

SNOW: Spatio-Temporal Scene Understanding with World Knowledge for Open-World Embodied Reasoning

Tin Stribor Sohn^{1,3,*}, Maximilian Dillitzer^{2,3,*}, Jason Corso⁴,
Johannes Bach³, and Eric Sax¹

¹ Karlsruhe Institute of Technology, Karlsruhe, Germany

² University of Applied Science Esslingen, Esslingen, Germany

³ Dr. Ing. h.c. F. Porsche AG, Weissach, Germany

⁴ University of Michigan, Ann Arbor MI 48109, USA

Abstract. Autonomous robotic systems require spatio-temporal understanding of dynamic environments to ensure reliable navigation and interaction. While Vision-Language Models (VLMs) provide open-world semantic priors, they lack grounding in 3D geometry and temporal dynamics. Conversely, geometric perception captures structure and motion but remains semantically sparse. We propose **SNOW** (Scene Understanding with Open-World Knowledge), a training-free and backbone-agnostic framework for unified 4D scene understanding that integrates VLM-derived semantics with point cloud geometry and temporal consistency. SNOW processes synchronized RGB images and 3D point clouds, using HDBSCAN clustering to generate object-level proposals that guide SAM2-based segmentation. Each segmented region is encoded through our proposed **Spatio-Temporal Tokenized Patch Encoding (STEP)**, producing multimodal tokens that capture localized semantic, geometric, and temporal attributes. These tokens are incrementally integrated into a **4D Scene Graph (4DSG)**, which serves as 4D prior for downstream reasoning. A lightweight SLAM backend anchors all STEP tokens spatially, providing the global reference alignment, and ensuring unambiguous spatial grounding across time. The resulting 4DSG forms a queryable, unified world model through which VLMs can directly interpret spatial scene structure and temporal dynamics. Experiments on a diverse set of benchmarks demonstrate that *SNOW* enables precise 4D scene understanding and spatially grounded inference, setting new state-of-the-art performance in several settings, highlighting the importance of structured 4D priors for embodied reasoning and autonomous robotics.

Keywords: Scene Understanding · Spatio-Temporal Reasoning · Robotic Agents

*Equal contribution.

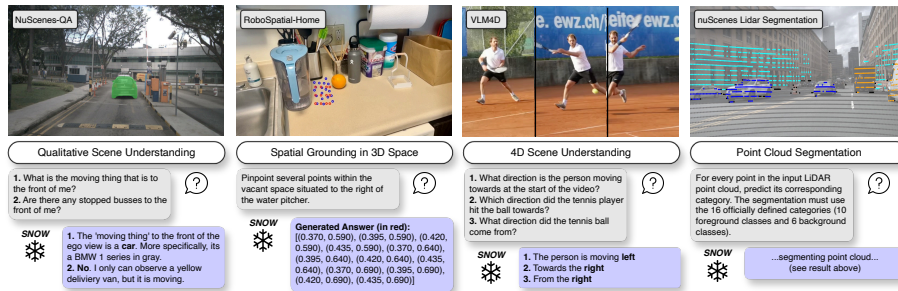


Fig. 1: Overview of SNOW. SNOW constructs a unified 4D Scene Graph (4DSG) by integrating VLM-derived open-world semantics with 3D geometric structure and temporal continuity. Spatio-Temporal Tokenized Patch Encoding (STEP) tokens encode object-level semantic, spatial, and temporal attributes, which accumulate into a persistent 4D representation that enables grounded reasoning and generalizes across diverse 4D benchmarks without additional training.

1 Introduction

Robotic systems operating in unstructured, dynamic environments must reason not only about which objects are present, but also how they are situated in 3D space and how they evolve over time. This requires the integration of open-world semantics with geometrically precise and temporally coherent scene representations [47]. Existing approaches expose a fundamental disconnect: Vision-Language Models (VLMs), relying on tokenized image patches, provide rich semantic priors and general world knowledge [20, 36, 58], yet their reasoning remains weakly grounded in spatial geometry and temporal continuity [5, 45, 52, 53, 65]. Conversely, geometric perception systems capture structure and motion, but are limited in semantic expressiveness and open-vocabulary flexibility. A unified 4D representation within VLMs is therefore required to connect semantic abstraction with persistent spatial and temporal grounding.

To address this challenge, we introduce **SNOW** (**S**cene **U**nderstanding with **O**pen-**W**orld **K**nowledge), a *training-free* framework that constructs a structured 4D representation from synchronized RGB images and point cloud observations. Consecutive point clouds are grouped via HDBSCAN clustering [1] to form object-level proposals, which guide SAM2 [46] in targeted segmentation. Through calibrated projection and fusion, each segmented region is associated with its geometric shape and temporal identity. We encode each object-level region using our new **Spatio-Temporal Tokenized Patch Encoding (STEP)**, a compact multimodal token representation capturing localized semantics, geometry, and time. Accumulating STEP tokens across frames yields a structured **4D Scene Graph (4DSG)**, where object entities are persistently indexed across space and time by a SLAM backend [15, 24]. The 4DSG provides a queryable 4D prior: VLMs can directly condition inference on spatially grounded semantics and temporally coherent object tracks, without fine-tuning or architectural

modification. This representation enables spatio-temporal reasoning, stable scene interpretation under motion, and consistent integration of new observations. The contributions of this work are as follows:

- We propose **SNOW**, a training-free, backbone agnostic framework that fuses open-world semantic priors from VLMs with temporally consistent 3D perception for unified 4D scene understanding.
- We introduce the **STEP encoding**, a multimodal object-level tokenization scheme that jointly encodes semantic, geometric, and temporal information.
- We construct a persistent **4DSG** that serves as a structured and queryable spatio-temporal representation through which VLMs can perform grounded reasoning in 4D.
- We demonstrate that structured 4D priors substantially improve spatial grounding, temporal coherence, and open-vocabulary scene understanding, achieving new state-of-the-art results on multiple benchmarks.

By linking semantic world knowledge of VLMs to explicit 4D structure, SNOW establishes a general foundation for grounded reasoning in autonomous and embodied systems.

2 Related Work

Recent advances in VLMs demonstrate strong progress in semantic reasoning and language-guided interaction [6, 7]. However, spatial and temporal reasoning capabilities remain comparatively underdeveloped [5, 43, 45, 52, 53, 65, 68]. Spatial reasoning is critical for localization and relational grounding, yet most VLMs either rely on compressed image embeddings or approximate geometric priors [18, 19, 37]. Temporal reasoning is often reduced to frame-level extensions of image models, where sequential dependencies are captured without maintaining explicit spatial structure [26, 27, 29]. Consequently, existing approaches struggle to jointly represent evolving 3D environments consistently across both space and time.

