PROS4D: MODELING EVOLVING 4D WORLDS IN LATENT SPACE FOR DYNAMIC SPATIAL REASONING

Anonymous authors
Paper under double-blind review



Figure 1: **Pros4D** supports observed-dynamics understanding with latent observed states and future-dynamics prediction through autoregressive rollout of latent future states.

ABSTRACT

Dynamic spatial reasoning from monocular video enables vision-language models (VLMs) to understand how 3D environments evolve over time, supporting physical world understanding and embodied intelligence. Existing efforts primarily address observed-dynamics understanding, whereas agents in dynamic environments also require future-dynamics prediction beyond the observation boundary. Moreover, external geometric processing and textual descriptions can increase inference complexity or lose information when representing continuous dynamics. These limitations motivate a shared internal state that preserves observed dynamics and remains predictive beyond the input video. Accordingly, we present **Pros4D**, a framework for modeling evolving 4D worlds in latent space that unifies retrospective understanding of observed 4D dynamics with prospective reasoning for future-dynamics prediction. Specifically, **Pros4D-Engine** derives geometry-grounded supervision and constructs paired examples for both tasks, while **Pros4D-Bench** evaluates them across eight spatial task categories spanning dynamic cameras, objects, and scenes. At the model level, Pros4D compresses observed dynamics into latent observed states and autoregressively rolls them forward into latent future states. **4D Latent Alignment (4DLA)** grounds these states through complementary visual-semantic and geometric supervision, while joint language supervision connects them to task reasoning. Experiments across benchmarks show consistent improvements over strong baselines in both capabilities.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Dynamic spatial reasoning aims to understand how cameras, objects, and their spatial relations evolve over time in a 3D environment (Huang et al., 2026b; Zhou et al., 2025a). From monocular video, this requires an agent to disentangle observer and object motion, preserve spatial relations across time, and reason about how these dynamics may continue beyond the observed segment (Morbitzer et al., 2026; Chen et al., 2026b). Such capability is essential for connecting visual perception with embodied decision making, where an agent must not only understand what has happened but also predict what may happen next (Hou et al., 2026; Huang et al., 2026c; Hafner et al., 2023). Although vision-language models (VLMs) have made substantial progress in spatial understanding (Wu et al., 2025a; Yin et al., 2026; Yang et al., 2025a; Gao et al., 2026; Ma et al., 2025b), extending this ability from observed dynamics to subsequent 4D evolution remains an open challenge.

Existing 4D reasoning methods primarily focus on observed-dynamics understanding. Explicit approaches augment VLMs with geometric modules or tool-mediated computation (Chen et al., 2026a; Cho et al., 2026; Dai et al., 2026), or express intermediate dynamics through text (Yin et al., 2026; Hua et al., 2026). Geometric augmentation adds inference-time components and perception dependencies, while verbalizing dynamics can lose fine-grained motion and geometric information. More importantly, supervision centered on observed evidence does not require a state that continues evolving once observation ends. Dynamic spatial reasoning therefore calls for an internal representation that preserves observed dynamics while remaining predictive beyond the input video.

An implicit latent state provides a natural basis for such a representation. As an intuitive analogy, people integrate successive observations into an internal mental model to interpret ongoing motion and imagine what may happen next. Continuous latent representations can similarly preserve visual information and support state updates without verbalizing each change. Prior work explores latent visual computation for fine-grained perception (Li et al., 2026; Shen & Elhoseiny, 2026) and persistent spatial memory (Liu et al., 2026; Wang et al., 2026b). Video world models further combine latent dynamics forecasting with semantic guidance, linking motion cues to context useful for prediction (Zhang et al., 2026a). These advances motivate shared states that preserve observed 4D dynamics and support geometrically grounded evolution beyond the observation boundary.

Following this principle, we present **Pros4D**, a framework for modeling evolving 4D worlds in latent space that unifies retrospective understanding of observed 4D dynamics with prospective reasoning for future-dynamics prediction. To support both capabilities, we develop **Pros4D-Engine**, an automated data engine that derives motion and spatial supervision from estimated 3D geometry and constructs paired examples for the two tasks. We further introduce **Pros4D-Bench**, which evaluates these tasks across eight spatial categories spanning dynamic cameras, objects, and scenes.

Within this framework, Pros4D performs **4D Latent State Reasoning (4D-LSR)** by compressing observed dynamics into continuous latent observed states and autoregressively rolling them forward into latent future states for future-dynamics prediction. **4D Latent Alignment (4DLA)** grounds both types of state through complementary visual-semantic and geometric supervision from observed and held-out frames, while joint language supervision connects these states to task reasoning. At inference, 4D-LSR operates entirely within the VLM using only the observed video prefix, without external geometric tool calls during question answering. Extensive experiments across multiple dynamic spatial reasoning benchmarks demonstrate consistent improvements over strong baselines in both retrospective understanding and prospective reasoning across 3B and 8B model scales.

Our contributions are summarized as follows.

- We propose **Pros4D**, a unified latent reasoning framework that supports retrospective understanding and prospective reasoning through a shared evolving 4D state in one VLM.
- We develop **Pros4D-Engine**, an automated data engine for paired supervision of observed-dynamics understanding and future-dynamics prediction, together with **Pros4D-Bench** for evaluating both tasks across eight spatial categories under a shared observation boundary.
- We introduce **4D Latent State Reasoning (4D-LSR)**, which reasons through latent observed and future states, and **4D Latent Alignment (4DLA)**, which grounds these evolving states with complementary visual-semantic, geometric, and language supervision.

2 Related Work

2.1 Latent Space Reasoning

Latent reasoning studies how intermediate computation can proceed without verbalizing every step. Textual chains make reasoning explicit but bind it to discrete language (Wei et al., 2022). Work on latent reasoning relaxes this constraint through pause tokens, progressive internalization of reasoning steps, and continuous hidden states (Goyal et al., 2024; Deng et al., 2024; Hao et al., 2024). These approaches separate computation from its verbal expression. For multimodal problems, intermediate representations must also retain task-relevant visual evidence. This shifts the focus from how many reasoning steps to generate toward what information those steps preserve across modalities.

Visual grounding supervises latent tokens with visual features or aligns intermediate states with perceptual targets (Yang et al., 2026; Li et al., 2026; Jeon et al., 2026). Predictive embedding objectives distill expert multimodal trajectories without explicit tool calls at inference (Adhikari & Lapata, 2026). Latent imagery also supports dynamic spatial understanding through visual-semantic supervision of camera and object motion (Chen et al., 2026b). These studies motivate grounding intermediate computation in relevant evidence. Perceptual similarity alone, however, does not constrain the geometry and temporal correspondence needed to describe scene evolution. We therefore learn ordered states that preserve geometry and temporal changes through joint visual-semantic and geometric alignment of observed and future states within a shared representation of evolving scenes.

