{\texttt{Depth}}$. We follow Cheng et al. (2024)'s protocol to test whether a VLM effectively leverages auxiliary 3D information.

Methods	mAP (↑)	Acc. (%)
CLIP (Radford et al., 2021)	58.9	-
RegionCLIP (Zhong et al., 2022)	58.3	-
LLaVA-7B (Liu et al., 2023)	-	40.0
Shikra-7B (Chen et al., 2023c)	-	53.9
GPT4RoI-7B (Zhang et al., 2023)	-	64.0
PVIT-7B (Chen et al., 2023a)	-	64.5
ASM-7B (Wang et al., 2024g)	69.3	-
RegionGPT-7B (Guo et al., 2024)	70.0	80.6
DynRefer (Zhao et al., 2025)	-	81.2
SpatialRGPT-8B (Cheng et al., 2024)	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 (Zhong et al., 2022) and RegionGPT (Guo et al., 2024).

well as a newly curated EmbodiedScan (Wang et al., 2024d) 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 (Zhu et al., 2024) 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- and multi-view inputs without distinction between spatial region prompts and global queries.

2.4 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. This reflects SR-3D 's ability to handle varying annotation densities while preserving spatial alignment. For 3D multi-view 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 (Leroy et al., 2024) or CUT3R (Wang et al., 2025c). 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.

3 Experiments

We first evaluate SR-3D on 2D benchmarks (Section 3.1) 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 3.2. Finally, we show ablation studies in Section 3.4 to analyze the role of pretraining and 3D positional encoding.

3.1 Evaluation on 2D Benchmarks

Region-level Question Answering. We evaluate our model's object classification performance on the COCO-2017 (Lin et al., 2014) dataset using mean Average Precision (mAP) and classification accuracy as metrics. Following prior work on region-level recognition (Zhong et al., 2022; Guo et al., 2024; Cheng et al., 2024), 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 (Cheng et al., 2024), 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) (Zhao et al., 2025) as a baseline at the same resolution. Their proposed strategies are also complementary to our approach.

Methods	Scan2Cap				ScanQA					SQA3D
	B-4 ↑	Rouge ↑	Cider ↑	Meteor ↑	B-4 ↑	Rouge ↑	Cider ↑	Meteor ↑	EM ↑	EM ↑
Task-specific Specialist										
VoteNet+MCAN (Yu et al., 2019)	-	-	-	-	6.2	29.8	54.7	11.4	17.3	-
ScanRefer+MCAN (Yu et al., 2019)	-	-	-	-	7.9	30.0	55.4	11.5	18.6	-
ScanQA (Azuma et al., 2022)	-	-	-	-	10.1	33.3	64.9	13.1	21.0	-
3D-VisTA (Zhu et al., 2023)	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 (Liu et al., 2024b)	-	-	-	-	-	37.3	72.3	15.0	-	-
NaviLLM (Zheng et al., 2024)	-	-	-	-	12.0	38.4	75.9	15.4	23.0	-
LLaVA-Video-7B† (Zhang et al., 2024c)	-	-	-	-	3.1	44.6	88.7	17.7	-	-
NaVILA (Cheng et al., 2025)	-	-	-	-	16.9	49.3	102.7	20.1	28.6	-
3D Large Multi-modal Models										
3D-LLM$_{(flamingo)}$ (Hong et al., 2023)	-	-	-	-	7.2	32.3	59.2	12.2	20.4	-
3D-LLM$_{(BLIP2-flant5)}$ (Hong et al., 2023)	-	-	-	-	12.0	35.7	69.4	14.5	20.5	-
LL3DA (Chen et al., 2024b)	36.8	55.1	65.2	26.0	13.5	37.3	76.8	15.9	-	-
Chat-3Dv2 (Wang et al., 2023b)	-	-	-	-	14.0	-	87.6	-	-	54.7
LEO (Huang et al., 2024c)	36.9	57.8	68.4	27.7	13.2	49.2	101.4	20.0	24.5	50.0
Scene-LLM (Fu et al., 2024a)	-	-	-	-	12.0	40.0	80.0	16.6	27.2	54.2
ChatScene (Huang et al., 2024b)	36.3	58.1	77.2	28.0	14.3	41.6	87.7	18.0	21.6	54.6
LLaVA-3D (Zhu et al., 2024)	41.1	63.4	79.2	30.2	14.5	50.1	91.7	20.7	27.0	55.6
Video-3D LLM (Zheng et al., 2025)	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 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.

