ROSE: Real-Time Open-World Scene Understanding from Monocular Video via Compact Multimodal 4D Scene Graphs

Ziyue Qiu[1], Yong Wang[2], and Jin Pan[3]

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health
ziyueqiu@hsph.harvard.edu
[2] University of the Chinese Academy of Sciences
yongwang17@mails.ucas.ac.cn
[3] The Chinese University of Hong Kong
jpan@link.cuhk.edu.hk

Abstract. Vision-Language Models (VLMs) have advanced rapidly in visual and semantic understanding, yet they remain weak at dynamic 4D spatial grounding in video. On spatio-temporal reasoning benchmarks, raw-video input often yields low and variable accuracy. 4D Scene Graphs (4DSGs) can supply missing structure, but two interface gaps remain: identity facets (appearance, geometry, motion) are often encoded through language-side descriptors rather than modality-separated inputs, and most training targets static spatial relations instead of dynamic 4D attributes such as trajectories and velocity. We introduce Visual Spatio-Temporal Anchor (**VISTA**), a multimodal node encoding that represents each tracked object as a persistent, queryable entity, delegating appearance to anchor crops processed by the VLM's vision encoder and geometry and motion to explicitly structured tokens. We introduce Real-Time Open-World Scene Understanding (**ROSE**), a training-free monocular pipeline (single GPU) that runs at 10 Hz. Beyond inference-time prompting, the 4DSGs produced by ROSE can serve as structured training-time context: fine-tuning a VLM on 4DSG-augmented data teaches it to use trajectory and velocity tokens for dynamic reasoning, yielding a new state of the art on DSR-Bench (67.5, +8.6 over the previous best), with gains across all 13 subtasks. ROSE achieves competitive accuracy across all three benchmarks under real-time, single-GPU monocular constraints.

Keywords: 4D Scene Understanding · Scene Graphs · Monocular Depth · Real-Time Perception · Video Segmentation · Embodied Reasoning

1 Introduction

Vision-Language Models (VLMs) have achieved substantial gains in semantic understanding, yet their ability to reason about *where things are in 3D and how they move over time* remains unreliable. Tasks such as judging relative direction, estimating object speed, or confirming whether an entity exists in a dynamic

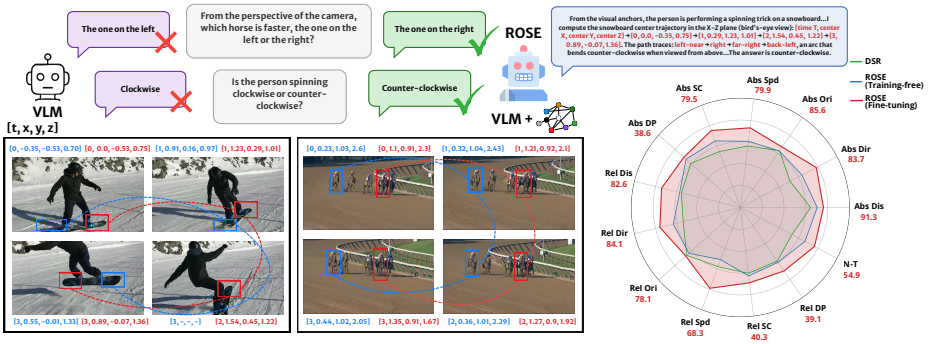

Fig. 1: Overview of ROSE. Given monocular video, ROSE constructs a multimodal 4D scene graph with persistent object tracking, 3D trajectories, and anchor crops, enabling a VLM to answer dynamic spatial questions. Notably, ROSE is applicable even under extreme out-of-view scenarios, e.g., the blue box disappears at 2s. Right: DSR-Bench [57] radar chart showing gains across all 13 sub-task types. We compare the previous state-of-the-art method DSR [57] with ROSE under both training-free and fine-tuning paradigms.

scene require joint spatial and temporal grounding that current VLMs lack. On 4D spatio-temporal benchmarks, most models, including strong open-weight ones, perform poorly when given raw video (Table 3). Scaling model capacity alone does not reliably solve dynamic spatial grounding [6, 39].

We argue that the key missing ingredient is not richer semantics but *computable 4D structure*. Recent work on 4D Scene Graphs (4DSGs) [13,36] suggests this: augmenting VLMs with object-centric spatio-temporal graphs yields measurable accuracy gains on spatial reasoning benchmarks. Yet two design-level gaps remain open (§2.3). First, a **usage gap**: prior work trains VLMs with static spatial relations [6, 7] or leverages scene graphs for embodied planning [20], but few explicitly use dynamic 4D attributes (trajectories, velocity, temporal order) as training-time context during fine-tuning. Second, a **representation gap**: existing 4DSGs encode identity facets (appearance, geometry, motion) predominantly through language-side descriptors (captions, natural-language spatial descriptions), instead of routing each facet through a dedicated modality channel. As a result, spatial attributes are not exposed as explicit structured numeric tokens. Fine-grained visual detail that the VLM vision encoder could exploit is also weakened.

To address these gaps, we introduce **VISTA** (Visual Spatio-Temporal Anchor), a modality-aligned node encoding that enables dynamic 4DSGs to function both as inference-time context and as structured training-time context. Our primary contribution is a training paradigm: we fine-tune VLMs with 4DSG context, enabling them to acquire improved dynamic reasoning from structured 4D supervision. VISTA represents each tracked object as a persistent, queryable en-

tity: appearance is routed through native image tokens, while shape and motion are routed through structured text tokens (Fig. 1, 3). We instantiate VISTA with **ROSE**, a training-free pipeline that constructs VISTA-based 4DSGs from monocular RGB video on a single GPU in real time (Fig. 2).

Our contributions are as follows:

1. **4DSGs as structured training-time context.** We show that including ROSE 4DSGs as training-time context, not merely inference-time prompts, enables a VLM to learn dynamic spatial reasoning. On DSR-Bench, ROSE (Fine-tuning) reaches **67.5** (+8.6 over DSR), with gains across all 13 subtasks (§4.2, Table 1).
2. **ROSE: real-time monocular 4DSG construction.** We propose a training-free pipeline on a single GPU that couples monocular depth and pose estimation [26] with open-world object discovery and persistent tracking, constructing VISTA-based 4DSGs at 10 Hz from monocular video under real-time, single-GPU constraints (§3.1–3.3).
3. **VISTA: a modality-aligned node encoding.** Each tracked object is represented as a multimodal node that delegates appearance to an anchor crop processed by the VLM's own vision encoder, and geometry and motion to structured text tokens. This addresses the representation gap by routing visual detail and spatial attributes through dedicated channels; ablation confirms that structured spatial tokens are a major accuracy driver, with anchor crops contributing complementary visual grounding (§3.4, Table 6).