2.1 Spatial Grounding and Representations

Consider this broad grouping of spatial extensions: image, point cloud, and hybrid modality-based methods [64]. Image-based approaches leverage multi-view reconstruction, depth estimation, or Bird’s-Eye-View abstractions to approximate 3D structure [10, 23, 54, 71, 73]. While effective in controlled or dense settings, they often depend on specialized training and pre-aligned feature spaces. Point-based reasoning directly incorporates 3D geometry, either through projection into 2D depth maps [66, 74] or with dedicated 3D encoders [4, 9, 41, 72]. Such methods achieve fine-grained localization, but require multi-stage retraining and are bound to specific modalities. Hybrid designs combine images and point clouds to balance semantic richness with metric fidelity [16, 21, 30, 70], yet this couples performance to tailored data pipelines and trained backbone models.

2.2 Temporal Grounding and Representations

Temporal extensions have followed a similar trend. Early work uniformly samples video frames and feeds them into pretrained image-language backbones [26, 29, 35], while more advanced models introduce temporal-aware embeddings or memory mechanisms to capture longer sequences [17, 28, 32]. These approaches improve event localization and activity recognition but treat videos as 2D temporal streams, neglecting the underlying 3D geometry. Consequently, temporal understanding remains largely decoupled from spatial reasoning.

2.3 Spatio-Temporal Grounding

Spatio-temporal grounding attempts to bridge this gap by jointly modeling objects in space and time. Classical computer vision pipelines detect and track spatio-temporal tubes [49, 60, 67], while recent VLM-based variants incorporate temporal encoding into spatial features [25, 39, 51]. Further related works have explored structured representations for 4D reasoning, yet SNOW differs fundamentally in its design and capabilities. 3DGraphLLM [63] encodes local, discrete spatial and semantic relations derived from a predefined VL-SAT relation set. Similarly, PSG-4D [61] is constrained by predefined relations and objects, lacking an explicit, compact geometry-shape token interface for embodied reasoning. Critically, all mentioned approaches require extensive training for modality alignment and remain dependent on fixed backbones. Furthermore, we distinguish our work from static mapping approaches like DSM [57]; while DSM utilizes temporal observations to build a static geometric map enriched with 3D attributes for grounding, SNOW provides embodied-relevant, fine-grained 4D attributes—including precise 3D centroids, occupancy, and evolving 3D Gaussian shapes—as a unified set of tokens per object. By representing these attributes directly in the VLMs’ native token language, SNOW achieves unambiguous spatial grounding across time in a fully *training-free* and *backbone-agnostic* manner.

2.4 Gap Towards Unified, Training-free Spatio-Temporal Understanding

Overall, existing approaches are limited by (i) the reliance on extensive training to align modalities, (ii) dependence on specific backbones, and (iii) a lack of explicit metric geometry in the temporal domain. This motivates training-free methods that maintain generality while supporting diverse point cloud sources (*e.g.*, LiDAR, radar, RGB-D) and preserving spatio-temporal consistency.

3 Method

Robotic perception requires the integration of semantic richness, geometric precision, and temporal consistency. While point clouds provide accurate 3D structure, they are semantically sparse. Conversely, VLMs offer open-vocabulary semantics but lack grounding in metric space and temporal reasoning. To bridge

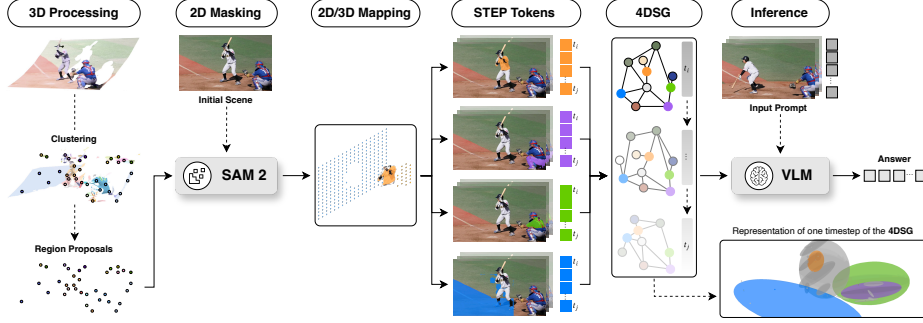


Fig. 2: High-level pipeline of SNOW. The method clusters point clouds, samples representative points, and employs them as point prompts for SAM2-based segmentation. The resulting STEP tokens form a unified spatio-temporal scene graph (*i.e.*, 4DSG), which serves as a persistent 4D world model, queryable by VLMs.

this gap, we propose SNOW, a *training-free* and *backbone-agnostic* framework for 4D spatio-temporal scene understanding. SNOW operates on synchronized RGB images and point clouds from MapAnything [24] or from LiDAR sensors. All sensors are assumed to be temporally synchronized and geometrically calibrated; experiments on sensor misalignment and noise are provided in Appendix C. The approach leverages 3D point clouds to guide SAM2-based segmentation [46], enforces cross-view and temporal consistency, and organizes all observations into a tokenized 4DSG that serves as a persistent 4D prior to VLMs (*cf.* Fig. 2). A SLAM backend [15, 24] is used for maintaining a globally consistent reference frame, ensuring unambiguous spatial alignment. On a single NVIDIA H100 GPU, the pipeline processes about 1.2 frames per second (*cf.* App. D). As detailed in our runtime analysis, the VLM backbone remains the primary bottleneck, accounting for 78.32% of total pipeline latency.

3.1 Point Cloud Clustering and Sampling

Given an input point cloud at time t , $P^t = \{p_i^t\}_{i=1}^N$ with each $p_i^t \in \mathbb{R}^3$, we initialize the set of unmapped points as $U^t \leftarrow P^t$. We cluster U^t in metric space using HDBSCAN [1], which identifies regions of high point density and prunes unstable clusters, producing a set of data-driven spatial clusters:

$$\mathcal{R}^t = \{R_1^t, \dots, R_K^t\}. \quad (1)$$

We uniformly sample m representative points $V_k^t = \{v_{k1}^t, \dots, v_{km}^t\} \subset R_k^t$ from each cluster R_k^t , which act as region proposals for subsequent mask generation. For our experiments we set $m = 4$, providing a balance between mask accuracy and computational efficiency; however, this parameter is fully adjustable for scenes containing fine-grained or cluttered structures.