We further evaluate SR-3D on the BLINK$_{\texttt{Depth}}$ benchmark (Fu et al., 2024b) using the region-prompts as in SpatialRGPT (Cheng et al., 2024), which tests point-level depth understanding in VLMs. BLINK$_{\texttt{Depth}}$ 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 (Cheng et al., 2024), 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 on spatial-related tasks? To answer these, we evaluate our model on general VLM benchmarks covering Spatial (xAI, 2024; Ray et al., 2025; Du et al., 2024; Fu et al., 2024b), Math (Lu et al., 2024), General Understanding (Hudson & Manning, 2019; Kembhavi et al., 2016; Yue et al., 2024; Li et al., 2024; 2023), and OCR-related (Singh et al., 2019; Masry et al., 2022; Mathew et al., 2021) tasks. As shown in Table 1, compared to the base model NVILA-Lite-8B (Liu et al., 2025b), 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 (xAI, 2024). We also provide qualitative examples from RealWorldQA in Figure 3, showcasing cases where NVILA-Lite fails while SR-3D succeeds. These results demonstrate that SR-3D enhances spatial reasoning while preserving general capabilities.

3.2 Evaluation on 3D Benchmarks

General 3D Question Answering. We report results on three classic 3D vision-language understanding tasks: 3D dense captioning on Scan2Cap (Chen et al., 2021), ScanQA (Azuma et al., 2022), and SQA3D (Ma et al., 2023) in Table 4. 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 SR-3D against strong baselines, including task-specific specialist models for each benchmark and leading methods from both 2D and 3D large multimodal models (LMMs). SR-3D significantly outperforms state-of-the-art single-task and task-specific fine-tuned models on 3D dense captioning and 3D QA tasks. Our design operates in a canonicalized 3D space, making it naturally compatible with geometric foundation models and well-suited for extending to casual in-the-wild videos. Motivated by this, we further evaluate robustness under reconstructed inputs using Cut3R (Wang et al., 2025c), comparing results against ground-truth point clouds. As shown in Table 9, SR-3D maintains performance close to ground truth, highlighting its resilience to reconstruction artifacts, while Video3dLLM (Zheng et al., 2025) exhibits a clear performance drop.

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 (OpenAI, 2024)	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 (OpenAI, 2024)	52.1	54.1	57.5	62.4	42.4	53.7	72.4	72.8	55.8	67.0
NVILA-Video-8B (Liu et al., 2025b)	48.8	38.9	53.7	52.1	36.0	45.9	59.2	54.3	6.6	40.0
Region VLMs										
GPT-4o (OpenAI, 2024)+SoM	46.1	39.9	39.3	52.1	43.2	44.1	52.4	47.8	40.0	46.7
NVILA-Video-8B (Liu et al., 2025b)+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. SR-3D outperforms all baselines, highlighting the importance of strong region understanding and spatial awareness. Notably, SoM struggles with multi-frame inputs, reflecting the inherent difficulty of multi-frame visual grounding.