Across VLM4D, RoboSpatial-Home, and DSR-Bench, ROSE with VISTA achieves strong accuracy on directional and motion-dependent subtasks from monocular input at 10 Hz, attaining state of the art on DSR-Bench (67.5) and competitive performance on VLM4D and RoboSpatial-Home, without relying on ground-truth depth or multi-view supervision (§4.2, Table 3).

2 Related Work

2.1 Spatial Grounding and Representations

Spatial grounding for VLMs places geometry in the 2D token stream, an explicit 3D representation, or an object-level graph. Image-centric methods add depth or region-level supervision [6,7] but are largely per-frame and under-model temporal order, which video benchmarks struggle to isolate [12]. 3D-centric methods lift inputs to point/object representations [14, 15, 29], yet many rely on stronger sensors or multi-view input [51], and scene-graph lines rarely target monocular persistent tracking [19,42,47]. For monocular pipelines the bottleneck is temporal consistency: framewise depth [31,49,50] versus video-consistent depth+pose [26] that reduces dependence on classical SLAM and structure-from-motion [33, 34, 38, 41], while first-person benchmarks show spatial grounding remains hard [9, 27, 43]. These trade-offs motivate persistent object-level interfaces in a shared world frame.

2.2 Temporal Grounding and Representations

Video LLMs use sampling, aggregation, and memory [10,25,44,45,53] to capture temporal semantics, but geometry stays mostly implicit, reflected by weaker results on geometry-sensitive benchmarks [57, 58]. Tracking systems (SAM1-3, FastSAM, DEVA, MASA) give strong mask continuity [5, 8, 22, 23, 32, 54], yet without world-frame 3D coordinates they still struggle on distance, direction and velocity reasoning, motivating explicit 3D lifting coupled with tracking.

2.3 Spatio-Temporal Scene Understanding

Recent systems combine spatial geometry with temporal continuity into structured representations, but differ along two principal axes: interface explicitness (explicit graph vs. latent geometry) and usage phase (inference-only vs. trainable). Reported setups also differ in sensor assumptions, from LiDAR- and point-cloud-based pipelines [36, 51] to image-based systems [13], and in interface design choices. Inference-time 4DSG prompting systems (e.g., SNOW, DAAAM) improve reasoning but expose scene information through token/text-serialized interfaces with different emphases (spatial QA vs. long-horizon memory/retrieval) [13,36]. Related datasets, benchmarks, and perception components in this space include spatial QA datasets and benchmarks [3,30,37], feed-forward 3D reconstruction [21], multi-object tracking [2], and localized captioning [24]. Training-side work mostly targets static spatial supervision or scene-graph prediction [6, 7, 18, 20, 46, 52], while unified 4D VLMs are emerging [56]. DSR [57] shows that geometric priors help dynamic QA, but it distills them into learned geometry tokens whose interface is not explicitly inspectable.

Two interface-level gaps remain. First, a *representation gap*: prior systems serialize identity and geometry as text/token prompts [4, 6, 7, 13, 36], and even object-token 3D LLMs [15, 16] fold appearance into token streams rather than routing it through the VLM vision encoder at query time, which weakens fine-grained grounding [28, 48]. Second, a *usage gap*: dynamic 4DSGs are rarely used as structured training-time context. ROSE targets an explicit, monocular, single-device interface that addresses both: a modality-aligned node encoding (VISTA) for the representation gap, and dynamic 4DSGs as training-time context for the usage gap, instantiated from monocular video with learned dense depth [26] in real time on a single GPU.

3 Method

From monocular RGB video alone, ROSE builds a 4DSG whose every node is a VISTA encoding of one tracked object, $F_k = (\mathcal{I}_k,\ e_k,\ \{(t, c_k^t)\}_{t\in\mathcal{T}_k})$ (§3.4): an anchor image $\mathcal{I}_k$, a 3D extent e_k, and a world-frame centroid trajectory $\{(t, c_k^t)\}_{t\in\mathcal{T}_k}$. Three stages produce it: DA3 returns temporally consistent depth and camera poses (§3.1), a two-pass FastSAM+SAM3 schedule returns persistent open-world tracks (§3.2), and per-mask back-projection lifts each track to 3D

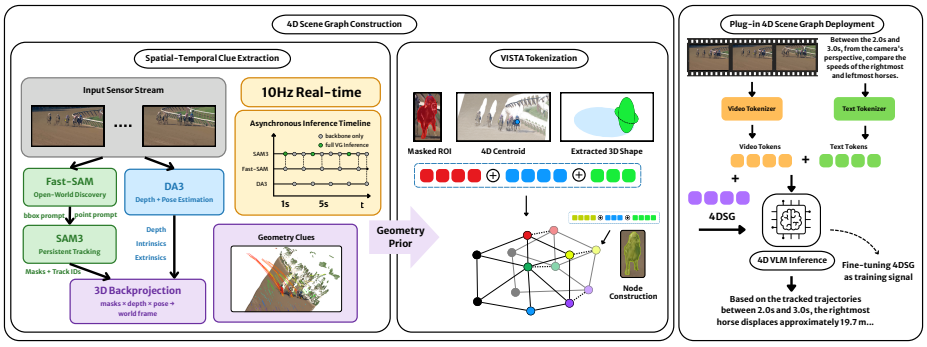

Fig. 2: ROSE pipeline. ROSE takes monocular video and runs two parallel streams: class-agnostic segmentation with persistent tracking, and metric depth with camera pose estimation. Their outputs are fused into a 4DSG, where each object is encoded as a VISTA node with an anchor crop, a 3D extent, and trajectory tokens $[x, y, z, t]$. The resulting 4DSG, together with sampled keyframes, is used by a VLM for grounded spatio-temporal QA.

(§3.3). Depth/pose and segmentation/tracking run as two concurrent streams; because the tracking stream dominates latency, ROSE concentrates acceleration there and overlaps the rest around it, reaching real-time single-GPU throughput (§3.5).