2.2 Visual-Spatial Understanding

Spatial reasoning requires consistent object locations and relations across viewpoints. Existing work improves this capability through geometric supervision, structured coordinates, and reconstructed 3D features (Chen et al., 2024; Liu et al., 2025a; Fan et al., 2026). Complementary approaches retain spatial evidence in internal representations or persistent memory (Yin et al., 2025; Liu et al., 2026), while benchmark evaluations test whether that information supports relational and quantitative reasoning (Yang et al., 2025b; Ma et al., 2025a). The common requirement is to preserve scene structure beyond appearance. When cameras and objects move, this requirement becomes temporal: a useful representation must maintain entity identity across frames, preserve relations in a consistent spatial reference, and connect successive observations to changes in the underlying scene.

Dynamic spatial benchmarks expose challenges of viewpoint change, object motion, and evolving relations (Zhou et al., 2025b; Zhang et al., 2025c; Huang et al., 2026b). Geometry-guided reasoning and perceptual distillation strengthen representations of observed dynamics (Zhou et al., 2025a; Yang et al., 2025a), while training on synthesized future views can transfer motion-conditioned spatial knowledge into VLMs (Zhang et al., 2026b). These directions motivate learning spatial evolution within the model. Our focus is a shared sequence of geometrically aligned latent states that supports both understanding the observed interval and predicting its continuation from the same video prefix, with future evidence used for training supervision but excluded from inference.

3 Method

Reasoning about a dynamic scene requires a representation that captures observed motion and remains predictive beyond the observation boundary. We propose **Pros4D** to meet this need by unifying retrospective understanding and prospective reasoning through evolving 4D latent states, as illustrated in Figure 1. The framework couples **Pros4D-Engine**, which constructs geometry-grounded questions and traces for **4D Latent State Reasoning (4D-LSR)** in Section 3.1, with **4D Latent Alignment (4DLA)**, which grounds latent observed and future states through complementary visual-semantic and geometric supervision in Section 3.2.

3.1 Pros4D-Engine

Video preprocessing. We use raw SpatialVID videos (Wang et al., 2026a) without their annotations. For sampled frames $V = \{I_t\}_{t=1}^{T}$, boundary b separates the observed video prefix $V_{\text{obs}} = \{I_t\}_{t \le b}$ from the future continuation $V_{\text{fut}} = \{I_t\}_{t > b}$. Both tasks receive only V_{obs}; their supervision comes from the respective intervals. Appendix A details sampling.

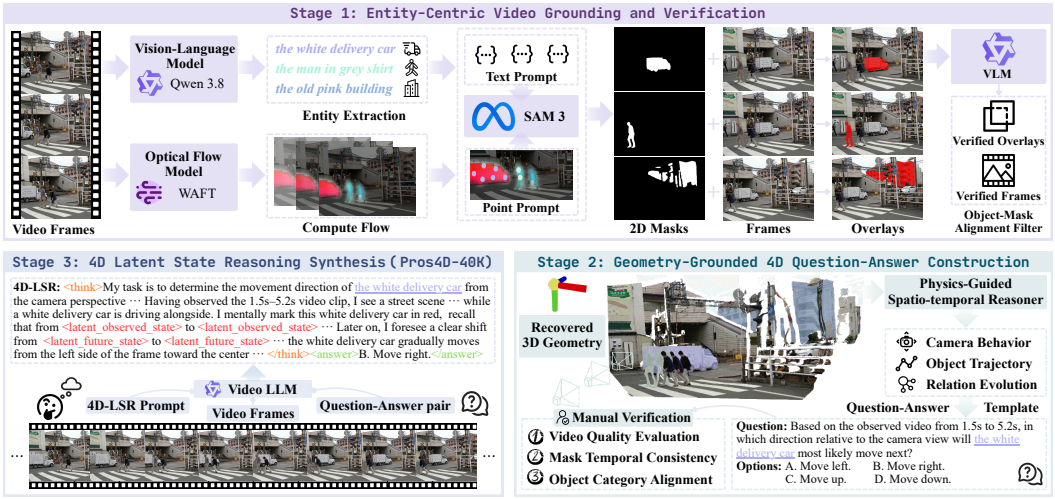

Figure 2: **Overview of Pros4D-Engine**. Given raw videos, Stage 1 grounds and verifies entities to obtain reliable object-mask alignments. Stage 2 recovers 3D geometry and constructs question-answer pairs for camera motion, object trajectories, and relation evolution. Stage 3 synthesizes 4D latent-state reasoning traces, pairing textual reasoning with observed- and future-state supervision.

Entity-centric video grounding and verification. Reliable state supervision requires consistent entity identity across frames. Qwen3.8 (Qwen Team, 2026) identifies moving objects and static landmarks, while WAFT (Wang & Deng, 2026) supplies optical-flow cues for point prompts. Together with entity text prompts, these cues guide SAM3 (Carion et al., 2026) to track masks $\{M_t\}$. The overlay $J_t = I_t + \alpha M_t \odot (\boldsymbol{\gamma} - I_t)$ highlights the target with color $\boldsymbol{\gamma}$ and opacity α, where $\odot$ denotes element-wise multiplication. This preserves the surrounding scene while identifying the entity whose dynamics should be modeled. An object-mask alignment filter retains verified frames and overlays, establishing entity correspondence for geometric reasoning (Figure 2).

Geometry-grounded 4D question-answer construction. DA3 (Lin et al., 2025a) recovers geometry from the verified frames. Masked depth is mapped to a shared 3D frame; object motion directions follow the camera view. Camera centers $\mathbf{c}_t$, object centroids $\mathbf{p}_t$, and landmark positions $\mathbf{s}_t$ anchor a physics-guided reasoner for camera behavior, object motion, and changing relations. For example, Scene Spatial Relationship follows $d_t = \|\mathbf{p}_t - \mathbf{s}_t\|_2$ to distinguish approach, separation, and transitions between them. Robust temporal statistics apply the same eight task definitions to both intervals. The VLM verbalizes these labels rather than deciding answers. Questions and distractors are then manually checked for video quality, mask consistency, and entity correspondence. Appendix B gives the task rules and templates.

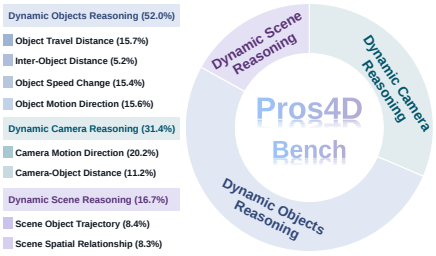

Figure 3: **Pros4D-Bench** task composition across dynamic camera, object, and scene reasoning categories.