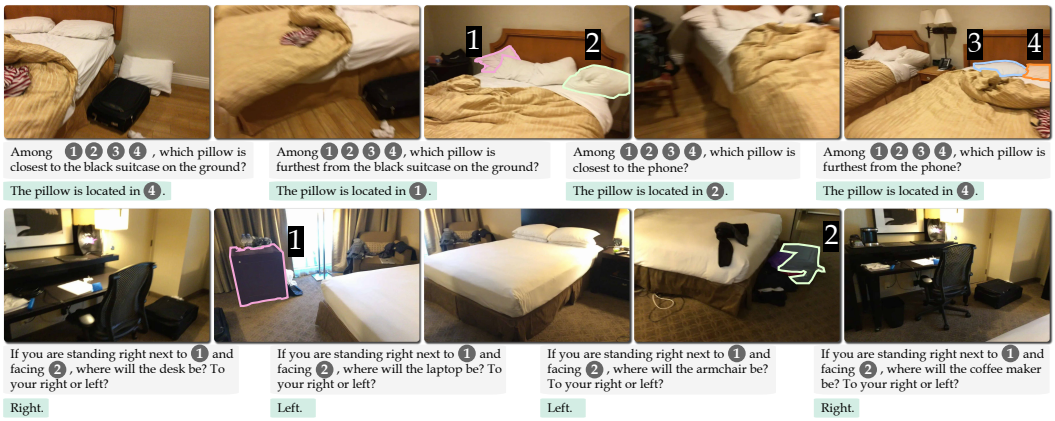

Figure 4: SR-3D results on region-level multi-view spatial understanding. We show extreme cases where the same region prompts are used across samples but with different target objects. SR-3D answers all queries correctly, showing strong evidence that it truly understands 3D spatial relationships. Note that the overlayed region ID tags are only for visualization in the paper to improve readability and are not used during inference.

3.3 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, a region-level spatial benchmark curated from ScanNet (Dai et al., 2017), ARKitScenes (Baruch et al., 2021), and Matterport (Chang et al., 2017) video scan datasets with 3D ground truth. Specifically, we utilize preprocessed oriented bounding box annotations from EmbodiedScan (Wang et al., 2024d), 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 (OpenAI, 2024) 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 (OpenAI, 2024) as the representative, as it is

Methods	Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	Avg.
	Quantitative				Qualitative				
Random	-	-	-	-	25.0	36.1	28.3	25.0	-
Human Level†	94.3	47.0	60.4	45.9	94.7	95.8	95.8	100	79.2
Proprietary Models (API)									
GPT-4o (OpenAI, 2024)	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0
Gemini-1.5 Flash (Georgiev et al., 2024)	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8	42.1
Gemini-1.5 Pro (Georgiev et al., 2024)	**56.2**	30.9	64.1	43.6	51.3	46.3	**36.0**	34.6	45.3
Open-source Models									
InternVL2-8B (Chen et al., 2024e)	31.3	29.0	48.9	44.2	38.0	33.4	28.9	46.4	37.5
InternVL2-40B (Chen et al., 2024e)	41.3	26.2	48.2	27.5	47.6	32.7	27.8	44.7	37.0
LongVILA-8B (Xue et al., 2025)	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5	21.6
VILA-1.5-8B Lin et al. (2024)	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8	28.9
VILA-1.5-40B Lin et al. (2024)	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9	31.2
LongVA-7B (Zhang et al., 2024a)	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7	29.2
LLaVA-Video-7B (Zhang et al., 2024b)	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6	35.6
LLaVA-Video-72B (Zhang et al., 2024b)	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
LLaVA-OneVision-7B (Li et al., 2025)	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4	32.4
LLaVA-OneVision-72B (Li et al., 2025)	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2
SR-3D-8B	54.9	**53.8**	**74.5**	**65.1**	**63.5**	**81.8**	33.5	**75.9**	**62.9**

Table 6: Results on multi-view global spatial scene understanding evaluated on VSI-Bench (Yang et al., 2025b). † indicates methods tested on the `Tiny` subset. SR-3D achieves strong performance, especially on the relative direction task, providing clear evidence that the model effectively leverages the 3D positional encoding.