3.2 Mask Generation and STEP Encoding

All points of the input cloud P^t are first projected into the image plane of each single camera c :

$$(x_i^{\text{img}}, y_i^{\text{img}}) = \pi(p_i^t, I_c^t), \quad (2)$$

where $\pi(\cdot)$ denotes the perspective projection using camera intrinsics and extrinsics. Within the same process, the projected region proposals $\{V_k^t\}^{\text{img}}$ are used as point prompts for SAM2 [46], which returns object masks

$$m_{k,c}^t \subset I_c^t. \quad (3)$$

4D Feature Encoding. Next, we associate the points from the 3D point cloud with their corresponding masks in image space. Each 3D point $(x_i^{\text{img}}, y_i^{\text{img}})$ is assigned to mask $m_{k,c}^t$ if its projection lies within the support of $m_{k,c}^t$ (i.e., $(x_i^{\text{img}}, y_i^{\text{img}}) \in m_{k,c}^t$). Each object mask $m_{k,c}^t$ is processed by our new **Spatio-Temporal Tokenized Patch Encoding (STEP)**, which merges semantic, geometric, and temporal data into unified tokens (cf. Fig. 3). The core philosophy behind STEP is that it is fundamentally instance-centric, object-level, and explicitly grounded in language. This design stems from a key human intuition: the physical objects we perceive are natively tied to linguistic concepts. By establishing this direct link, grounding effectively bridges the gap between natural language and instance-centric 4D scene understanding. Consequently, STEP provides a rich, expressive object abstraction that goes far beyond traditional scene graphs restricted to rigid, predefined categories and discrete relations. Instead of relying on static symbolic labels, each token explicitly captures continuous 3D geometry, spatial statistics, and temporal persistence. This allows SNOW to reason accurately over changing object shapes over time without requiring any additional training. The STEP encoding pipeline consists of the following steps:

1. **Mask Isolation:** We isolate each object mask $m_{k,c}^t$ by highlighting all pixels inside the mask (cf. Fig. 3 left).
2. **Image Partitioning:** The masked image is divided into a grid matching the model’s native patch size (here, 16×16).
3. **Token Filtering:** We calculate the Intersection-over-Union (IoU) between the mask and each grid cell. Cells with $\text{IoU} \geq 0.5$ are kept as *image patch tokens*, denoted as the set $\{\tau_{k,1}^t, \dots, \tau_{k,m}^t\}$.
4. **Attribute Appending:** To give these visual patches precise physical context, we attach three complementary feature tokens:
 - A *centroid token* $c_k^t = (\bar{x}, \bar{y}, \bar{z})$ that marks the exact 3D center of the object (derived from the 3D point cloud).
 - A *shape token* $s_k^t = (\Sigma, \{(a_{\min}, a_{\max}) \mid a \in \{x, y, z\}\})$ based on a 3D Gaussian modeling of the object’s spatial distribution. Here, the 3×3 covariance matrix Σ captures the 3D orientation and directional variance of the object’s geometry, while $a_{\min}$ and $a_{\max}$ define its axis-aligned boundaries. This statistical representation captures the true volumetric spread and rotation of an object without collapsing it into a rigid bounding box, while simultaneously filtering sensor noise and outliers.

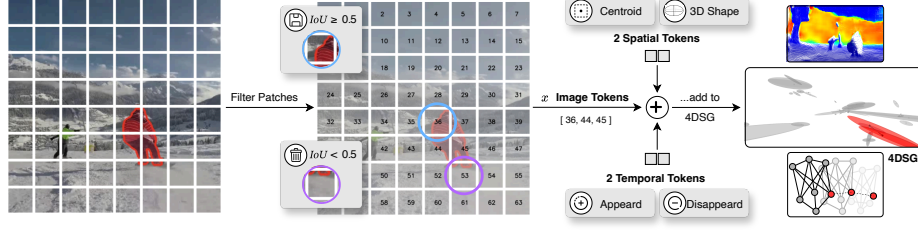


Fig. 3: STEP token assignment process. Masks with at least 50% IoU containment retain their image tokens, which are enriched with 3D centroid, Gaussian shape, and extent tokens, as well as temporal appearance and disappearance tokens. The resulting STEP tokens are assembled into a 4DSG, serving as SNOW’s persistent 4D prior.

- A pair of *temporal tokens* $\theta_k^t = (t_{\text{start}}, t_{\text{end}})$ that track exactly when the object first appears and disappears in the scene.

The complete token set for object k at time t is therefore

$$S_k^t = \{\tau_{k,1}^t, \dots, \tau_{k,m}^t, c_k^t, s_k^t, \theta_k^t\}. \quad (4)$$

These STEP tokens, as dense, structured representation, jointly capture semantic appearance, geometric structure, and temporal context, forming the atomic building blocks of the 4DSG and enabling precise 4D reasoning on evolving objects and shape that component stacking alone cannot achieve.

After STEP encoding, the unmapped point set U^t is updated and reprocessed for up to N_{iter} iterations. To enhance global consistency, an H_{hop} reasoning procedure operates on the tokenized representations, detecting implausible geometries and reassigning them to U^t (*cf.* Tab. 1).

3.3 4D Scene Graph Construction

At each time step t , SNOW constructs a spatial scene graph $\mathcal{G}^t = (\mathcal{V}^t, \mathcal{E}^t)$, where each node $v_k^t \in \mathcal{V}^t$ corresponds to a STEP-token set S_k^t . A special node encodes the ego-actors’ pose and position for self-awareness. Spatio-temporal association is modeled over a sliding window of T frames, yielding a temporally coherent sequence $\mathcal{F}_k = \{S_k^{t-T}, \dots, S_k^t\}$. We use $T = 10$ in our experiments (approximately 10 seconds of context), achieving a balanced trade-off between temporal grounding and inference efficiency. The parameter can be freely adjusted to accommodate different backbone models or resource settings. Newly observed instances detected in U^t are initialized with fresh STEP tokens, while disappeared ones are terminated by marking their final timestamp θ_k^t . The resulting sequences $\{\mathcal{F}_k\}$ form the node-level temporal representation used in the 4DSG. Aggregating these yields the unified 4DSG:

$$\mathcal{M}^t = (\mathcal{G}^{t-T:t}, \{S_k^{t-T:t}\}). \quad (5)$$

Table 1: H_{hop} reasoning refinement. Example of hop-based filtering on the 4DSG, removing implausible motions and outliers before answering spatio-temporal queries.

H_{hop} Reasoning Refinement Example
Q: What object moved the most in the last 2 seconds? A: Strictly by world-coordinate displacement between $t = 1$ and $t = 2$, object_id 32 (pedestrian) moved the most: about 32.2 m ($\Delta x \approx +0.13$ m, $\Delta y \approx +32.06$ m, $\Delta z \approx -3.73$ m). This magnitude is implausible for a standing person and the shape/extents changed markedly, suggesting an ID switch or association error. Excluding that outlier and considering consistent tracks, the next largest (and likely true) motion is object_id 8 (car), which moved about 22.4 m ($\Delta x \approx -18.61$ m, $\Delta y \approx +12.42$ m, $\Delta z \approx -0.88$ m) over the 2 s timestep (≈ 11.2 m/s).