4D latent state reasoning synthesis. Given a question q, its geometry-derived answer a, and the video, a VLM synthesizes a trace r within `<think>` and `<answer>` tags. Observed evidence uses `<latent_observed_state>`; prediction traces additionally use `<latent_future_state>` for the future continuation. Each placeholder corresponds to an ordered target overlay J_i, highlighting the queried entity. The trace thus specifies both what to infer and which entity-specific states should support each reasoning step. Checks on answer preservation and placeholder-to-image correspondence yield Pros4D-40K, whose samples $(V_{\text{obs}}, q, r, a, \{J_i\})$ couple the question, reasoning trace, and answer with ordered state targets for both temporal tasks.

Pros4D-Bench. Pros4D-Bench evaluates both capabilities on videos separated from training by video identifier. Its categories span dynamic cameras, objects, and scenes (Figure 3). A shared

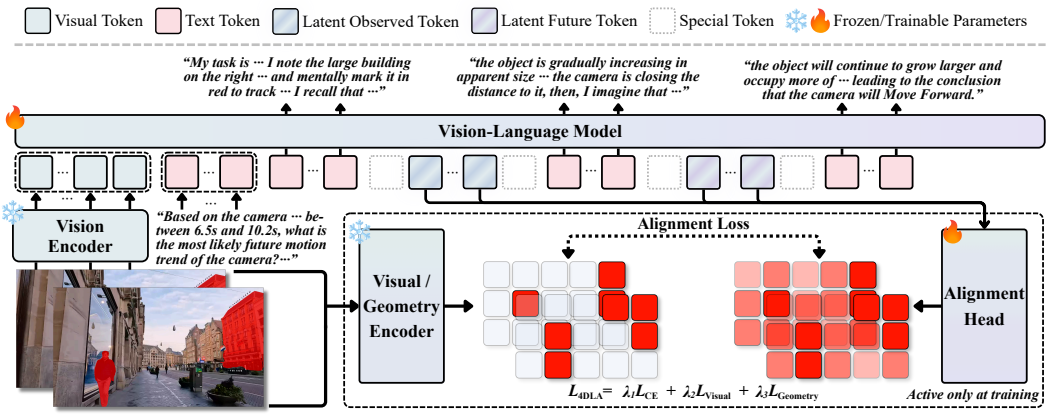

Figure 4: **4D Latent Alignment (4DLA)**. Visual-semantic and geometric supervision ground the interleaved observed and future states used by 4D-LSR. Alignment is used only during training.

observed video prefix tests interpretation and prediction under the same observation boundary. Appendix C compares benchmark coverage; Table 14 reports the task distribution.

3.2 Learning to Reason with Evolving 4D States

Latent state representation. 4D-LSR represents each state by K continuous hidden vectors (Figure 4). With temporal type $\tau \in \{\text{obs}, \text{fut}\}$, a *latent observed token* encodes observed evidence, while a *latent future token* represents predicted continuation. An ordered group forms a latent observed state or latent future state $\mathbf{Z}_i^\tau = [\mathbf{z}_{i,1}^\tau, \ldots, \mathbf{z}_{i,K}^\tau]$, replacing its placeholder in r with

$$B_i^\tau = \big(\texttt{<|latent_start|>}, \mathbf{Z}_i^\tau, \texttt{<|latent_end|>}\big), \tag{1}$$

where learnable boundary tokens delimit continuous computation. Both state types share this protocol; their temporal roles follow the trace and associated targets.

Visual-semantic alignment. To ground each state in its queried entity, the frozen native visual encoder f_{vis} encodes overlay J_i as patch features. Ordered mean pooling yields its target block

$$E_i = \text{Pool}_K\big(f_{\text{vis}}(J_i)\big) = [\mathbf{e}_{i,1}, \ldots, \mathbf{e}_{i,K}] \in \mathbb{R}^{K \times d}, \tag{2}$$

where d is the decoder hidden dimension. During teacher-forced training, continuous positions are filled with detached target embeddings. The preceding causal hidden state $\mathbf{h}_{i,k}$ predicts $\mathbf{e}_{i,k}$ through

$$\mathcal{L}_{\text{Visual}} = \frac{1}{N} \sum_{i,k} \big(1 - \cos\big(\mathbf{h}_{i,k}, \text{sg}(\mathbf{e}_{i,k})\big)\big), \tag{3}$$

where N counts valid latent positions and sg denotes stop-gradient. Supervising both state types embeds observed evidence and its continuation in a shared visual space.

Geometric alignment. To preserve spatial structure, we align prediction states H_i^ℓ at an intermediate layer ℓ with frozen Pi3X (Wang et al., 2026c) features. A trainable head g_ϕ maps each latent block to the grid resolution of the geometry encoder f_{geo},

$$\widehat{G}_i = g_\phi(H_i^\ell), \qquad G_i = f_{\text{geo}}(J_i), \qquad \mathcal{L}_{\text{Geometry}} = \frac{1}{M} \sum_{i=1}^{M} \mathcal{D}_g\big(\widehat{G}_i, \text{sg}(G_i)\big). \tag{4}$$

M counts target states. $\mathcal{D}_g$ constrains direction and magnitude through cosine distance and Smooth L1 penalties on values and log norms. Appendix A details the head and coefficients.

Joint training. Language supervision links states to reasoning and answers. Cross entropy $\mathcal{L}_{\text{CE}}$ supervises text and boundary tokens, excluding continuous positions. The joint objective is

$$\mathcal{L}_{\text{4DLA}} = \lambda_1 \mathcal{L}_{\text{CE}} + \lambda_2 \mathcal{L}_{\text{Visual}} + \lambda_3 \mathcal{L}_{\text{Geometry}}. \tag{5}$$

The decoder, token embeddings, and alignment head are jointly trained with both encoders frozen. Weights $\lambda_1, \lambda_2, \lambda_3$ balance reasoning with visual-semantic and geometric grounding.

Table 1: **Pros4D-Bench results (accuracy, %)**. Retrospective and Prospective evaluate observed-dynamics understanding and future-dynamics prediction. Eight task columns cover dynamic camera, object, and scene reasoning. Bold and underlined values mark the best and second best scores.