3.1 Monocular Depth and Pose Estimation

3D lifting needs per-frame metric depth, and a shared world frame needs per-frame camera pose. Instead of per-frame depth models [31,49,50], whose predictions drift from frame to frame, we process the whole clip jointly with DA3 [26]:

$$\mathrm{DA3}\big(\{I_t\}_{t=0}^{N-1}\big) \;\rightarrow\; \big\{(D_t,\; K_t,\; T_t^{wc},\; \gamma_t)\big\}_{t=0}^{N-1}, \tag{1}$$

yielding, per frame, a depth map $D_t \in \mathbb{R}^{H\times W}$, intrinsics $K_t \in \mathbb{R}^{3\times 3}$, world-to-camera pose $T_t^{wc} \in \mathrm{SE}(3)$, and per-pixel depth confidence $\gamma_t \in [0,1]^{H\times W}$; the inverse $T_t^{cw} = (T_t^{wc})^{-1}$ is the ego pose, and γ_t gates unreliable pixels at lifting (§3.3). Superscripts wc and cw denote world-to-camera and camera-to-world throughout.

Joint inference exploits inter-frame cues for temporally consistent geometry and removes external SLAM; we normalize all poses to frame 0, which then defines the world origin ($T_0^{wc} = I$). Clips that exceed GPU memory are split into overlapping chunks and re-aligned across boundaries by SIM3 [40]. DA3 runs on its own device stream, concurrent with segmentation/tracking (§3.5), and the reported 10 Hz is end-to-end pipeline throughput.

3.2 Two-Pass Segmentation and Tracking

VISTA nodes need stable, long-lived identities, but initialize-and-propagate tracking sees only first-frame detections and drops objects that enter later. ROSE instead runs a two-pass schedule over one shared SAM3 [5] memory bank: a discovery pass adds object prompts and a propagation pass tracks them jointly, with SAM3 associating per-frame detections to memory slots by bipartite matching.

Initialization and Full Propagation. FastSAM [54] emits class-agnostic masks with boxes and confidence scores at periodic anchor frames. We de-duplicate these proposals across anchors by bounding-box IoU and mask intersection-over-minimum, keeping one box per physical object, then seed a single SAM3 multiplex session at the earliest non-empty anchor with the surviving boxes batched as box prompts. SAM3 propagates all boxes jointly across the N frames in one forward pass and caches the per-frame masks $\hat{m}_k^t$ for reuse by discovery and lifting.

Mid-Video Object Discovery. Every δ_{stride} frames FastSAM re-runs, and each new detection m_j^t is matched against the cached masks; it counts as novel when its best IoU with them falls below δ_{disc}:

$$\text{novel}(m_j^t) = \mathbb{1}\big[\max_i \text{IoU}(m_j^t, \hat{m}_i^t) < \delta_{\text{disc}}\big]. \tag{2}$$

A novel detection enters the live session as a point prompt at its mask centroid (normalized coordinates), without resetting the memory bank, subject to per-frame and per-video discovery caps and a minimum-area filter (§4.1). SAM3 then partial-propagates only the newly added objects and merges them into the cache, yielding per-object, per-frame masks $\{(\hat{m}_k^t, \hat{s}_k^t, r_k)\}$ with tracking confidence $\hat{s}_k^t \in [0,1]$ and originating run r_k.

3.3 3D Lifting and Track Management

Back-Projection. Each mask $\hat{m}_k^t$ is lifted to the world frame from the DA3 depth and pose (§3.1): we drop low-confidence pixels ($\gamma_t[v,u] < \tau_\gamma$, with τ_γ in §4.1) and back-project the rest,

$$\mathbf{p}_{\text{world}}(u,v) = T_t^{cw} \begin{bmatrix} D_t[v,u] \cdot K_t^{-1}[u,\ v,\ 1]^\top \\ 1 \end{bmatrix}, \tag{3}$$

giving a per-detection point set P_k^t, from which we read a centroid $c_k^t \in \mathbb{R}^3$ and a per-axis shape descriptor s_k^t:

$$c_k^t = \frac{1}{|P_k^t|} \sum_{\mathbf{p} \in P_k^t} \mathbf{p}, \qquad s_k^t = (\mu_a, \sigma_a, a_{\min}, a_{\max})_{a \in \{x,y,z\}}. \tag{4}$$

The extent token uses the robust bounds $a_{\min}, a_{\max}$ (Eq. 5), while μ_a, σ_a drive a per-axis 3σ clip that removes depth outliers before those bounds are read.

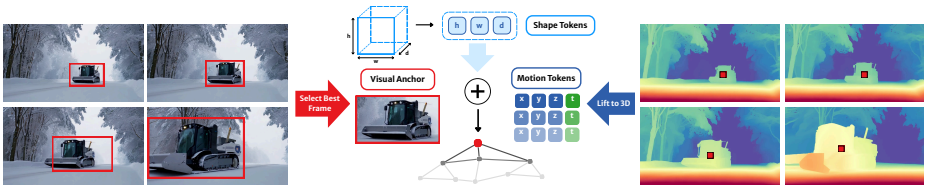

Fig. 3: VISTA node encoding. Each tracked object is represented as a persistent multimodal node: an anchor crop for appearance, shape tokens for 3D extent, and motion tokens $[x, y, z, t]$ for trajectory. This design routes appearance, geometry, and motion through separate channels.

Track Management. Discovery can assign several ids to one object, so a three-step consolidation runs over all tracks. *(i) Cross-run merge.* Detections from different SAM3 runs are merged transitively into one identity (a Union-Find over pairwise matches) when they agree on 2D overlap (IoU $> \delta_{\text{iou}}$), 3D proximity (centroid distance $< \delta_{\text{dist}}$), and temporal continuity (gap $< \delta_{\text{gap}}$); non-metric backbones swap δ_{dist} for the depth-normalized gate $\delta_{\text{dist,rel}}$. *(ii) Lifecycle.* A track stays active while observed, becomes lost after a gap of δ_{lost} frames, and is archived after δ_{arch} further frames, beyond which it is frozen from merging. *(iii) Re-identification.* Long-occlusion splits are rejoined the same way, on anchor-crop cosine similarity ($> \delta_{\text{sim}}$), recovering objects that disappear and reappear. Thresholds are listed in §4.1.

3.4 Visual Spatio-Temporal Anchor (VISTA) Encoding

VISTA encodes each track k as a backbone-agnostic multimodal node with three modality-separated components (Fig. 3), built from the depth, poses, tracks, and point sets above:

Visual Anchor $\mathcal{I}_k$. $\mathcal{I}_k$ is the track's highest-confidence SAM3 frame, ties broken by crop brightness as an exposure prior, cropped to the padded bounding box so surrounding context is kept rather than masked to the silhouette. It is fed to the VLM as a native image token, so recognition reuses the VLM's own vision encoder with no captioning model; keeping box context instead of the masked silhouette markedly improves identification, especially for smaller VLMs (§4.4).