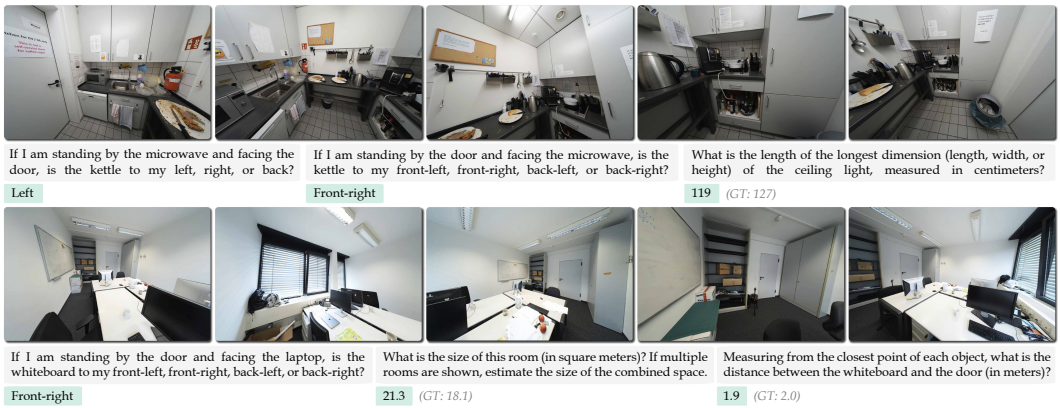

Figure 5: VSI-Bench (Yang et al., 2025b) results. SR-3D answers spatial questions correctly even without region prompts, handles fine-grained directional queries such as distinguishing `front-left` from `front-right`, and accurately answers metric-scale spatial questions like distance queries.

one of the most 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 and NVILA-Video (Liu et al., 2025b) 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 Qiu et al. (2024) and Wang et al. (2024a) 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 (Yang et al., 2025b), a recently proposed benchmark that quantitatively evaluates the visual-spatial intelligence of VLMs based on egocentric videos. To avoid potential effects from noisy or inconsistent labels, training samples from the ScanQA series are excluded. We follow the original setting and use accuracy as the evaluation metric for qualitative questions and Mean Relative Accuracy (MRA)

	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. SR-3D-2D achieves reasonable accuracy without explicit 3D supervision.

3D PE	PT	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 (PT). The results indicate that both 3D positional embeddings and single-view pre-training are crucial, and further scaling up pre-training is likely to yield additional gains.

	3D Source	C ↑	B-1 ↑	B-4 ↑	M ↑	R ↑	EM ↑
Video-3D LLM	GT	102.1	47.1	16.2	19.8	49.0	30.1
Video-3D LLM	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 ground-truth and Cut3R-reconstructed point clouds, compared with Video-3D LLM (Zheng et al., 2025). SR-3D exhibits a smaller performance drop than the baseline when shifting from ground-truth to reconstructed inputs.

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.

3.4 Analysis and Ablation Study

Zero-shot Generalization. In this analysis, we ask: Can a foundational 2D VLM trained only on single-view images perform zero-shot spatial reasoning on multi-view 3D scenes? To test this, we evaluate its zero shot performance on SR-3D-Bench across the Tall/Short, Big/Small, Height, and Distance categories. We exclude width because it is defined differently in single-view and multi-view settings: in single-view images, it refers to the horizontal extent in the image plane (Cheng et al., 2024), while in multi-view scenes, it denotes the maximum object dimension. Table 7 presents the results, showing that the single-view model performs strongly. This indicates that our unified representation transfers knowledge from single-view images effectively, even though the model has not seen 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. Four model variants are compared, with/without pre-training and with/without 3D positional embeddings. As shown in Table 8, single-view pre-training provides substantial gains by allowing the model to transfer spatial knowledge, while 3D embeddings offer limited improvements at the current scale. These findings highlight the need for larger-scale settings to fully exploit positional representations for spatial reasoning.

4 Related Work

Our work builds upon recent advancements in region-level understanding (Yuan et al., 2024b; Guo et al., 2024), spatial reasoning (Chen et al., 2024a; Yang et al., 2025b), and 3D large multimodal models (Hong et al., 2023; Chen et al., 2024b). The most closely related methods are LLaVA-3D (Zhu et al., 2024) and Video-3D LLM (Zheng et al., 2025), which also integrate 3D position-aware features into 2D VLMs. However, these approaches often rely on separate processing pathways for 2D/3D data or require fine-tuning on specialized 3D video data, which risks overfitting position encodings to specific tasks. In contrast, we propose a unified architecture and a shared 3D representation space for both images and videos, fostering better alignment and improving generalization across spatial understanding tasks. A comprehensive literature review is in Appendix B.