Methods	Avg.	Retrospective	Prospective	Cam. Dir.	Cam.-Obj. Dist.	Obj. Dir.	Inter-Obj. Dist.	Obj. Speed	Obj. Dist.	Scene Rel.	Scene Traj.
		Temporal		Dynamic Camera			Dynamic Objects			Dynamic Scene	
Spatial Understanding Models											
SpaceR-7B (Ouyang et al., 2025)	46.9	49.5	42.6	86.3	39.1	41.4	24.5	53.6	29.9	33.9	19.4
VST-7B-RL (Yang et al., 2025c)	41.1	44.2	35.8	74.9	37.9	32.6	25.2	45.8	23.3	31.9	23.0
Spatial-SSRL-7B (Liu et al., 2025b)	39.3	43.2	32.6	73.4	32.8	35.0	25.8	30.0	31.4	34.7	18.3
SpatialReasoner (Ma et al., 2025b)	39.4	41.3	36.1	55.0	35.5	30.5	33.5	58.7	22.2	31.5	31.3
SpatialLadder-3B (Li et al., 2025a)	42.7	45.7	37.6	70.8	33.4	25.8	27.1	65.2	25.6	31.0	30.6
MLLM-4D-8B (Yin et al., 2026)	51.7	59.3	38.7	92.1	**61.8**	24.3	43.2	65.9	30.1	**38.7**	24.6
4DThinker-3B (Chen et al., 2026b)	41.5	43.8	37.4	67.5	33.4	43.5	30.3	42.8	24.2	32.3	31.7
4D-RGPT-8B (Yang et al., 2025a)	42.1	46.2	35.2	81.5	27.8	21.1	32.9	59.0	27.5	29.8	19.8
Video Understanding Models											
LLaVA-OV-1.5-8B (An et al., 2025)	33.8	41.4	20.8	69.8	34.0	21.3	29.7	22.0	18.9	34.7	21.0
LLaVA-OV-2-8B (An et al., 2026)	41.9	43.3	39.4	45.7	32.8	42.0	38.7	66.5	31.1	36.7	26.2
VideoRefer-7B (Yuan et al., 2025)	27.6	33.6	17.3	60.4	19.7	16.2	18.7	7.6	25.8	31.5	22.2
GLM-4.1V-9B (Team et al., 2026)	51.0	56.2	42.1	92.4	23.3	42.9	29.0	60.7	42.8	34.3	30.6
InternVideo2.5-8B (Wang et al., 2025c)	42.8	51.3	28.1	90.9	21.8	23.7	21.9	52.1	30.9	28.2	22.6
InternVideo3-8B (Yan et al., 2026)	46.0	52.7	34.6	76.2	49.3	20.5	29.7	60.5	39.8	33.9	23.4
VideoLLaMA3-7B (Zhang et al., 2025a)	47.7	50.1	43.6	85.8	28.4	40.3	23.2	71.1	26.3	28.6	26.2
Training-free Models											
Spatial-Claw (Cho et al., 2026)	53.4	58.3	45.1	92.9	56.7	30.9	43.9	67.8	36.0	32.3	28.6
See&Trek (Li et al., 2025b)	45.0	48.3	39.4	86.5	20.9	25.8	32.9	67.8	23.5	37.5	26.6
DSR Suite (Zhou et al., 2025a)	46.6	50.6	39.8	84.7	29.3	33.5	31.0	66.1	30.3	29.0	24.2
DynTrace (Gao et al., 2026)	54.6	57.5	49.6	93.2	56.4	33.7	<u>44.5</u>	69.8	40.3	25.0	32.5
Different Base Models (Pros4D)											
Qwen2.5-VL-3B (Qwen Team, 2025a)	45.4	50.9	36.0	77.1	37.6	29.6	33.5	71.7	21.2	35.5	22.6
Pros4D-3B	<u>62.0</u>	<u>63.9</u>	<u>58.8</u>	**93.9**	53.7	<u>44.1</u>	37.4	**76.9**	<u>58.5</u>	37.5	<u>48.4</u>
Qwen3-VL-8B (Qwen Team, 2025b)	46.4	49.7	40.7	86.0	36.1	33.0	21.3	63.5	23.7	35.5	26.6
Pros4D-8B	**62.9**	**64.9**	**59.5**	<u>93.6</u>	<u>57.0</u>	**44.3**	**45.2**	<u>76.0</u>	**59.5**	<u>38.3</u>	**49.2**

Autoregressive inference. Given only (V_{obs}, q), text and latent states are generated causally. Continuous positions $t \in \mathcal{T}_{\text{lat}}$ reuse the preceding hidden state as input embedding $\mathbf{u}_t$,

$$\mathbf{u}_t = \begin{cases} \mathbf{h}_{t-1}, & t \in \mathcal{T}_{\text{lat}}, \\ \text{Embed}(y_t), & t \notin \mathcal{T}_{\text{lat}}, \end{cases} \tag{6}$$

where y_t is a discrete token; text generation resumes after K continuous steps. Latent observed states support observed-dynamics understanding, and latent future states extend reasoning beyond the observation boundary for future-dynamics prediction, without future frames or external geometry. Appendix D defines the state controls.

4 Experiments

Experimental setup. We evaluate Pros4D on seven benchmarks to assess its two reasoning capabilities and their transfer across tasks. Pros4D-Bench jointly evaluates observed-dynamics understanding and future-dynamics prediction, while Dyn-Bench (Huang et al., 2026b), VLM4D (Zhou et al., 2025b), R4D-Bench (Yang et al., 2025a), and STI-Bench (Li et al., 2025d) test dynamic spatial generalization. OpenX-VQA (Chen et al., 2025a) and STRIDE-QA-Bench (Ishihara et al., 2026) extend this evaluation to manipulation videos and to estimation and prediction in driving scenes. We report accuracy for multiple-choice QA and prescribed success rates for STRIDE-QA-Bench. Appendix C describes the datasets and metrics.

Baselines and implementation. We compare spatial VLMs, video models, and methods using external tools. Pros4D builds on Qwen2.5-VL-3B and Qwen3-VL-8B, with one epoch of training on Pros4D-40K, frozen visual encoders, and $K = 16$ tokens per state. The decoder and alignment head are optimized with 4DLA. Evaluation uses autoregressively generated states (Appendix D).

Table 2: **Dyn-Bench and VLM4D results (accuracy, %).** Columns report overall scores and relation or video subsets.