Shape Tokens e_k. The extent e_k is the per-axis median of the per-frame box sizes $s_k^t[a_{\max}] - s_k^t[a_{\min}]$ over the whole track (from s_k^t, Eq. 4):

$$e_k = \text{median}_{t \in \mathcal{T}_k} \left(s_k^t[a_{\max}] - s_k^t[a_{\min}] \right)_{a \in \{x,y,z\}} \in \mathbb{R}^3. \tag{5}$$

The median suppresses occlusion- and depth-noise outliers and collapses the per-frame 12D descriptor to a single 3D vector.

Motion Tokens $\{(t, c_k^t)\}$. The motion token is the centroid trajectory: each retained timestamp $t \in \mathcal{T}_k$ contributes an explicit 4D tuple $[x, y, z, t]$ whose $c_k^t = [x, y, z]$ is the world-frame centroid (Eq. 4).

4DSG Assembly. A node bundles the three components,

$$F_k = \big(\underbrace{\mathcal{I}_k}_{\text{image}},\ \underbrace{e_k,\ \{(t, c_k^t)\}_{t \in \mathcal{T}_k}}_{\text{structured text}} \big). \tag{6}$$

and the serialized 4DSG is the set over all tracks,

$$\mathcal{M} = \{F_k\}_{k=1}^{K}, \tag{7}$$

plus per-track metadata (coarse 2D image position, anchor reference) for the prompt template. Values are world-frame metric units (meters, seconds) with no per-scene normalization, and camera poses, used only upstream for lifting, are never serialized as an ego trajectory in $\mathcal{M}$. The VLM answers $\hat{y} = \mathrm{VLM}(q \mid \mathcal{M})$ for a query q; fixed keyframe and object budgets (N_{kf}, K; §4.1) hold $\mathcal{M}$ within the context window.

Relational structure. Alongside the per-object nodes, ROSE defines explicit pairwise edges. An edge R_{ij} between objects i and j is a time-indexed sequence of relative orientations: its entry at each shared timestamp t is the world-frame unit direction $(c_j^t - c_i^t)/\|c_j^t - c_i^t\|$ between their centroids. Because all nodes share one metric world frame with synchronized timestamps, each edge follows directly from the centroid trajectories serialized in $\mathcal{M}$, exposing cross-object relations (relative position, direction, distance) explicitly, following SNOW's object-level token-sequence interface [36].

3.5 Real-Time Execution on a Single GPU

Real-time construction forbids any dominant stage. The tracking stream (§3.2) is the bottleneck: memory-attention propagation and mask decoding fire many tiny kernels per frame and leave the GPU near half-idle, so our optimizations target it and overlap the rest around it. At their default settings each is either numerically exact or a controlled approximation confined to interpolated frames.

Joint Multi-Object Tracking. The SAM3 object-multiplex predictor tracks all active objects through one shared backbone and memory bank in a single forward pass, replacing per-object sessions whose cost grows linearly in track count; this is the dominant saving in crowded scenes.

Strided Propagation with Flow Interpolation. Propagation, not depth or lifting, sets the frame budget, so SAM3 propagates only every s-th frame; each skipped frame t inherits its masks by warping the nearest preceding keyframe mask along dense optical flow rather than copying it,

$$\hat{m}_k^t = \mathcal{W}\big(\hat{m}_k^{\kappa(t)};\ \mathbf{f}_{\kappa(t) \to t}\big), \qquad \kappa(t) = \lfloor t/s \rfloor \cdot s, \tag{8}$$

Algorithm 1 ROSE pipeline

Require: Monocular RGB video $\{I_t\}_{t=0}^{N-1}$, query q
Ensure: 4DSG $\mathcal{M}$ and answer $\hat{y}$
$\{D_t, K_t, T_t^{wc}, \gamma_t\} \leftarrow \mathrm{DA3}(\{I_t\})$ *§3.1: depth & pose*
$\{(\hat{m}_k^t, \hat{s}_k^t, r_k)\} \leftarrow \text{TwoPass-FastSAM-SAM3}(\{I_t\})$ *§3.2: segment & track*
$\{(c_k^t, s_k^t)\} \leftarrow \mathrm{BackProject}(\hat{m}_k^t, D_t, K_t, T_t^{cw}, \gamma_t)$ *§3.3: 3D lifting*
$\{F_k\} \leftarrow \mathrm{VISTA}\big(\mathcal{I}_k, e_k, \{(t, c_k^t)\}_{t \in \mathcal{T}_k}\big)$ *§3.4: encode 4DSG*
$\mathcal{M} \leftarrow \{F_k\}_{k=1}^{K}$ *assemble serialized 4DSG*
$\hat{y} \leftarrow \mathrm{VLM}(q \mid \mathcal{M})$ *answer*
return $\mathcal{M}, \hat{y}$

with $\mathbf{f}_{\kappa(t)\to t}$ the flow from keyframe $\kappa(t)$ to t and $\mathcal{W}$ the warp. Depth and lifting still run every frame, so centroids and velocities track true motion without stair-stepping, and propagation cost scales as $1/s$ while the optical flow stays cheap and parallel across skipped frames.

Feature Reuse and Bounded Re-Propagation. Image-encoder features are cached and shared across full and partial passes, so late-object propagation and periodic re-grounding skip redundant detection. A mid-video discovery (§3.2) re-propagates only within a bounded backward window, not from frame 0, and a compact memory bank (a fixed frame count sampled at a temporal stride) caps per-frame attention while keeping a long effective horizon.

Kernel-Level Acceleration. Two Hopper-targeted kernels are on by default. The long-sequence modules (image backbone and memory-attention encoder) run FlashAttention-3 with FP8, while the short-sequence mask-decoder heads keep SDPA-flash, where the FP8 cast does not pay off. Because `torch.compile` fuses operators but cannot capture CUDA graphs under the tracker's dynamic shapes, launch overhead persists on a launch-bound, half-idle GPU; we therefore capture the steady-state memory-attention encoder as a shape-keyed CUDA graph and replay it, collapsing hundreds of launches into one. The replay is bit-exact, and FA3's FP8 attention is approximate but accuracy-neutral.