To ensure consistent spatial alignment across frames, $\mathcal{M}^t$ is anchored in a globally referenced coordinate system using a SLAM backend: MapAnything [24] for image-only input and KISS-SLAM [15] for LiDAR.

3.4 Inference with VLMs

Inference is performed directly on the unified 4DSG $\mathcal{M}^t$ through the STEP representation. Unlike prior scene graphs that store predefined categories and symbolic relations, SNOW interfaces with the VLM via its *native input interface*. Concretely, each object node is serialized into (i) standard visual patch tokens obtained from the masked image regions and (ii) tokens that explicitly parameterize the object’s metric geometry and temporal extent, including 3D Gaussian statistics, 3D centroid coordinates, and temporal interval. As both STEP-components are expressed directly in the model’s “token language,” the VLM can consume and reason over the 4DSG without additional alignment training or feature pooling. Formally, we express inference as:

$$\hat{y} = \text{VLM}(q \mid \mathcal{M}^{t-T:t}), \quad (6)$$

where $\mathcal{M}^t$ encodes the 4D dynamics within the temporal window $t - T : t$, q denotes the query, and $\hat{y}$ is the predicted output. Since the representation is *training-free* and *backbone-agnostic*, it can be integrated with different VLM architectures and sensing modalities (LiDAR, radar, RGB-D), ensuring flexibility across robotic domains. The full pipeline of SNOW is summarized in Algorithm 1.

4 Experiments

4.1 Experimental Setting

We evaluate SNOW on five complementary benchmarks targeting semantic, spatial, and temporal understanding. In addition to the primary evaluations on

Table 2: NuScenes-QA evaluation. Accuracy scores ($\uparrow$) include Existence (Ext.), Count (Cnt.), Object, (Obj.), Status (Sts.), Comparison (Cmp.), and overall Accuracy (Acc.). **Bold** indicates the highest score and underline the second highest score.

Method	Ext.	Cnt.	Obj.	Sts.	Cmp.	Acc.
LLaMA-AdapV2 [13]	19.3	2.7	7.6	10.8	1.6	9.6
LLaVA1.5 [31]	45.8	7.7	7.8	9.0	52.1	26.2
LiDAR-LLM [62]	74.5	15.0	37.8	45.9	57.8	48.6
OccLLaMA3.1 [54]	80.9	19.2	46.3	47.8	66.6	54.5
BEVDet+BUTD [42]	83.7	22.0	48.8	52.0	67.7	57.0
OpenDriveVLA-0.5B [70]	83.9	22.0	50.2	<u>57.0</u>	68.4	58.4
OpenDriveVLA-3B [70]	<u>84.0</u>	22.3	<u>50.3</u>	56.9	<u>68.5</u>	<u>58.5</u>
OpenDriveVLA-7B [70]	84.2	<u>22.7</u>	49.6	54.5	68.8	58.2
SNOW (Ours)	82.3	27.4	53.2	80.5	61.0	60.1

NuScenes-QA [42], RoboSpatial-Home [48], ERQA [50], and VLM4D [69], we include NuScenes-based LiDAR segmentation [11] to assess downstream performance of spatial accuracy and temporal consistency, with results reported in Appendix B.1. An ablation on VLM4D further isolates the effect of integrating 4D STEP tokens into the reasoning process. All evaluations follow the official scoring protocols of their respective benchmarks. Overall, the suite is designed to highlight SNOW’s versatility in accurately grounding information in 3D and 4D. We note the limited temporal scope of RoboSpatial-Home and use it specifically to demonstrate strong *spatial grounding* enabled by explicit object-level, metric 3D features in STEP. For all experiments, we configure SNOW with a local observation window of $T = 10$ frames, $N_{\text{iter}} = 1$, and $H_{\text{hop}} = 1$. Further information on algorithm and parameter choices can be seen in Appendix H. The video predictor of SAM2_Hiera_Large [46] is employed for semantic segmentation, while Gemma3-4B-IT [14] serves as our primary backbone. We further demonstrate SNOW’s *backbone-agnostic* nature by evaluating with Qwen3-VL-8B-Instruct [59] and Llama-4-Scout-17B-16E [38] in Appendix B.2. MapAnything [24] serves as the primary SLAM backend, while KISS-SLAM [15] is used for experiments that provide access to LiDAR sensors.

4.2 Main Results

NuScenes-QA Evaluation. As shown in Table 2, SNOW establishes a new state-of-the-art on NuScenes-QA [42] with an overall accuracy of 60.1% despite operating entirely without training or fine-tuning. The most pronounced improvement appears in the *Status* category (+23.5%), indicating that SNOW’s STEP-tokenized spatio-temporal 4DSG enables explicit reasoning over dynamic object states such as motion, orientation, or occlusion. Additional gains emerge in *Count* (+4.7%) and *Object* (+2.9%), reflecting enhanced multi-entity grounding and robust object identity preservation. While performance in the *Existence* category

Table 3: RoboSpatial-Home grounded VQA evaluation. Models are evaluated on three dimensions, with Avg. reporting their mean. “+RS” denotes models finetuned on the RoboSpatial dataset. SNOW operates in a fully training-free setting, using its 4DSG as a persistent spatial representation. **Bold** indicates the highest score and underline the second highest score; higher is better ($\uparrow$).

Method	Configuration	Context	Compatibility	Avg.
VILA [48]	57.8	0.0	69.0	42.3
VILA +RS [48]	65.9	15.6	78.0	53.2
LLaVA-NeXT [48]	68.3	0.0	70.5	46.3
LLaVA-NeXT +RS [48]	78.9	19.7	80.1	59.6
SpaceLLaVA [48]	61.0	2.5	61.0	41.5
SpaceLLaVA +RS [48]	71.6	13.1	72.4	52.4
RoboPoint [48]	69.9	19.7	70.5	53.4
RoboPoint +RS [48]	78.0	31.1	81.0	63.4
Molmo [48]	58.6	0.1	18.1	25.6
GPT-4o [48]	77.2	5.7	58.1	47.0
NaviMaster [34]	–	21.7	–	–
Gemma-3-4B-IT [14]	61.8	3.3	67.6	43.1
SNOW (Ours)	<u>84.6</u>	<u>54.9</u>	78.1	<u>72.3</u>
SNOW +RS (Ours)	85.4	59.0	<u>80.0</u>	74.6

is comparable to prior work, the evaluation shows a lower score in the *Comparison* category. We empirically identify this drop as a retrieval bottleneck rather than a fundamental flaw in our 4D representation. Because the *Comparison* category involves complex, multi-hop referential chains, heavily trained baselines exploit dataset-specific cues (*e.g.*, OpenDriveVLA [70] incorporates NuScenes-QA in Stage 2 training, and BEVDet+BUTD [42] explicitly trains its QA-head on this benchmark). To test this retrieval hypothesis, we introduced a simple, orientation-aware prompt to better exploit the existing 4DSG data by deriving object pose directly from the STEP shape, motion, and semantic features. On a 1,000-sample subset, this zero-shot adjustment increased *Comparison* performance from 63.4% to 74.3% (on this subset), outperforming the heavily trained state-of-the-art baseline by +5.5%. This demonstrates that our representation is inherently richer and captures genuine 4D scene understanding without linguistic over-fitting. Further analysis regarding the robustness of these results under sensor noise and temporal misalignment is provided in Appendix C.