5 Conclusion

We introduce SR-3D, a foundational vision language model for 3D-aware spatial reasoning. By unifying single- and multi-view data, our approach adapts strong 2D priors from pretrained VLMs into a 3D-aware representation for complex spatial tasks. Additionally, our tile-and-stitch method extracts high-resolution region features, enabling flexible region prompts. Experiments on 2D and 3D benchmarks show state-of-the-art performance, validating SR-3D's ability to unify and enhance spatial reasoning, unlocking the potential of 3D-aware VLMs.

Ethics Statement

SR-3D is developed as a general-purpose visual assistant, similar to other vision language models (OpenAI, 2024; xAI, 2024; Georgiev et al., 2024). While it offers potential benefits for tasks in robotics, AR/VR, and other domains, it also shares common concerns associated with large language and multimodal models. These include the risk of output hallucinations, inherited biases from pretrained models, and the environmental impact of scaling to larger architectures. Evaluating spatial reasoning performance remains challenging (Cheng et al., 2024), and further research is needed to ensure robustness and reliability, particularly in safety-critical domains such as robotics. In potential applications to VR/AR smart glasses, future work should also address privacy and security concerns. Our work serves as a research prototype, and we do not claim deployment readiness. No human subjects were involved in this study, and no personally identifiable information was collected. The supplementary demonstration video on the website uses publicly available YouTube footage and is provided solely for academic research purposes, not for commercial use.

Reproducibility Statement

We have taken several measures to ensure the reproducibility of our work. SR-3D builds upon an open-sourced vision language model (Liu et al., 2025b) as the base, and all datasets used in our experiments are publicly available, with no in-house or proprietary data involved. In the main paper and appendix, we provide detailed descriptions of the data curation pipeline, model architecture, and training hyperparameters. To further support reproducibility, we will release our curated datasets, benchmark, source code, and pretrained model weights as open-sourced software.

References

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. 1, 5

Anthropic. Claude-3-family, 2024. URL https://www.anthropic.com/news/claude-3-family. 5

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *CVPR*, 2022. 4, 6, 19

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. 21

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. 1

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. 7

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *rXiv:2407.07726*, 2024. 21

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. 19

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. 18

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. 5

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. 7

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*, 2024a. 9, 18, 21

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*, 2023a. 5

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. 19

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*, 2023b. 18

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv:2306.15195*, 2023c. 5, 18

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*, 2024b. 2, 6, 9, 18, 19

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*, 2024c. 19

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*, 2024d. 2

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*, 2024e. 8

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *CVPR*, 2021. 4, 6, 19

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. 2, 4, 5, 6, 9, 10, 21

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. 6

XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023. 5

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. 7

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. 5

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. 6

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. 18

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint*, 2024a. 6, 18, 19

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*, 2024b. 6

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. 8, 10

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, 4, 5, 9, 18

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. 18

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. 1, 6, 9, 18, 19

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *ECCV*, 2024a. 1

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*, 2024b. 2, 6, 19

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*, 2024c. 1, 2, 6, 18, 19

Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv:2507.23478*, 2025. 18

Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *CVPR*, 2019. 6

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In *ECCV*, 2016. 6

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *ECCV*, 2024. 2, 5, 18

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. 8

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking Multimodal LLMs with Generative Comprehension. In *CVPR*, 2024. 6

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. 6

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. 18

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, 2024. 8

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. 5

Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Shawn Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3d scenes. In *NeurIPS*, 2024. 18

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1, 5

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024a. 5

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*, 2025a. 18

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*, 2025b. 1, 2, 4, 6, 7, 8, 10

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*, 2024b. 6, 19

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. 6

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. 1

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*, 2024a. 18

Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. *arXiv:2412.07825*, 2024b. 18

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. 4, 6, 19

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *CVPR*, 2024. 18

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. *arXiv:2502.06787*, 2025. 18

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. 6

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A Dataset for VQA on Document Images. In *WACV*, 2021. 6