Cross-Stage Overlap. Depth/pose and tracking run on separate device streams; per-frame 3D lifting (§3.3) is threaded across CPU cores, since back-projection releases the interpreter lock; and the pairwise mask IoU for discovery and merging is one batched matrix product on the GPU. Across clips, a persistent model pool amortizes weight loading and compilation and overlaps one clip's CPU-side 4DSG assembly with the next clip's GPU work.

4 Experiments

We evaluate ROSE on three benchmarks: DSR-Bench [57] (dynamic spatial reasoning; 1,484 QA pairs), VLM4D [58] (4D spatio-temporal reasoning), and RoboSpatial-Home [37] (indoor grounded spatial QA).

4.1 Experimental Setup

Implementation. All experiments run on a single NVIDIA H200 GPU. The 4DSG construction pipeline runs at 10 Hz (VLM latency reported separately in §4.3) with peak VRAM well within the H200's memory budget. We use DA3-Giant (metric depth) for depth/pose, FastSAM-s with the SAM3 object-multiplex tracker for segmentation/tracking, and Qwen2.5-VL-32B as the primary VLM backbone. The efficiency techniques of §3.5 are enabled by default: propagation stride s=3 with optical-flow interpolation, a seven-frame memory bank at temporal stride four, FlashAttention-3 on the long-sequence Hopper attention modules, a compiled mask-decoder transformer with shape-keyed CUDA-graph replay of the memory-attention encoder, and parallel 3D lifting. Detailed hyperparameters are in the supplementary appendix.

Baselines. We compare against SNOW [36]. Frontier VLMs on raw video include GPT-4o, GPT-5, Gemini-2.5-Pro, and Claude-Sonnet-4. We also include Llama-4-Maverick/Scout-17B, Qwen2.5-VL-72B, and InternVideo2.5-8B. We additionally compare to DSR [57] (task-specific training) and other task-specific baselines where available.

Evaluation Controls. To ensure fair comparison, we cap the keyframe and object budgets at N_{kf}=32 and K=32. In DSR-Bench, a token-matched null baseline (**LenCtrl**) controls for input-length effects.

4.2 Experimental Results

DSR-Bench Evaluation: 4DSG as a Training-Time Representation. Our key hypothesis is that dynamic 4DSGs can serve as structured training-time context, included alongside training examples, so that a VLM learns to use them for dynamic spatial reasoning. We test this directly on DSR-Bench with DSR-Train (50k QA pairs). The three conditions share DSR's Qwen2.5-VL-7B backbone and training recipe [57] and differ only in 4DSG exposure. DSR+LenCtrl replaces the serialized 4DSG $\mathcal{M}$ (Eq. 7) with a token-matched null; ROSE (Training-free) supplies $\mathcal{M}$ in the prompt at inference only, so a VLM trained without it must read it zero-shot; ROSE (Fine-tuning) additionally fine-tunes the VLM on DSR-Train examples that carry $\mathcal{M}$, then supplies $\mathcal{M}$ identically at test time. Unless otherwise stated, all DSR-Bench training/generalization results in §4.2–4.2 use this DSR backbone.

LenCtrl is near DSR (58.9 → 59.5, +0.6), so token length alone explains little; ROSE (Training-free) improves to 62.3 (+3.4 over DSR), and **ROSE (Fine-tuning)** reaches **67.5**, establishing a new DSR-Bench state of the art (+8.6 vs. DSR, +5.2 vs. ROSE (Training-free)). The +5.2 gap between ROSE (Training-free) and ROSE (Fine-tuning) shows that 4DSGs are not merely inference-time prompts: with 4DSG context during training, the VLM learns to map trajectory and velocity tokens to motion reasoning. The largest gains are on direction subtasks (Abs. Dir 73.8→83.7, Rel. Dir 76.1→84.1), while Direction Prediction

Table 1: DSR-Bench evaluation with token-matched controls. Accuracy (%) across all 13 DSR subtask types. Abs/Rel: absolute/relative viewpoint. Subtasks: Distance (Dis), Direction (Dir), Orientation (Ori), Speed (Spd), Speed Comparison (SC), Direction Prediction (DP), Non-Template (N-T). All baseline numbers from [57]. **Bold**: best; underline: second best.

	Absolute						Relative							
Method	Dis	Dir	Ori	Spd	SC	DP	Dis	Dir	Ori	Spd	SC	DP	N-T	Avg↑
GPT-4o [17]	18.8	29.2	26.8	29.7	21.5	26.2	24.1	23.8	17.2	32.3	22.8	24.4	34.7	26.4
GPT-5 [35]	21.1	41.5	48.7	34.5	33.3	34.7	17.2	44.3	41.9	21.2	25.0	30.9	26.7	30.8
Gemini-2.5-Pro [11]	20.0	44.6	53.6	27.3	38.7	30.5	23.2	32.9	43.2	17.1	28.5	27.9	34.3	31.7
Qwen2.5-VL-7B [1]	18.8	15.3	14.6	42.8	29.0	19.4	31.8	19.3	11.1	22.2	19.2	20.2	30.1	23.5
VG-LLM [55]	55.2	32.3	58.5	57.1	51.6	32.2	56.0	36.3	32.0	30.3	32.1	29.1	27.9	38.4
DSR [57]	87.0	73.8	84.1	73.8	72.0	35.5	75.8	76.1	77.7	60.6	37.1	35.1	46.4	58.9
DSR + LenCtrl	86.7	74.2	84.4	73.6	71.5	35.8	75.9	76.9	78.0	60.3	36.8	35.3	46.8	59.5
ROSE (Training-free)	89.2	78.3	84.0	75.4	75.8	37.6	76.4	79.5	76.7	58.1	38.0	35.6	51.3	62.3
ROSE (Fine-tuning)	**91.3**	**83.7**	**85.6**	**79.9**	**79.5**	**38.6**	**82.6**	**84.1**	**78.1**	**68.3**	**40.3**	**39.1**	**54.9**	**67.5**

Table 2: Held-out subtask generalization on DSR-Bench. Held-out Avg is accuracy on the removed family; Overall Avg is full-test accuracy; Δ compares to "ROSE (Training-free)" under the same split.

Method	**Held-out Avg↑**	**Overall Avg↑**	Δ vs Training-free
DSR baseline (12-family FT split, no 4DSG)	22.9	58.9	–
ROSE (Training-free, 12-family FT split)	38.7	62.3	0.0
ROSE (Fine-tuning, 12-family FT split)	45.4	67.5	5.2

(DP) improves more modestly (35.5→38.6), consistent with DP requiring forward temporal prediction beyond directly observed history tokens.