RoboSpatial-Home Evaluation. RoboSpatial-Home evaluates grounded spatial reasoning in real indoor environments. The benchmark tests three complementary dimensions of spatial understanding: (i) *Spatial Context* measures whether a model can identify suitable free or support surfaces by predicting a point location in the scene; (ii) *Spatial Compatibility* evaluates whether a region can feasibly support a given object, formulated as binary feasibility judgments; and (iii) *Spatial Configuration* assesses relative object-to-object spatial relationships.

We compare SNOW against pretrained and RoboSpatial-finetuned models. Unlike prior approaches requiring task-specific finetuning or spatial alignment training, SNOW performs zero-shot grounding via its STEP-based 4DSG. In this setup, native Gemma3-only backbone results (43.1%) underscore SNOW’s substantial zero-shot gain (72.3%/+29.2%), establishing a new state-of-the-art average performance on RoboSpatial-Home (*cf.* Tab. 3). Most notably, SNOW improves *Spatial Context* by a substantial margin of +23.8%, the most challenging dimension requiring continuous-point spatial grounding, outperforming all prior systems including those explicitly finetuned on RoboSpatial (+RS) (*cf.* Fig. 4). SNOW further achieves +7.4% in *Spatial Configuration* and remains competitive in *Spatial Compatibility* (−2.9%), indicating consistent generalization across complementary spatial reasoning tasks. To demonstrate that SNOW can also serve as a strong backbone for supervised learning, we fine-tune the Gemma-3 backbone using STEP tokens on ScanNet [8], annotated via the RoboSpatial annotation pipeline (denoted as SNOW +RS). With this modification, performance reaches 74.6%. As shown in Table 3, task-specific supervision particularly enhances *Spatial Context* reasoning (+4.1% over zero-shot), validating STEP tokens as a high-quality foundation for both zero-shot and supervised embodied learning. These findings demonstrate that structured 4D scene representations and STEP-based grounding enable strong spatial understanding. Qualitative success and failure cases are provided and discussed in Appendix E.1.

ERQA Benchmark Evaluation. ERQA [50] evaluates embodied agents with respect to their competence in physically grounded environments. The benchmark covers a diverse set of capabilities essential for real-world deployment, such as spatial reasoning or referential grounding (*e.g.*, pointing), and high-level task planning. The quantitative results are summarized in Table 4. Despite its compact size, SNOW (utilizing a 4B parameter backbone) achieves an accuracy of 61.5%, outperforming much larger frontier models and specialized reasoning agents. Specifically, SNOW surpasses Gemini 2.0 Pro (48.3%) and Qwen3 VL 235B Thinking (52.5%) by a substantial margin. Most notably, compared to the Gemma3-4B-IT (backbone-only) baseline (40.7%), our framework provides an absolute improvement of +20.8%, demonstrating that the injection of 4D STEP tokens allows small-scale models to rival or exceed the performance of frontier architectures. Our method is only surpassed by the latest generation of reasoning-heavy models—o3 and GPT-5—which benefit from significantly larger parameter counts and intensive test-time compute.

VLM4D Evaluation. Table 5 presents a comprehensive comparison of SNOW against state-of-the-art models on the VLM4D benchmark. SNOW achieves 73.1% ego-centric and 72.8% exo-centric reasoning accuracy, corresponding to an average reasoning gain of +9.4% despite being the smallest model with only 4B parameters. In directional reasoning, SNOW attains 71.2%, surpassing the best prior model by a significant margin of +16.4%. For false positive (FP) reasoning, SNOW scores 77.8%, which is comparable to high-performing baselines, confirming that 4D spatio-temporal modeling is primarily beneficial for scenario

Table 4: ERQA evaluation. Accuracy (%) is provided for multiple-choice questions on embodied reasoning capabilities in physical environments. **Bold** and underline denote first and second highest scores ($\uparrow$).

Method	Accuracy
Gemini 2.0 Flash [50]	46.3
Gemini 2.0 Pro Experimental [50]	48.3
Qwen3 VL 32B Thinking [33]	52.3
Qwen3 VL 235B A22B Thinking [33]	52.5
GPT-4o [50]	47.0
o3 [33]	<u>64.0</u>
GPT-5 [33]	65.7
Gemma3-4B-IT [14]	40.7
SNOW (Ours)	61.5

Table 5: VLM4D evaluation. Accuracy (%) is reported for egocentric (Ego), exocentric (Exo), directional (Dir), and false positive (FP) reasoning. Final columns provide reasoning averages and overall benchmark scores. **Bold** indicates the highest score and underline the second highest score.

Model	Ego	Exo	Avg	Dir	FP	Avg	All
GPT-4o [69]	55.5	62.2	60.0	49.5	53.3	49.9	57.5
Gemini-2.5-Pro [69]	<u>64.6</u>	<u>62.9</u>	<u>63.5</u>	<u>54.8</u>	<u>80.0</u>	<u>57.3</u>	<u>62.0</u>
Claude-Sonnet-4 [69]	52.6	52.1	52.2	44.0	86.7	48.3	51.3
Llama-4-Maverick-17B [69]	52.6	54.3	53.8	53.3	51.1	53.0	53.6
Llama-4-Scout-17B [69]	48.6	56.2	53.7	53.3	75.6	55.5	54.1
Qwen2.5-VL-72B [69]	54.3	52.5	53.1	49.5	<u>80.0</u>	52.6	53.0
InternVideo2.5-8B [69]	57.2	50.5	52.7	44.3	46.7	44.5	50.7
SNOW (Ours)	73.1	72.8	72.9	71.2	77.8	71.9	72.6

comprehension rather than for FP detection. Overall, SNOW achieves an overall benchmark score of 72.6%, outperforming all baselines substantially by +10.6% on average. These results quantitatively underline that the integration of 4D STEP tokens significantly enhances the model’s ability to reason about space and time, particularly in ego-, exo-centric, and directional contexts. Qualitative examples are provided in Appendix E.2 to further illustrate SNOW’s spatio-temporal understanding along with success and failure cases.