OpenAI. Gpt-4o, 2024. URL https://openai.com/index/hello-gpt-4o/. 1, 5, 7, 8, 10

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. 18

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. 1

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. 8

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. 5

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, 18

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. 6

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. 6

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. 18

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. 18

Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *ECCV*, 2024a. 8, 18

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *ICCV*, 2025a. 18

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025b. 18

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*, 2024b. 1

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025c. 2, 5, 6, 18

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

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*, 2024d. 5, 7, 21

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*, 2024e. 18

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*, 2024f. 18

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*, 2024g. 5

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*, 2024h. 18

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*, 2023a. 18

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*, 2025d. 18

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*, 2023b. 6, 18, 19

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv:2505.23747*, 2025. 18

xAI. Grok-1.5, 2024. 4, 6, 10

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. 18

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

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. 8

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. 18

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*, 2025a. 18

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*, 2025b. 8, 9, 18, 20

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *NeurIPS*, 2024. 2, 3

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. 18

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. 5

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. 18

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 6

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*, 2024a. 18

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*, 2024b. 2, 9, 18

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. 6

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*, 2024a. 8

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. 5, 18

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, 2024b. 5, 8

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024c. 6, 19

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. 5, 18

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *CVPR*, 2024. 6, 19

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *CVPR*, 2025. 2, 6, 9, 18, 19

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. 5

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. 18

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, 5, 6, 9, 18, 19

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. 6, 19

APPENDIX: TABLE OF CONTENTS

A Applications

Our method is flexible in two key ways. First, since SR-3D is trained in a normalized 3D space, it naturally integrates with existing 3D foundation models (Wang et al., 2025c; Leroy et al., 2024; Wang et al., 2025d;b) for pointmap estimation. This design allows the input to extend beyond 3D scans, and SR-3D can also process in the wild videos such as YouTube footage. Second, SR-3D removes the need for costly 3D annotations or dense per-frame labeling. Instead, users can provide lightweight region inputs by drawing on a single frame, which the model then propagates across the video for spatial reasoning.

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.

B Comprehensive Literature Review

Region-level Vision-Language Models. Region-level VLMs enhance fine-grained visual understanding by focusing on specific regions in images and videos. Early methods (Peng et al., 2024; Chen et al., 2023c;b; Wang et al., 2024e) 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 (Yang et al., 2023), 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 (Wang et al., 2024h;f; 2023a; Zhou et al., 2023; Zhang et al., 2023; Rasheed et al., 2024; Zhao et al., 2025), with RegionGPT (Guo et al., 2024) and Osprey (Yuan et al., 2024b) 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 (Wang et al., 2024a; Yu et al., 2024; Fei et al., 2024; Ye et al., 2024; Heo et al., 2025) 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 (Chen et al., 2024a; Ma et al., 2024a; Cai et al., 2025; Yuan et al., 2024a; Ma et al., 2024b; Tang et al., 2024; Song et al., 2024; Xu et al., 2025; Marsili et al., 2025; Liu et al., 2025a; Yang et al., 2025a; Liao et al., 2024). While most previous work has focused on spatial understanding from 2D images, multi-view spatial reasoning remains less explored. Recently, VSI-Bench (Yang et al., 2025b) 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 (Wang et al., 2023b; Man et al., 2024; Linghu et al., 2024; Hong et al., 2023; Fu et al., 2024a; Chen et al., 2024b; Huang et al., 2024c; Wang et al., 2025a; Huang et al., 2025; Wu et al., 2025). Various 3D representations have been explored to integrate position information into LLMs. 3D-LLM (Hong et al., 2023) and Scene-LLM (Fu et al., 2024a) use multi-view images with object segmentation masks to construct pixel-aligned point representations, while LL3DA (Chen et al., 2024b) directly employs a point cloud encoder to extract 3D scene features. LEO (Huang et al., 2024c) and Chat3D (Wang et al., 2023b) 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 (Zhu et al., 2024) and Video-3D-LLM (Zheng et al., 2025), which also incorporate 3D position-aware features into 2D vision-language models. However, LLaVA-3D processes 3D and 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.