Held-out Generalization on DSR-Bench. Beyond per-subtask gains, we next test whether fine-tuning with 4DSG as training-time context improves *generalization*, not just seen-pattern fitting.

Held-out Subtask Generalization (within DSR). Table 2 reports one explicit held-out split: we hold out one task family from DSR-Train and fine-tune on the remaining 12 families. We report **Held-out Avg** (the held-out family only) and **Overall Avg** (full DSR-Bench test set) to measure transfer beyond seen templates.

Held-out accuracy rises from 22.9 to 38.7 with ROSE (Training-free), and to 45.4 with ROSE (Fine-tuning), while overall accuracy also increases (62.3 → 67.5), supporting genuine transfer rather than template memorization.

VLM4D Evaluation. Raw-video VLMs mostly cluster at 50–62% weighted overall; GPT-5 is the only raw-video outlier at 72.3%. ROSE reaches 70.1%, 17.1 points above the larger Qwen2.5-VL-72B raw-video baseline (53.0) and 3.7 points below SNOW (73.8). Since both ROSE and SNOW are 4DSG pipelines (ROSE from monocular video, SNOW from RGB plus point clouds), this overall

Table 3: VLM4D evaluation. Accuracy (%) for egocentric (Ego-C.), exocentric (Exo-C.), directional (Direct.), and false-positive (FP) reasoning across frontier VLMs on raw video and 4DSG-based systems. Reported Avg./Overall scores follow benchmark weighted averages across categories; the two Avg. columns are over (Ego-C., Exo-C.) and (Direct., FP). **Bold**: best; underline: second best.

Model	Ego-C.↑	Exo-C.↑	Avg.↑	Direct.↑	FP↑	Avg.↑	Overall↑
GPT-5 [35]	**76.0**	72.5	**73.6**	68.8	64.4	68.3	72.3
Gemini-2.5-Pro [11]	64.6	62.9	63.5	54.8	80.0	57.3	62.0
Claude-Sonnet-4	52.6	52.1	52.2	44.0	86.7	48.3	51.3
Llama-4-Maverick-17B	52.6	54.3	53.8	53.3	51.1	53.0	53.6
Llama-4-Scout-17B	48.6	56.2	53.7	53.3	75.6	55.5	54.1
Qwen2.5-VL-72B [1]	54.3	52.5	53.1	49.5	80.0	52.6	53.0
InternVideo2.5-8B	57.2	50.5	52.7	44.3	46.7	44.5	50.7
SNOW [36]	73.0	**72.8**	72.9	71.8	77.9	76.5	**73.8**
ROSE (Ours)	72.8	63.6	66.6	**79.8**	**91.9**	**80.9**	70.1

Table 4: RoboSpatial-Home grounded VQA evaluation. Scores (%) are reported for Spatial Configuration (Cfg.), Spatial Context (Ctxt.), Spatial Compatibility (Cpt.), and their weighted mean (Avg.). "+RS" denotes models fine-tuned on RoboSpatial. **Bold**: best; underline: second best. All baseline numbers from [36].

Method	Cfg.↑	Ctxt.↑	Cpt.↑	Avg.↑
VILA +RS	65.9	15.6	78.0	53.2
LLaVA-NeXT +RS	78.9	19.7	80.1	59.6
SpaceLLaVA +RS	71.6	13.1	72.4	52.4
RoboPoint +RS	78.0	31.1	**81.0**	63.4
3D-LLM +RS	55.2	8.2	52.3	37.6
LEO +RS	64.2	10.0	57.1	43.8
GPT-4o	77.2	5.7	58.1	47.0
SNOW [36]	**84.55**	54.92	78.10	**72.29**
ROSE (Ours)	80.5	**70.5**	66.1	70.6

gap is modest; ROSE already exceeds SNOW on direction (79.8 vs. 71.8) and false-positive detection (91.9 vs. 77.9), while exocentric queries remain the main weakness (63.6 vs. 72.8), consistent with their stronger dependence on camera-motion and external-frame alignment under monocular estimation.

RoboSpatial-Home Evaluation. On RoboSpatial-Home, ROSE achieves the best context score (70.5, +15.6 over SNOW's 54.9) and the second-best weighted average (70.6 vs. 72.3). This context gain is consistent with explicit world-frame trajectory and geometry tokens, which make cross-object spatial context more directly queryable than text-only identity descriptions. Configuration is also second-best (80.5). Compatibility (66.1) lags dedicated +RS fine-tuned methods, which we attribute to ROSE not being fine-tuned on RoboSpatial data and this category's stronger dependence on domain-specific visual priors than on geometric grounding.

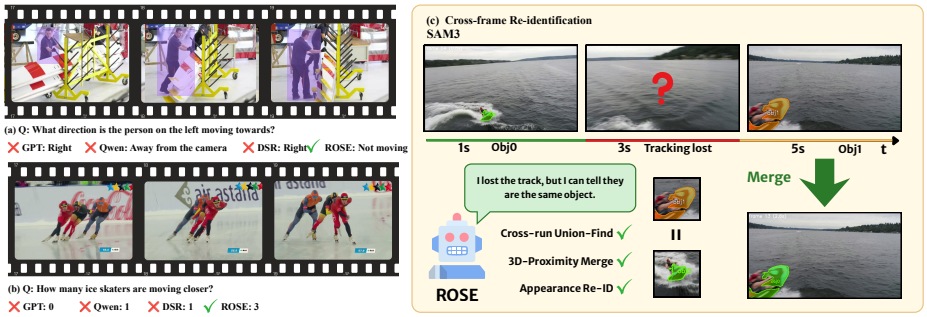

Fig. 4: Qualitative examples: dynamic QA and cross-frame re-identification. Left: on two dynamic spatial reasoning cases, ROSE produces correct answers while baseline VLMs fail. Right: ROSE recovers a temporarily occluded object track and re-identifies it under the correct persistent ID via appearance similarity and temporal consistency constraints.

Table 5: Average inference time on DSR-Bench (DSR). without-4DSG: keyframes only. with-4DSG: same keyframes + 4DSG. #TxtTok and #Images are total prompt text tokens and image inputs; T_p is prompt-time latency; ΔT_p is the added latency from including 4DSG. Overhead is the multiplicative factor relative to without-4DSG. Latency is averaged over repeated runs under the same hardware and decoding settings; 4DSG build time is excluded.