4.3 Ablation Study

To complement the benchmark results, the ablation is designed to isolate the impact of *4D tokenization*—*i.e.*, the proposed STEP representation—on downstream 4D reasoning. In our formulation, temporal reasoning is intrinsically coupled to spatial correspondence, since *time* is operationalized through tracking ob-

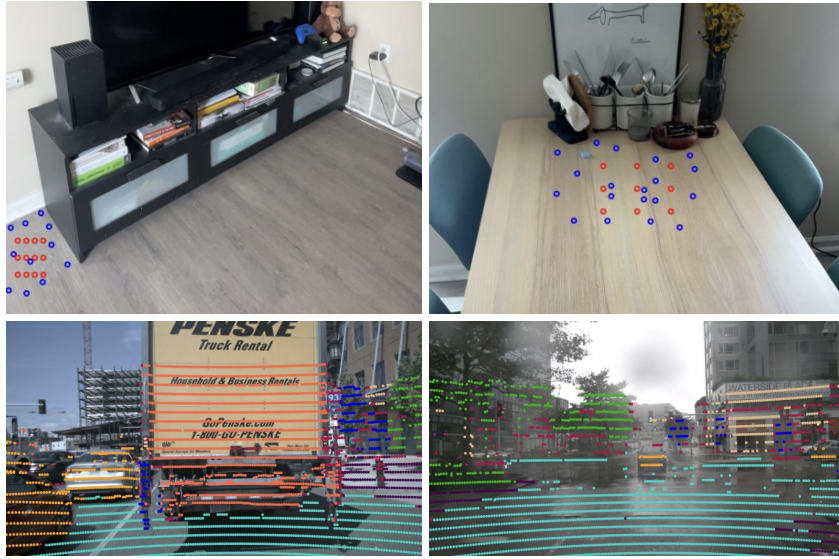


Fig. 4: SNOW on RoboSpatial-Home and open-vocabulary LiDAR segmentation. For RoboSpatial-Home: red is model prediction; blue is ground truth.

ject identities across frames. A purely temporal variant without spatial grounding is therefore ill-posed: the minimal meaningful representation already includes image-plane localization ($2D+t$). The only principled extension beyond this level is the introduction of temporally consistent metric 3D structure ($3D+t$), yielding a true 4D representation. Accordingly, the ablation focuses on the transition from $2D+t$ to 4D (STEP), directly testing our central hypothesis that temporally indexed 3D-consistent tokens are the primary driver of improved 4D reasoning. Furthermore, most state-of-the-art spatio-temporal reasoning pipelines rely on 2D tracking with temporal association while lacking persistent 3D structure (*cf.* Sec. 2). The $2D+t$ condition therefore serves as a strong and practically relevant reference point, clarifying the source of SNOW’s performance gains relative to existing approaches. Additional per-component ablations are provided in Appendix G for completeness.

A1: VLM-Only Baseline. The backbone receives RGB frames over the temporal window without explicit multi-view association or structured object tokens. This isolates the perception–language prior of the VLM.

A2: 2D Temporal Tracking. Object instances are associated over time in the image plane. Appearance and disappearance timestamps are maintained, but no 3D spatial consistency is enforced (dominant paradigm in existing methods).

A3: Full 4D STEP Representation. Spatially consistent 3D embeddings are fused with temporal instance links into unified STEP tokens, forming temporally indexed object trajectories in 4D space-time.

Table 6: Ablation study of SNOW on VLM4D. Each configuration isolates the contributions of temporal linking across frames (2D+t), and the 4D STEP tokens (4D-STEP) over the baseline model. Performance is reported using VLM4D accuracy metrics ($\uparrow$). **Bold** indicates the highest score and underline the second highest score.

ID	2D+t	4D-STEP	Ego	Exo	Avg	Dir	FP	Avg	All
A1	$\times$	$\times$	42.1	39.2	40.1	36.2	66.7	39.3	39.9
A2	$\checkmark$	$\times$	<u>61.0</u>	<u>55.0</u>	<u>57.0</u>	<u>51.2</u>	<u>71.1</u>	<u>53.3</u>	<u>56.1</u>
A3	$\times$	$\checkmark$	73.1	72.8	72.9	71.2	77.8	71.9	72.6

The progression A1→A2→A3 therefore reflects increasing structural priors: from unstructured perception, to 2D temporal continuity, to full 4D grounding. While A2 already improves performance (+16.2%) by enforcing identity persistence, the transition to STEP yields the largest gains (+32.7%; +16.5%), particularly in spatial reasoning, confirming that temporally consistent 3D grounding is the primary driver of SNOW’s performance on 3D and 4D tasks.

5 Conclusion

We presented **SNOW**, a *training-free, backbone-agnostic* framework for 4D spatio-temporal scene understanding in robotic environments. By clustering point clouds, point-prompting SAM2 for segmentation, and enriching objects with geometric and temporal attributes, SNOW unifies 3D structure, open-vocabulary semantics, and temporal dynamics into a coherent representation. Its tokenized 4DSG enables compact yet expressive encoding of object-level information and maintains temporal continuity through globally aligned semantic information. This design provides several advantages: (i) it supports plug-and-play integration with diverse VLMs and sensing modalities, (ii) generalizes across static and dynamic environments without retraining, and (iii) offers persistent memory for spatio-temporal reasoning and grounding. SNOW achieves consistent improvements across 3D and 4D understanding benchmarks, demonstrating that structured STEP tokenization can serve as a universal interface between geometric perception and foundation models. Beyond perception, SNOW offers an object-centric foundation for embodied agents, enabling unified scene interpretation, semantic mapping, and temporal abstraction in physically grounded world models.

Limitations and Future Work. The current implementation accumulates long STEP token sequences, increasing inference latency in large-scale scenes. Additionally, while the 4DSG captures global motion, it may underrepresent fine-grained dynamics or object morphing. Future work will explore (i) explicit point tracking for local motion modeling, (ii) latent-space fusion for compact token integration, (iii) encoders with learned 4D representations where SNOW acts as a data-acquisition pipeline, and (iv) STEP token ordering studies and temporal compression. These directions aim to extend SNOW toward scalable, real-time 4D scene understanding.