C 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 (Zheng et al., 2025).

	Cider ↑	Bleu-4 ↑	Meteor ↑	Rouge ↑
Scan2Cap (Chen et al., 2021)	39.1	23.3	22.0	44.5
3DJCG (Cai et al., 2022)	49.5	31.0	24.2	50.8
D3Net (Chen et al., 2022)	62.6	35.7	25.7	53.9
3D-VisTA (Zhu et al., 2023)	66.9	34.0	27.1	54.3
LL3DA (Chen et al., 2024b)	65.2	36.8	26.0	55.1
LEO (Huang et al., 2024c)	68.4	36.9	27.7	57.8
ChatScene (Huang et al., 2024b)	77.2	36.3	28.0	58.1
LLaVA-3D (Zhu et al., 2024)	79.2	41.1	30.2	63.4
Video-3D LLM (Zheng et al., 2025)	83.8	42.4	28.9	62.3
SR-3D	**97.9**	**44.7**	**31.5**	**67.3**

Table 10: Full results on Scan2Cap (Chen et al., 2021) validation set.

	EM	Bleu-1 ↑	Bleu-2 ↑	Bleu-3 ↑	Bleu-4 ↑	Rouge ↑	Meteor ↑	Cider ↑
ScanQA (Azuma et al., 2022)	21.1	30.2	20.4	15.1	10.1	33.3	13.1	64.9
3D-VisTA (Zhu et al., 2023)	22.4	–	–	–	10.4	35.7	13.9	69.6
Oryx-34B (Liu et al., 2024b)	–	38.0	24.6	–	–	37.3	15.0	72.3
LLaVA-Video-7B (Zhang et al., 2024c)	–	39.7	26.6	9.3	3.2	44.6	17.7	88.7
3D-LLM (Flamingo) (Hong et al., 2023)	20.4	30.3	17.8	12.0	7.2	32.3	12.2	59.2
3D-LLM (BLIP2-flant5) (Hong et al., 2023)	20.5	39.3	25.2	18.4	12.0	35.7	14.5	69.4
Chat-3D (Wang et al., 2023b)	–	29.1	–	–	6.4	28.5	11.9	53.2
NaviLLM (Zheng et al., 2024)	23.0	–	–	–	12.5	38.4	15.4	75.9
LL3DA (Chen et al., 2024b)	–	–	–	–	13.5	37.3	15.9	76.8
Scene-LLM (Fu et al., 2024a)	27.2	43.6	26.8	19.1	12.0	40.0	16.6	80.0
LEO (Huang et al., 2024c)	–	–	–	–	11.5	39.3	16.2	80.0
Grounded 3D-LLM (Chen et al., 2024c)	–	–	–	–	13.4	–	–	72.7
ChatScene (Huang et al., 2024b)	21.6	43.2	29.1	20.6	14.3	41.6	18.0	87.7
LLaVA-3D (Zhu et al., 2024)	27.0	–	–	–	14.5	50.1	20.7	91.7
Video-3D LLM (Zhang et al., 2024c)	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 (Azuma et al., 2022) validation set.

	What	Is	How	Can	Which	Others	Avg.
SQA3D (Ma et al., 2023)	31.6	63.8	46.0	69.5	43.9	45.3	46.6
3D-VisTA (Zhu et al., 2023)	34.8	63.3	45.4	69.8	47.2	48.1	48.5
LLaVA-Video(Zhang et al., 2024c)	42.7	56.3	47.5	55.3	50.1	47.2	48.5
Scene-LLM (Fu et al., 2024a)	40.9	69.1	45.0	70.8	47.2	52.3	54.2
LEO (Huang et al., 2024c)	–	–	–	–	–	–	50.0
ChatScene (Huang et al., 2024b)	45.4	67.0	52.0	69.5	49.9	55.0	54.6
LLaVA-3D (Zhu et al., 2024)	–	–	–	–	–	–	55.6
Video-3D LLM (Zheng et al., 2025)	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 (Ma et al., 2023) testing set.