Setting	#TxtTok	#Images	T_p (s)↓	ΔT_p (s)↓	Overhead↓
Qwen2.5-VL-32B (without-4DSG)	4.5k	32	3.95	0.00	-
Qwen2.5-VL-32B (with-4DSG)	14.3k	32 + 20 anchors	12.45	8.50	3.15x

Case Study. In Fig. 4, we show representative dynamic cases: ROSE handles continuously moving targets for QA and maintains identity consistency under temporary tracking loss through cross-frame re-identification and merge.

4.3 Inference Speed: SG vs no-SG

This experiment isolates the cost of attaching the 4DSG to the VLM prompt, which is separate from 4DSG construction: construction runs at 10 Hz (§3.5) and its build time is excluded here. We report a single operating point, DSR-Bench with Qwen2.5-VL-32B at N_{kf}=32 keyframes, where the full serialized $\mathcal{M}$ adds about 20 anchor crops (one per tracked object) and ∼10k text tokens on top of the keyframe-only prompt. The two rows share backbone, keyframes, and decoding settings and differ only in whether $\mathcal{M}$ is appended; T_p is the one-time prompt-time (prefill) latency per query.

At this operating point the 4DSG adds 8.50 s of prefill (a 3.15× increase), dominated by the anchor crops, which the vision encoder expands into image tokens, rather than by the additional text. The cost is incurred once per query at prefill and leaves the 10 Hz construction rate unaffected; it scales with the object

Table 6: Ablation: representation dimensionality (VLM4D, balanced 200-question subset, Qwen2.5-VL-32B). Rows progressively add structure from raw video to full 4DSG; (c) vs. (d) isolates object-only masked anchors vs. context-preserving bounding-box anchors.

	Configuration	Ego-C↑	Exo-C↑	Direct↑	FP↑	Overall↑
(a)	VLM only (raw frames)	40	48	56.36	53.33	49.5
(b)	VLM + 2D anchors (no depth/traj.)	46	48	58.2	66	56
(c)	VLM + 4DSG (masked crops)	56	48	65	91.1	64.5
(d)	VLM + 4DSG (Ours)	**70**	**58**	**76.4**	**91.1**	**73.5**

and keyframe budget, decreasing as K or N_{kf} is reduced. The accompanying accuracy gain is not attributable to prompt length, since the token-matched LenCtrl control (Table 1) recovers a negligible fraction of it. This regime suits accuracy-critical use such as offline evaluation or deliberative planning rather than latency-constrained control loops.

4.4 Ablation Study

Representation Dimensionality. All variants use Qwen2.5-VL-32B on a balanced 200-question VLM4D subset (100 real + 100 synthetic). We compare four inputs: (a) raw video only, (b) +2D anchors (no depth/trajectory), (c) +4DSG with *masked crops* (object-only anchor images masked to the segmentation region), and (d) +4DSG with *contextual crops* (bounding-box anchors with surrounding context, ours). Table 6 shows a monotonic gain 49.5→56.0→64.5→73.5. Notably, adding 3D geometry and trajectories ((b)→(c)) raises FP from 66.0 to 91.1, consistent with explicit world-frame trajectories helping the model reject false object-presence hypotheses. The final step (c)→(d) adds +9.0 overall, mainly on Ego-C (+14) and Direct (+11.4), while FP stays unchanged (91.1), indicating that added visual context primarily improves viewpoint/direction disambiguation once existence has already been largely resolved by the 4DSG geometry-motion tokens.

4.5 Failure Case Analysis

To characterize where ROSE breaks, not only where it succeeds, we examine two representative failures (Fig. 5). The first is a granularity limit: ROSE reasons at the object level and cannot resolve sub-object parts, so questions hinging on part-level detail, such as counting hand contacts, exceed what the current representation supports. The second is an open-world perception limit: under heavy clutter, occlusion, and low contrast, the segmentation front-end misses or loses objects, and those errors carry through to the 4DSG and the final answer. These are genuine shortcomings rather than isolated mistakes, concentrated in fine-grained and open-world perception, which we revisit in the limitations below.

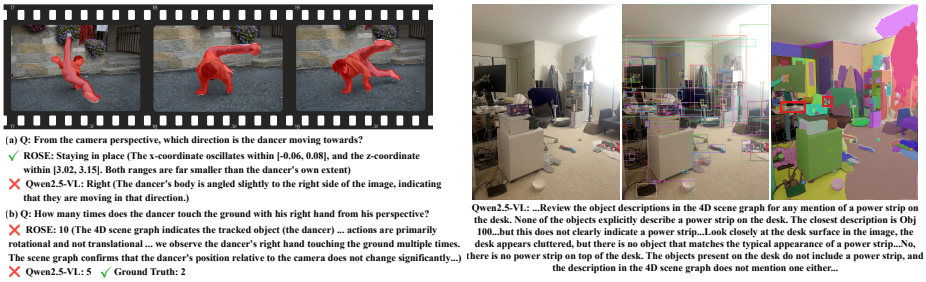

Fig. 5: Failure cases. Left: on *breakdance-flare*, SAM3 yields a single whole-body mask, giving body-level tracking but no hand granularity to count contacts. Right: in heavy clutter the 4DSG drops three objects (red boxes) as FastSAM misses the desk surface and shoes and SAM3 loses the low-contrast power strip.

5 Conclusion

Our main finding is that dynamic 4DSGs are most effective when used as structured training-time context, not only as inference-time prompts: fine-tuning with ROSE 4DSGs yields state-of-the-art DSR-Bench performance and broad gains across dynamic reasoning subtasks. VISTA addresses the representation gap by separating appearance from geometry/motion in a modality-aligned node design, delegating visual grounding to anchor crops and spatio-temporal attributes to structured text tokens. We demonstrate that such representations can be constructed from monocular video on a single GPU in real time, with open-world object discovery and persistent tracking. Across VLM4D, RoboSpatial-Home, and DSR-Bench, we observe consistent gains in direction/context-sensitive metrics, indicating that the improvements hold across three independently built benchmarks. We expect these findings to motivate further investigation into 4D scene graphs as structured supervision signals for dynamic reasoning, and their deployment in single-GPU real-time settings.

Limitations and Future Work. (i) Monocular geometry remains less scale-accurate than LiDAR, which can affect fine-grained outdoor metric reasoning; stronger scale-consistent geometry models are a natural next step. (ii) Long-absence identity recovery is still challenging, and more robust re-identification would improve long-horizon scene consistency. (iii) Perception coverage is a central limit: heavy clutter, occlusion, and low contrast make the front-end drop objects, and object-level reasoning leaves part-level queries out of reach, so finer-grained and more robust open-world segmentation, together with better efficiency-accuracy balancing in dense scenes, is a key direction. (iv) Extending ROSE from staged real-time processing to fully online frame-by-frame updates, together with lightweight and quantized VLM integration, points toward broader single-device deployment.