Acknowledgments

This research was funded, in part, by the U.S. Government under ARPA-H contract 1AY2AX000062 and DARPA contract HR0011-24-9-0429. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

References

1. Campello, R.J.G.B., Moulavi, D., Sander, J.: Density-based clustering based on hierarchical density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (eds.) *Advances in Knowledge Discovery and Data Mining*. pp. 160–172. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: Sam 3: Segment anything with concepts (2026), <https://arxiv.org/abs/2511.16719>
3. Chen, R., Liu, Y., Kong, L., Chen, N., Zhu, X., Ma, Y., Liu, T., Wang, W.: Towards label-free scene understanding by vision foundation models (2023), <https://arxiv.org/abs/2306.03899>
4. Chen, Y., Yang, S., Huang, H., Wang, T., Xu, R., Lyu, R., Lin, D., Pang, J.: Grounded 3d-llm with referent tokens (2024), <https://arxiv.org/abs/2405.10370>
5. Chu, Z., Chen, J., Chen, Q., Yu, W., Wang, H., Liu, M., Qin, B.: Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models (2024), <https://arxiv.org/abs/2311.17667>
6. Cui, C., Ma, Y., Cao, X., Ye, W., Wang, Z.: Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*. pp. 902–909 (January 2024)
7. Cui, C., Ma, Y., Cao, X., Ye, W., Wang, Z.: Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE Intelligent Transportation Systems Magazine* **16**(4), 81–94 (2024). <https://doi.org/10.1109/MITS.2024.3381793>
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes (2017), <https://arxiv.org/abs/1702.04405>
9. Deng, J., He, T., Jiang, L., Wang, T., Dayoub, F., Reid, I.: 3d-llava: Towards generalist 3d llms with omni superpoint transformer (2025), <https://arxiv.org/abs/2501.01163>
10. Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., Li, X.: Holistic autonomous driving understanding by bird’s-eye-view injected multi-modal large models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 13668–13677 (06 2024)

11. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nusenes: A large-scale benchmark for lidar panoptic segmentation and tracking. arXiv preprint arXiv:2109.03805 (2021)
12. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nusenes: A large-scale benchmark for lidar panoptic segmentation and tracking. arXiv preprint arXiv:2109.03805 (2021)
13. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model (2023), <https://arxiv.org/abs/2304.15010>
14. Gemma Team: Gemma 3 technical report (2025), <https://arxiv.org/abs/2503.19786>
15. Guadagnino, T., Mersch, B., Gupta, S., Vizzo, I., Grisetti, G., Stachniss, C.: Kiss-slam: A simple, robust, and accurate 3d lidar slam system with enhanced generalization capabilities (2025), <https://arxiv.org/abs/2503.12660>
16. Guo, Z., Zhang, R., Zhu, X., Tang, Y., Ma, X., Han, J., Chen, K., Gao, P., Li, X., Li, H., Heng, P.A.: Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following (2023), <https://arxiv.org/abs/2309.00615>
17. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-llm: Memory-augmented large multimodal model for long-term video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13504–13514 (06 2024)
18. He, Q., Peng, J., Jiang, Z., Wu, K., Ji, X., Zhang, J., Wang, Y., Wang, C., Chen, M., Wu, Y.: Unim-ov3d: Uni-modality open-vocabulary 3d scene understanding with fine-grained feature representation (2024), <https://arxiv.org/abs/2401.11395>
19. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models (2023), <https://arxiv.org/abs/2307.12981>
20. Hwang, J.J., Xu, R., Lin, H., Hung, W.C., Ji, J., Choi, K., Huang, D., He, T., Covington, P., Sapp, B., Zhou, Y., Guo, J., Anguelov, D., Tan, M.: Emma: End-to-end multimodal model for autonomous driving (2024), <https://arxiv.org/abs/2410.23262>
21. Ji, J., Wang, H., Wu, C., Ma, Y., Sun, X., Ji, R.: Jm3d & jm3d-llm: Elevating 3d representation with joint multi-modal cues. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 2475–2492 (2025). <https://doi.org/10.1109/TPAMI.2024.3523675>
22. Jiang, L., Shi, S., Schiele, B.: Open-vocabulary 3d semantic segmentation with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21284–21294 (June 2024)
23. Jiao, S., Fang, Y., Peng, B., Chen, W., Veeravalli, B.: Lavidia drive: Vision-text interaction vlm for autonomous driving with token selection, recovery and enhancement (2025), <https://arxiv.org/abs/2411.12980>
24. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zaus, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S.R., Richardt, C., Ramanan, D., Scherer, S., Kotschieder, P.: Mapanything: Universal feed-forward metric 3d reconstruction (2025), <https://arxiv.org/abs/2509.13414>
25. Li, H., Chen, J., Wei, Z., Huang, S., Hui, T., Gao, J., Wei, X., Liu, S.: Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8592–8603 (June 2025)

26. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding (2024), <https://arxiv.org/abs/2305.06355>
27. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., Wang, L., Qiao, Y.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22195–22206 (06 2024)
28. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: Leonardi, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 323–340. Springer Nature Switzerland, Cham (2025)
29. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection (2024), <https://arxiv.org/abs/2311.10122>
30. Liu, D., Huang, X., Hou, Y., Wang, Z., Yin, Z., Gong, Y., Gao, P., Ouyang, W.: Uni3d-llm: Unifying point cloud perception, generation and editing with large language models (2024), <https://arxiv.org/abs/2402.03327>
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (June 2024)
32. Liu, R., Li, C., Tang, H., Ge, Y., Shan, Y., Li, G.: St-llm: Large language models are effective temporal learners. In: Leonardi, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 1–18. Springer Nature Switzerland, Cham (2025)
33. LLM Stats: ERQA Leaderboard: Embodied Reasoning Question Answering Benchmark. <https://llm-stats.com/benchmarks/erqa> (2025), accessed: 2026-03-03
34. Luo, Z., Yan, W., Gong, J., Wang, M., Zhang, Z., Wang, X., Xie, Y., Tan, X.: Navimaster: Learning a unified policy for gui and embodied navigation tasks (2025), <https://arxiv.org/abs/2508.02046>
35. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models (2024), <https://arxiv.org/abs/2306.05424>
36. Mao, J., Qian, Y., Ye, J., Zhao, H., Wang, Y.: Gpt-driver: Learning to drive with gpt (2023), <https://arxiv.org/abs/2310.01415>
37. Mao, Y., Zhong, J., Fang, C., Zheng, J., Tang, R., Zhu, H., Tan, P., Zhou, Z.: Spatialllm: Training large language models for structured indoor modeling (2025), <https://arxiv.org/abs/2506.07491>
38. Meta Team: Llama-4-scout-17b-16e. <https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E> (2025)
39. Munasinghe, S., Thushara, R., Maaz, M., Rasheed, H.A., Khan, S., Shah, M., Khan, F.: Pg-video-llava: Pixel grounding large video-language models (2023), <https://arxiv.org/abs/2311.13435>
40. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: Openscene: 3d scene understanding with open vocabularies (2023), <https://arxiv.org/abs/2211.15654>
41. Qi, Z., Fang, Y., Sun, Z., Wu, X., Wu, T., Wang, J., Lin, D., Zhao, H.: Gpt4point: A unified framework for point-language understanding and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26417–26427 (June 2024)

42. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. *Proceedings of the AAAI Conference on Artificial Intelligence* **38**(5), 4542–4550 (Mar 2024). <https://doi.org/10.1609/aaai.v38i5.28253>, <https://ojs.aaai.org/index.php/AAAI/article/view/28253>
43. Qiu, Y., Zhao, Z., Ziser, Y., Korhonen, A., Ponti, E.M., Cohen, S.B.: Are large language models temporally grounded? (2023), <https://arxiv.org/abs/2311.08398>
44. Qwen Team: Qwen3-vl-8b-instruct. <https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct> (2025)
45. Ranasinghe, K., Shukla, S.N., Poursaeed, O., Ryoo, M.S., Lin, T.Y.: Learning to localize objects improves spatial reasoning in visual-llms. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12977–12987 (June 2024)
46. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos (2024), <https://arxiv.org/abs/2408.00714>
47. Sohn, T.S., Reis, P., Dillitzer, M., Bach, J., Corso, J.J., Sax, E.: A framework for a capability-driven evaluation of scenario understanding for multimodal large language models in autonomous driving (2025), <https://arxiv.org/abs/2503.11400>
48. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospacial: Teaching spatial understanding to 2d and 3d vision-language models for robotics (2025), <https://arxiv.org/abs/2411.16537>
49. Tang, Z., Liao, Y., Liu, S., Li, G., Jin, X., Jiang, H., Yu, Q., Xu, D.: Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(12), 8238–8249 (2022). <https://doi.org/10.1109/TCSVT.2021.3085907>
50. Team, G.R.: Gemini robotics: Bringing ai into the physical world (2025), <https://arxiv.org/abs/2503.20020>
51. Wang, J., Zhang, Z., Liu, Z., Li, Y., Ge, J., Xie, H., Zhang, Y.: Spacevlm: Endowing multimodal large language model with spatio-temporal video grounding capability (2025), <https://arxiv.org/abs/2503.13983>
52. Wang, J., Ming, Y., Shi, Z., Vineet, V., Wang, X., Li, Y., Joshi, N.: Is a picture worth a thousand words? delving into spatial reasoning for vision language models (2024), <https://arxiv.org/abs/2406.14852>
53. Wang, Y., Zhao, Y.: Tram: Benchmarking temporal reasoning for large language models (2024), <https://arxiv.org/abs/2310.00835>
54. Wei, J., Yuan, S., Li, P., Hu, Q., Gan, Z., Ding, W.: Occlama: An occupancy-language-action generative world model for autonomous driving (2024), <https://arxiv.org/abs/2409.03272>
55. Wei, W., Ülger, O., Nejadasl, F.K., Gevers, T., Oswald, M.R.: 3d-avs: Lidar-based 3d auto-vocabulary segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8910–8920 (June 2025)
56. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler, faster, stronger (2024), <https://arxiv.org/abs/2312.10035>
57. Xie, Q., Liang, Z., Li, F., Zeng, L.: Dsm: Constructing a diverse semantic map for 3d visual grounding (2025), <https://arxiv.org/abs/2504.08307>

58. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.Y.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters* **9**(10), 8186–8193 (2024). <https://doi.org/10.1109/LRA.2024.3440097>
59. Yang, A., Li, A., Yang, B., et al.: Qwen3 technical report (2025), <https://arxiv.org/abs/2505.09388>
60. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Tubedetr: Spatio-temporal video grounding with transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 16442–16453 (June 2022)
61. Yang, J., CEN, J., PENG, W., Liu, S., Hong, F., Li, X., Zhou, K., Chen, Q., Liu, Z.: 4d panoptic scene graph generation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 69692–69705. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/dc6319dde4fb182b22fb902da9418566-Paper-Conference.pdf
62. Yang, S., Liu, J., Zhang, R., Pan, M., Guo, Z., Li, X., Chen, Z., Gao, P., Guo, Y., Zhang, S.: Lidar-llm: Exploring the potential of large language models for 3d lidar understanding (2023), <https://arxiv.org/abs/2312.14074>
63. Zemskova, T., Yudin, D.: 3dgraphllm: Combining semantic graphs and large language models for 3d scene understanding (2025), <https://arxiv.org/abs/2412.18450>
64. Zha, J., Fan, Y., Yang, X., Gao, C., Chen, X.: How to enable llm with 3d capacity? a survey of spatial reasoning in llm (2025), <https://arxiv.org/abs/2504.05786>
65. Zhang, J., Cai, M., Xie, T., Lee, Y.J.: Countercurate: Enhancing physical and semantic visio-linguistic compositional reasoning via counterfactual examples (2024), <https://arxiv.org/abs/2402.13254>
66. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8552–8562 (June 2022)
67. Zhang, Z., Zhao, Z., Zhao, Y., Wang, Q., Liu, H., Gao, L.: Where does it exist: Spatio-temporal video grounding for multi-form sentences. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2020)
68. Zhao, Y., Li, Z., Jin, Z., Zhang, F., Zhao, H., Dou, C., Tao, Z., Xu, X., Liu, D.: Enhancing the spatial awareness capability of multi-modal large language model (2023), <https://arxiv.org/abs/2310.20357>
69. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, X.E., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models (2025), <https://arxiv.org/abs/2508.02095>
70. Zhou, X., Han, X., Yang, F., Ma, Y., Knoll, A.C.: Opendrivevla: Towards end-to-end autonomous driving with large vision language action model (2025), <https://arxiv.org/abs/2503.23463>
71. Zhou, Y., Huang, L., Bu, Q., Zeng, J., Li, T., Qiu, H., Zhu, H., Guo, M., Qiao, Y., Li, H.: Embodied understanding of driving scenarios. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) *Computer Vision – ECCV 2024*. pp. 129–148. Springer Nature Switzerland, Cham (2025)
72. Zhu, C., Wang, T., Zhang, W., Chen, K., Liu, X.: Scanreason: Empowering 3d visual grounding with reasoning capabilities. In: Leonardis, A., Ricci, E., Roth, S.,

- Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 151–168. Springer Nature Switzerland, Cham (2025)
73. Zhu, C., Wang, T., Zhang, W., Pang, J., Liu, X.: Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness (2025), <https://arxiv.org/abs/2409.18125>
74. Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P.: Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2639–2650 (October 2023)
75. Zou, P., Zhao, S., Huang, W., Xia, Q., Wen, C., Li, W., Wang, C.: Adaco: Overcoming visual foundation model noise in 3d semantic segmentation via adaptive label correction (2024), <https://arxiv.org/abs/2412.18255>