D More Qualitative 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.

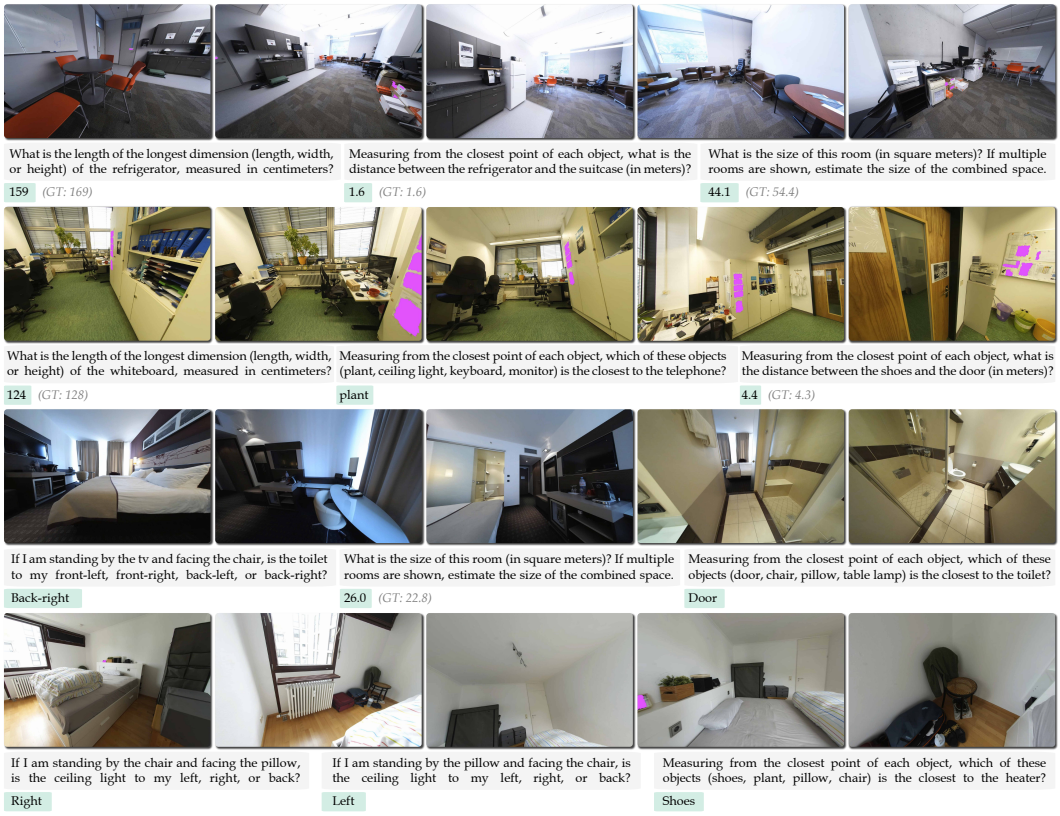

Figure 6: More results on VSI-Bench (Yang et al., 2025b). We highlight SR-3D 's outputs and include ground-truth values for numerical answers.

E 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.

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 on 3D region and 3D global tasks.

F Statistics of SR-3D-Bench

Our benchmark follows template designs from prior works on spatial reasoning in vision-language models, including SpatialRGPT (Cheng et al., 2024) and SpatialVLM (Chen et al., 2024a). To further increase the complexity and diversity of spatial reasoning tasks, we incorporate situated annotations from the EmbodiedScan (Wang et al., 2024d) 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.

G Implementation Details of SR-3D

We use PaliGemma (Beyer et al., 2024) as our visual backbone with an input size of 448 and a patch size of 14, paired with a Qwen-2-7B (Bai et al., 2023) 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 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.

H 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 of 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 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.

I USE OF LLM

To improve the clarity and presentation of this manuscript, we used large language models for minor editorial suggestions on grammar and sentence structure. The core scientific ideas, experimental work, and original text were authored exclusively. We critically evaluated every change proposed by the LLM to guarantee that the final manuscript is a faithful and accurate representation of our research and findings.