References

1. Bai, S., qin Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. ArXiv **abs/2502.13923** (2025), https://api.semanticscholar.org/CorpusID:276449796
2. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE international conference on image processing (ICIP). pp. 3464–3468. Ieee (2016)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
4. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024)
7. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
8. Cheng, H.K., Oh, S.W., Price, B., Schwing, A., Lee, J.Y.: Tracking anything with decoupled video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1316–1326 (2023)
9. Cheng, S., Fang, K., Yu, Y., Zhou, S., Li, B., Tian, Y., Li, T., Han, L., Liu, Y.: Videgothink: Assessing egocentric video understanding capabilities for embodied ai. arXiv preprint arXiv:2410.11623 (2024)
10. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Feng, B., Lai, Z., Li, S., Wang, Z., Wang, S., Huang, P., Cao, M.: Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? arXiv preprint arXiv:2505.14321 (2025)
13. Gorlo, N., Schmid, L., Carlone, L.: Describe anything anywhere at any moment. arXiv preprint arXiv:2512.00565 (2025)
14. Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K.M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., et al.: Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 5021–5028. IEEE (2024)

15. Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C.: 3d-llm: Injecting the 3d world into large language models. Advances in Neural Information Processing Systems **36**, 20482–20494 (2023)
16. Huang, H., Chen, Y., Wang, Z., Huang, R., Xu, R., Wang, T., Liu, L., Cheng, X., Zhao, Y., Pang, J., et al.: Chat-scene: Bridging 3d scene and large language models with object identifiers. arXiv preprint arXiv:2312.08168 (2023)
17. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
18. Ji, B., Agrawal, S., Tang, Q., Wu, Y.: Enhancing spatial reasoning in vision-language models via chain-of-thought prompting and reinforcement learning. arXiv preprint arXiv:2507.13362 (2025)
19. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Image retrieval using scene graphs. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3668–3678 (2015). https://doi.org/10.1109/CVPR.2015.7298990
20. Ju, Y., Liang, Y., Wang, Y.J., Gireesh, N., Ju, Y., Lee, S., Gu, Q., Hsieh, E., Huang, F., Sreenath, K.: Momagraph: State-aware unified scene graphs with vision-language model for embodied task planning. arXiv preprint arXiv:2512.16909 (2025)
21. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. arXiv preprint arXiv:2509.13414 (2025)
22. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
23. Li, S., Ke, L., Danelljan, M., Piccinelli, L., Segu, M., Van Gool, L., Yu, F.: Matching anything by segmenting anything. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18963–18973 (2024)
24. Lian, L., Ding, Y., Ge, Y., Liu, S., Mao, H., Li, B., Pavone, M., Liu, M.Y., Darrell, T., Yala, A., et al.: Describe anything: Detailed localized image and video captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21766–21777 (2025)
25. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
26. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
27. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. arXiv preprint arXiv:2210.07474 (2022)
28. Neo, C., Ong, L., Torr, P., Geva, M., Krueger, D., Barez, F.: Towards interpreting visual information processing in vision-language models. ArXiv **abs/2410.07149** (2024), https://api.semanticscholar.org/CorpusID:273233293
29. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 815–824 (2023)

30. Qian, T., Chen, J., Zhuo, L., Jiao, Y., Jiang, Y.G.: Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4542–4550 (2024)
31. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence **44**(3), 1623–1637 (2020)
32. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
33. Schmid, L., Abate, M., Chang, Y., Carlone, L.: Khronos: A unified approach for spatio-temporal metric-semantic slam in dynamic environments. arXiv preprint arXiv:2402.13817 (2024)
34. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016). https://doi.org/10.1109/CVPR.2016.445
35. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
36. Sohn, T.S., Dillitzer, M., Corso, J.J., Sax, E.: Snow: Spatio-temporal scene understanding with world knowledge for open-world embodied reasoning. arXiv preprint arXiv:2512.16461 (2025)
37. Song, C.H., Blukis, V., Tremblay, J., Tyree, S., Su, Y., Birchfield, S.: Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15768–15780 (2025)
38. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. Advances in neural information processing systems **34**, 16558–16569 (2021)
39. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
40. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(4), 376–380 (1991). https://doi.org/10.1109/34.88573
41. Vizzo, I., Guadagnino, T., Mersch, B., Wiesmann, L., Behley, J., Stachniss, C.: Kiss-icp: In defense of point-to-point icp–simple, accurate, and robust registration if done the right way. IEEE Robotics and Automation Letters **8**(2), 1029–1036 (2023)
42. Wald, J., Dhamo, H., Navab, N., Tombari, F.: Learning 3d semantic scene graphs from 3d indoor reconstructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3961–3970 (2020)
43. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., et al.: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19757–19767 (2024)
44. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video

understanding. In: European conference on computer vision. pp. 396–416. Springer (2024)
45. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)
46. Wu, S., Fei, H., Yang, J., Li, X., Li, J., Zhang, H., Chua, T.s.: Learning 4d panoptic scene graph generation from rich 2d visual scene. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24539–24549 (2025)
47. Wu, S.C., Tateno, K., Navab, N., Tombari, F.: Incremental 3d semantic scene graph prediction from rgb sequences. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5064–5074 (2023)
48. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. arXiv preprint arXiv:2310.11441 (2023)
49. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
50. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
51. Yang, S., Liu, J., Zhang, R., Pan, M., Guo, Z., Li, X., Chen, Z., Gao, P., Li, H., Guo, Y., et al.: Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9247–9255 (2025)
52. Zhang, H., Li, Z., Liu, J.: Scenellm: Implicit language reasoning in llm for dynamic scene graph generation. Pattern Recognition **170**, 111992 (2026)
53. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
54. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J.: Fast segment anything. arXiv preprint arXiv:2306.12156 (2023)
55. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. arXiv preprint arXiv:2505.24625 (2025)
56. Zhou, H., Lee, G.H.: Uni4d-llm: A unified spatiotemporal-aware vlm for 4d understanding and generation. arXiv preprint arXiv:2509.23828 (2025)
57. Zhou, S., Chen, Y., Ge, Y., Huang, W., Lin, J., Shan, Y., Qi, X.: Learning to reason in 4d: Dynamic spatial understanding for vision language models. arXiv preprint arXiv:2512.20557 (2025)
58. 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. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8600–8612 (2025)