Methods	Dyn-Bench				VLM4D		
	Avg.	Inter-Obj.	Obj.-Scene	Cam.-Obj.	Avg.	Real	Syn.
Spatial Understanding Models							
SpaceR-7B (2025)	56.5	54.1	71.9	47.8	47.4	49.2	41.8
VST-7B-RL (2025c)	55.8	53.9	73.8	45.0	44.6	45.1	43.1
MLLM-4D-8B (2026)	56.5	49.6	64.8	55.5	59.1	59.3	58.4
4DThinker-3B (2026b)	46.5	43.8	58.5	40.2	38.0	38.7	35.9
4D-RGPT-8B (2025a)	48.1	46.5	63.5	39.0	56.3	54.9	60.9
Video Understanding Models							
LLaVA-OV-2-8B (2026)	56.9	53.4	72.7	48.7	57.0	59.2	50.6
GLM-4.1V-9B (2026)	56.0	53.5	68.7	49.2	54.1	55.9	48.5
InternVideo3-8B (2026)	59.3	55.8	**75.0**	51.2	56.9	54.9	63.1
VideoLLaMA3-7B (2025a)	53.7	47.3	71.1	46.3	49.8	55.6	32.1
Training-free Models							
Spatial-Claw (2026)	53.5	47.5	69.9	46.6	50.3	53.2	41.3
See&Trek (2025b)	51.5	51.1	68.2	40.7	47.8	44.6	57.5
DSR Suite (2025a)	55.3	52.7	73.9	44.7	48.4	48.6	47.6
Different Base Models (Pros4D)							
Qwen2.5-VL-3B (2025a)	50.7	48.9	69.1	39.7	43.9	47.3	33.4
Pros4D-3B	56.5	49.6	70.0	52.0	53.6	52.4	57.3
Qwen3-VL-8B (2025b)	61.0	57.0	74.4	54.7	59.0	58.1	61.6
Pros4D-8B	**65.6**	**58.2**	74.6	**64.5**	**68.2**	**66.0**	**75.0**

Table 3: **R4D-Bench** region-level results (accuracy, %).

Methods	Acc.
LLaVA-Video-7B (Zhang et al., 2025b)	39.7
SpatialReasoner (Ma et al., 2025b)	38.3
SpaceR-7B (Ouyang et al., 2025)	37.0
4D-RGPT-8B (Yang et al., 2025a)	42.2
ViLaSR-7B (Wu et al., 2025b)	39.6
Qwen3-VL-8B (Qwen Team, 2025b)	40.3
Pros4D-8B	**43.5**

Table 4: **STI-Bench** spatiotemporal results (accuracy, %).

Methods	Acc.
InternVL2.5-78B (Chen et al., 2025b)	38.5
VideoLLaMA3-7B (Zhang et al., 2025a)	35.0
SpaceR-7B (Ouyang et al., 2025)	35.1
VideoChat-R1 (Li et al., 2025c)	34.3
See&Trek (Li et al., 2025b)	31.2
DSR Suite (Zhou et al., 2025a)	35.2
Qwen2.5-VL-72B (Qwen Team, 2025a)	40.7
Qwen3-VL-8B (Qwen Team, 2025b)	40.5
Pros4D-8B	**43.6**

4.1 Retrospective Understanding and Prospective Reasoning

We first evaluate whether shared latent states support both interpretation and prediction. Table 1 shows gains of 15.2 and 18.8 points over the 8B backbone on observed-dynamics understanding and future-dynamics prediction, respectively. Pros4D-3B follows the same pattern, reaching 62.0 overall accuracy compared with 62.9 for Pros4D-8B. Both temporal tasks receive the same observed prefix, so the improvement on future questions reflects the ability to infer subsequent scene changes from the available visual and motion evidence. These gains across the observation boundary support our design because the shared representation both interprets observed dynamics and predicts how they continue. It therefore links visible evidence to scene evolution beyond the observed prefix.

Results across tasks reveal which spatial relations benefit from state modeling. Pros4D-8B gains 35.8 points on object travel distance and 22.6 on scene object trajectory over its backbone. Both require tracking queried entities while retaining a spatial reference for their motion, consistent with how 4D-LSR grounds states in the entities being queried. Appendix E illustrates how Pros4D connects changes in camera distance and object movement to the queried entities and their surroundings.

4.2 Generalization across Dynamic and Embodied Tasks

We next test transfer beyond Pros4D-Engine data. Table 2 reports gains of 4.6 and 9.2 points over the 8B backbone on Dyn-Bench and VLM4D, with overall improvements at the 3B scale as well. Dyn-Bench includes questions about relations between the camera and objects, which require interpreting motion relative to a changing viewpoint. Gains on both real and synthetic VLM4D videos further indicate transfer across diverse visual sources, scene configurations, and question formulations.

Table 5: **OpenX-VQA results (accuracy, %).** Parentheses give changes in percentage points over Qwen2.5-VL-3B. Spatial, goal, and interaction task abbreviations are defined in Appendix C.4.

Methods	Avg.	Spatial Reasoning				Goal Reasoning			Interaction Reasoning		
		RS	OS	SU	MV	TS-G	TS-S	TS-GL	AU	IP	TU
Qwen2.5-VL-3B	30.48	37.09	80.11	23.23	13.47	37.46	40.63	2.10	24.05	26.81	**46.94**
4D-RGPT-8B	29.70	45.20	42.70	18.60	17.70	30.40	38.10	19.90	18.90	32.80	34.00
MLLM-4D-8B	46.30	**70.30**	98.50	22.60	20.00	39.70	**70.50**	20.90	**37.90**	41.00	46.20
Pros4D-3B	**51.39** (+20.91)	47.09 (+10.00)	**100.00** (+19.89)	**30.09** (+6.86)	**29.51** (+16.04)	**39.80** (+2.34)	37.01 (−3.62)	**47.90** (+45.80)	30.76 (+6.71)	**79.41** (+52.60)	43.54 (−3.40)

Tables 3 and 4 test reasoning about regions and quantitative properties. Pros4D improves over its backbone by 3.2 points on R4D-Bench and 3.1 on STI-Bench, extending the benefits to these question formats. This transfer is consistent with geometric grounding that supplies spatial structure beyond the training tasks' answer vocabulary. Appendix C.3 gives complete Dyn-Bench and VLM4D results across methods and benchmark subsets, using the same metrics and reporting conventions.

Table 5 tests transfer to robot manipulation videos. Pros4D-3B improves average accuracy by 20.91 points over its backbone, with pronounced gains on questions about object states, goal locations, and interaction phases. These tasks relate scene content to the configuration and progress of an interaction, showing how reasoning with states transfers to videos of embodied tasks. Appendix C.4 defines the task abbreviations and the states, goals, and interactions queried in each task category.

Table 6: **STRIDE-QA-Bench results (success rate, %)**. Cells give backbone | Pros4D-3B scores with changes in percentage points. Time indices and localization metrics are defined in Appendix C.4. At t_1–t_3, LSR, distance, and direction share composite questions and success counts.

Time	LSR	Target Distance	Target Direction	Ego Speed	Target Speed
t_0	11.49 | 14.67 (+3.18)	33.25 | 43.03 (+9.78)	36.19 | 35.70 (-0.49)	29.58 | 42.30 (+12.72)	30.56 | 35.70 (+5.14)
t_1	41.81 | 44.74 (+2.93)	41.81 | 44.74 (+2.93)	41.81 | 44.74 (+2.93)	33.99 | 44.74 (+10.75)	33.99 | 38.63 (+4.64)
t_2	39.36 | 42.30 (+2.94)	39.36 | 42.30 (+2.94)	39.36 | 42.30 (+2.94)	33.25 | 44.50 (+11.25)	29.83 | 35.94 (+6.11)
t_3	34.23 | 39.85 (+5.62)	34.23 | 39.85 (+5.62)	34.23 | 39.85 (+5.62)	35.21 | 52.57 (+17.36)	29.58 | 34.23 (+4.65)

We use STRIDE-QA-Bench to examine prediction in driving scenes. Table 6 shows better estimates of ego and target speed at every evaluated time, together with improved future localization. At t_1–t_3, distance and direction are assessed from the same joint localization questions. The same answers succeed on both measures, so their individual success rates and joint success rate (LSR) are identical. Together with OpenX-VQA, these gains show that reasoning with evolving states transfers to both manipulation videos and driving scenes that require tracking changing spatial relations over time.

4.3 Ablation Studies

We conduct ablations on Pros4D-8B to examine state supervision and latent evolution. Obs. and Future denote accuracy on Pros4D-Bench's observed-dynamics understanding and future-dynamics prediction subsets. External Avg. weights accuracy on Dyn-Bench, VLM4D, R4D-Bench, and STI-Bench by their question counts. Appendix D lists these counts and defines the ablation settings.

Table 7: Supervision variants trained on the same data (accuracy, %). Text and latent supervision vary. Parentheses give gains over Raw-QA SFT.

Method	Visual-align.	Geometry-align.	Pros4D Avg.	Obs. Acc.	Future Acc.	External Avg.
Raw-QA SFT	×	×	53.4 (0.0)	55.7	49.5	55.8
Textual-CoT SFT	×	×	58.1 (+4.7)	60.0	54.9	57.2
4DLA	✓	✓	**62.9** (+9.5)	**64.9**	**59.5**	**59.7**

Effect of state supervision. Table 7 compares three supervision strategies on the same training data. Textual-CoT SFT improves overall accuracy by 4.7 points over Raw-QA SFT. Combining language supervision with visual-semantic and geometric alignment, 4DLA adds 4.8 points overall, with gains of 4.9 on observed-dynamics understanding, 4.6 on future-dynamics prediction, and 2.5 on the external aggregate. These gains support supervising latent states alongside textual reasoning. Figure 5(b) separates the contributions of the two alignment objectives.

Complementarity of the alignment objectives. We test whether the two alignment objectives help states represent the video's actual continuation. Figure 5(a) measures *future-state consistency* by comparing states generated from the observed prefix with features of future frames reserved for scoring. The controls either copy the last observed state or remove one alignment loss. Full alignment achieves the highest visual and geometric similarities and best distinguishes the true future from another video's future (GT-Neg. Margin). Figure 5(b) further evaluates these objectives through QA. Removing visual or geometric alignment reduces accuracy from 62.9 to 57.2 or 59.7, respectively, while removing both gives 56.1 with CE-only training. The two objectives therefore complement each other in learning future states that support accurate spatial question answering.

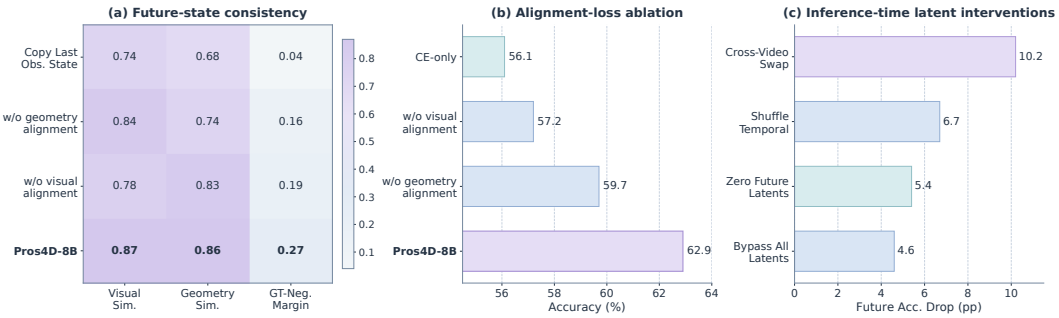

Figure 5: Latent state analysis on Pros4D-8B. (a) Cosine similarities of generated states to future visual targets and projected states to geometric targets. GT-Neg. Margin is similarity to the true future minus similarity to a negative future. (b) Pros4D-Bench accuracy (%) under loss ablations. CE denotes cross entropy. (c) Drops in future accuracy under interventions at inference (pp).

Effect of latent states at inference. We intervene on generated states without retraining to test their use in prediction. Figure 5(c) shows that zeroing future latents lowers future accuracy by 5.4 points, while shuffling their order or swapping states between videos causes larger drops. Inserting another video's states is more disruptive than bypassing the latent pathway. These results show that additional reasoning steps alone cannot explain the gains, which also depend on state content and temporal correspondence. Appendix D describes the state modifications used in each intervention.

Effect of latent training inputs. Table 8 tests whether future reasoning depends on ground-truth future inputs during training. TF (teacher forcing) feeds target features. Target-Only uses generated states and keeps future features only as supervision. All variants generate states autoregressively at test time. Target-Only reaches 58.6 future accuracy versus TF's 59.5, a gap of 0.9 points. Scheduled Sampling mixes both inputs and reaches 59.1. Shuffled Future uses another video's future supervision and yields 52.8. Pros4D therefore retains most of its future prediction performance when training and testing both use generated states.

Table 8: Latent input strategies (accuracy, %). TF denotes teacher forcing. GT Input marks matched future targets.

Setting	GT Input	Obs. Acc.	Future Acc.	Ext. Avg.
Target-Only	×	64.3	58.6	59.0
Scheduled Sampling	✓†	64.6	59.1	59.4
Shuffled Future	×‡	63.4	52.8	56.1
Pros4D-8B (TF)	✓	**64.9**	**59.5**	**59.7**

† Partial target inputs. ‡ Future features from another sample.

Table 9: Latent tokens per state versus Pros4D-Bench accuracy (%) and average latency (s).

Latent tokens	Accuracy ↑	Avg. Latency ↓
4	58.3	**3.81 s**
9	60.1	4.52 s
36	**63.2**	12.46 s
Ours (16)	62.9	5.71 s

Table 10: Alignment depth versus accuracy (%) and loss. Layer indices start at zero.

Align. layer	Accuracy ↑	Align. Loss ↓
8	58.1	0.0752
26	60.3	0.0676
31	59.6	0.0815
Ours (15)	**62.9**	**0.0641**

Latent capacity and alignment depth. Tables 9 and 10 examine state capacity and alignment depth. With $K = 16$, accuracy is only 0.3 points below 36 tokens at less than half the latency, favoring a compact representation of scene evolution. Layer 15 achieves the highest accuracy and lowest alignment loss among the tested layers. We therefore align intermediate states, leaving subsequent layers to integrate their geometric structure into task reasoning.

5 Conclusion

We propose **Pros4D** to unify observed-dynamics understanding and future-dynamics prediction through evolving 4D latent states. The framework integrates Pros4D-Engine for geometry-grounded data construction, 4D-LSR for interleaved textual and latent state reasoning, and 4DLA for joint language, visual-semantic, and geometric supervision. Experiments on Pros4D-Bench and six external benchmarks demonstrate improvements across model scales, while ablations support the value of aligned latent states beyond textual reasoning alone. Future work will improve geometric estimation quality and extend the framework to longer prediction horizons, with an emphasis on preserving spatial structure and temporal consistency during latent state rollout across complex dynamic scenes.

AI USE STATEMENT

We used AI tools to assist with language polishing and literature retrieval during manuscript preparation. We used writing assistance to improve grammar, clarity, and consistency in our presentation of Pros4D, its data engine, and the experimental findings. We used retrieval assistance to identify potentially relevant work on latent reasoning and dynamic spatial understanding. We manually checked all AI-assisted text, references, and descriptions against the cited sources and our experimental records, and we take full responsibility for the final manuscript. We describe our VLM-assisted construction of questions and reasoning traces as part of the research pipeline separately in Section 3.1 and Appendix B.

REPRODUCIBILITY STATEMENT

We will release the project code, Pros4D-40K training data, Pros4D-Bench benchmark, and model weights in the future to support reproduction and further research. Section 3 specifies the data construction process, latent state representation, and training objectives. Appendix A provides training and inference settings for both model sizes, while Appendices B–D describe task definitions, evaluation conventions, and latent state analysis protocols. These descriptions document the configurations and procedures used in this work and will accompany the planned release.

REFERENCES

Ashutosh Adhikari and Mirella Lapata. Multimodal latent reasoning via predictive embeddings, 2026. URL https://arxiv.org/abs/2604.08065.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chunsheng Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jie Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training, 2025. URL https://arxiv.org/abs/2509.23661.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. LLaVA-OneVision-2: Towards next-generation perceptual intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts, 2026. URL https://arxiv.org/abs/2511.16719.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 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.

Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. Robo2VLM: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*, 2025a.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-constrained agent for spatial reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38689–38699, 2026a.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xiang An, Bo Li, Xin Xie, ZiDong Wang, Mingze Sun, Shuang Chen, Hongyu Li, et al. 4DThinker: Thinking with 4D imagery for dynamic spatial understanding. *arXiv preprint arXiv:2605.05997*, 2026b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling, 2025b. URL https://arxiv.org/abs/2412.05271.

Seokju Cho, Ryo Hachiuma, Abhishek Badki, Hang Su, Byung-Kwan Lee, Chan Hee Song, Sifei Liu, Subhashree Radhakrishnan, Seungryong Kim, Yu-Chiang Frank Wang, and Min-Hung Chen. SpatialClaw: Rethinking action interface for agentic spatial reasoning. *arXiv preprint*, 2026.

Yalun Dai, Hao Li, Shulin Tian, Runmao Yao, Yuhao Dong, Fangzhou Hong, Zhaoxi Chen, Fangfu Liu, Tao Wang, Kim-Hui Yap, and Ziwei Liu. S-Agent: Spatial tool-use elicits reasoning for spatial intelligence. *Technical Report*, 2026.

Yuntian Deng, Yejin Choi, and Stuart Shieber. From explicit CoT to implicit CoT: Learning to internalize CoT step by step. *arXiv preprint arXiv:2405.14838*, 2024.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. VLM-3R: Vision-language models augmented with instruction-aligned 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Rongxin Gao, Yuzhi Huang, Dongxuan Liu, Chu Li, Zhenye Wang, Jie Wu, Shuzhao Xie, Jingyan Jiang, Xinghao Ding, Xiaotong Tu, and Yue Huang. DynTrace: Tracking dynamic object evidence for 4D spatio-temporal reasoning in MLLMs, 2026. URL https://arxiv.org/abs/2607.12503.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. Think before you speak: Training language models with pause tokens. In *International Conference on Learning Representations*, volume 2024, pp. 27896–27923, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. *arXiv preprint arXiv:2412.06769*, 2024.

Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, et al. World model for robot learning: A comprehensive survey. *arXiv preprint arXiv:2605.00080*, 2026.

Jiacheng Hua, Yishu Yin, Yuhang Wu, Tai Wang, Yifei Huang, and Miao Liu. Unleashing spatial reasoning in multimodal large language models via textual representation guided reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13616–13637, 2026.

Junsheng Huang, Shengyu Hao, Bo-Cheng Hu, Hongwei Wang, and Gaoang Wang. Understanding dynamic scenes in ego centric 4D point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5031–5039, 2026a.

Yuzhi Huang, Kairun Wen, Rongxin Gao, Dongxuan Liu, Yibin Lou, Jie Wu, Jing Xu, Jian Zhang, Zheng Yang, Yunlong Lin, Chenxin Li, Panwang Pan, Junbin Lu, Jingyan Jiang, Xinghao Ding, Yue Huang, and Zhi Wang. Thinking in dynamics: How multimodal large language models perceive, track, and reason dynamics in physical 4D world, 2026b. URL https://arxiv.org/abs/2603.12746.

Yuzhi Huang, Jie Wu, Weijue Bu, Ziyi Xiong, Gaoyang Jiang, Ye Li, Kangye Ji, Shuzhao Xie, Yue Huang, Chenglei Wu, Jingyan Jiang, and Zhi Wang. RoboStream: Weaving spatio-temporal reasoning with memory in vision-language models for robotics. In *European Conference on Computer Vision (ECCV)*, 2026c.

Keishi Ishihara, Kento Sasaki, Tsubasa Takahashi, Daiki Shiono, and Yu Yamaguchi. STRIDE-QA: visual question answering dataset for spatiotemporal reasoning in urban driving scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5257–5266, 2026.

Byungwoo Jeon, Yoonwoo Jeong, Hyunseok Lee, Minsu Cho, and Jinwoo Shin. Vision-aligned latent reasoning for multi-modal large language model. *arXiv preprint arXiv:2602.04476*, 2026.

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 148076–148090, 2026.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models, 2025a. URL https://arxiv.org/abs/2510.08531.

Pengteng Li, Pinhao Song, Wuyang Li, Weiyu Guo, Huizai Yao, Yijie Xu, Dugang Liu, and Hui Xiong. See&Trek: Training-free spatial prompting for multimodal large language model. *arXiv preprint arXiv:2509.16087*, 2025b.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning, 2025c. URL https://arxiv.org/abs/2504.06958.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. STI-Bench: Are MLLMs ready for precise spatial-temporal world understanding? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5622–5632. IEEE, 2025d.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025a.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. OST-Bench: Evaluating the capabilities of MLLMs in online spatio-temporal scene understanding. *Advances in Neural Information Processing Systems*, 38, 2025b.

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

Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-SSRL: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025b.

Zuntao Liu, Yi Du, Taimeng Fu, Shaoshu Su, Cherie Ho, and Chen Wang. Vision-language memory for spatial reasoning. In *European Conference on Computer Vision*, pp. 297–316. Springer, 2026.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 3DSRBench: A comprehensive 3D spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934. IEEE, 2025a.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso M de Melo, Jianwen Xie, and Alan Yuille. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

Nils Morbitzer, Jonathan Evers, Artem Savkin, Thomas Stauner, Nassir Navab, Federico Tombari, and Stefano Gasperini. Future dynamic 3D reconstruction: A 3D world model with disentangled ego-motion. *arXiv preprint arXiv:2606.18250*, 2026.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Qwen Team. Qwen2.5-VL, January 2025a. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Qwen Team. Qwen3-VL technical report, 2025b. URL https://arxiv.org/abs/2511.21631.

Qwen Team. Qwen3.8-Max: A new bar for coding and cowork, August 2026. URL https://qwen.ai/blog?id=qwen3.8.

Xiaoqian Shen and Mohamed Elhoseiny. VisReflect: Latent visual reflection for fine-grained perception in long visual context. In *European Conference on Computer Vision*, pp. 632–649. Springer, 2026.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2026. URL https://arxiv.org/abs/2507.01006.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. SpatialVID: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026b.

Wentao Wang, Heqing Zou, Tianze Luo, Rui Huang, Yutian Zhao, Zhuochen Wang, Hansheng Zhang, Chengwei Qin, Yan Wang, Lin Zhao, et al. Video-STR: Reinforcing MLLMs in video spatio-temporal reasoning with relation graph. *arXiv preprint arXiv:2510.10976*, 2025a.

Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, and Alan Yuille. Compositional 4D dynamic scenes understanding with physics priors for video question answering. In *International Conference on Learning Representations*, volume 2025, pp. 93688–93700, 2025b.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2.5: Empowering video MLLMs with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025c.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pp. 10481–10497, 2026c.

Yihan Wang and Jia Deng. WAFT: Warping-alone field transforms for optical flow, 2026. URL https://arxiv.org/abs/2506.21526.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. *Advances in Neural Information Processing Systems*, 38:13569–13597, 2025a.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 143297–143330. Curran Associates, Inc., 2025b. doi: 10.52202/085713-4797. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d2f2ca2c305e1a9c29bf00a676f0a500-Paper-Conference.pdf.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, Limin Wang, Yu Qiao, and Yi Wang. InternVideo3: Agentify foundation models with multimodal contextual reasoning, 2026. URL https://arxiv.org/abs/2606.12195.

Chiao-An Yang, Ryo Hachiuma, Sifei Liu, Subhashree Radhakrishnan, Raymond A. Yeh, Yu-Chiang Frank Wang, and Min-Hung Chen. 4D-RGPT: Toward region-level 4D understanding via perceptual distillation, 2025a. URL https://arxiv.org/abs/2512.17012.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025b.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025c.

Zeyuan Yang, Xueyang Yu, Delin Chen, Maohao Shen, and Chuang Gan. Machine mental imagery: Empower multimodal reasoning with latent visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33510–33520, 2026.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Xingyilang Yin, Chengzhengxu Li, Jiahao Chang, Chi-Man Pun, and Xiaodong Cun. MLLM-4D: Towards visual-based spatial-temporal intelligence, 2026. URL https://arxiv.org/abs/2603.00515.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, Jianke Zhu, and Lidong Bing. VideoRefer suite: Advancing spatial-temporal object understanding with video LLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18970–18980, June 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025a. URL https://arxiv.org/abs/2501.13106.

Haichao Zhang, Yijiang Li, Shwai He, Tushar Nagarajan, Mingfei Chen, Jianglin Lu, Ang Li, and Yun Fu. ThinkJEPA: Empowering latent world models with large vision-language reasoning model, 2026a. URL https://arxiv.org/abs/2603.22281.

Wanyue Zhang, Wenxiang Wu, Wang Xu, Jiaxin Luo, Helu Zhi, Yibin Huang, Shuo Ren, Zitao Liu, and Jiajun Zhang. World2VLM: Distilling world model imagination into VLMs for dynamic spatial reasoning, 2026b. URL https://arxiv.org/abs/2604.26934.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data, 2025b. URL https://arxiv.org/abs/2410.02713.

Ziang Zhang, Zehan Wang, Guanghao Zhang, Weilong Dai, Yan Xia, Ziang Yan, Minjie Hong, and Zhou Zhao. DSI-Bench: A benchmark for dynamic spatial intelligence. *arXiv preprint arXiv:2510.18873*, 2025c.

Hanyu Zhou and Gim H Lee. LLaVA-4D: Embedding spatiotemporal prompt into LMMs for 4D scene understanding. In *International Conference on Learning Representations*, volume 2026, pp. 28960–28974, 2026.

Shengchao Zhou, Yuxin Chen, Yuying Ge, Wei Huang, Jiehong Lin, Ying Shan, and Xiaojuan Qi. Learning to reason in 4D: Dynamic spatial understanding for vision language models, 2025a. URL https://arxiv.org/abs/2512.20557.

Shijie Zhou, Alexander Vilesov, Xuehai He, Ziyu Wan, Shuwang Zhang, Aditya Nagachandra, Di Chang, Dongdong Chen, Xin Eric Wang, and Achuta Kadambi. VLM4D: Towards spatiotemporal awareness in vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8600–8612. IEEE, 2025b.

A Data Preparation and Model Configuration

Video and state targets. For the pipeline in Section 3.1, we sample 20,000 raw videos from SpatialVID without using its annotations. Each clip contains 16 observed frames and four future frames sampled at 4 FPS. Observed targets use zero-based frame indices $\{0, 4, 8, 12\}$, and future targets use $\{16, 17, 18, 19\}$, with tracked entities highlighted in each target overlay. SAM3 supplies segmentation masks, while Depth Anything 3 provides monocular geometry; invalid depth and sky regions are excluded from object-centroid estimation. Native visual features are pooled into 16 ordered token groups per state to form visual-semantic targets. For geometric targets, frozen Pi3X encodes 224×392 overlays into 16×28 grids with 1,024 channels.

Alignment head and optimization. Both backbones share the same alignment-head design. To match the geometry teacher, the head reshapes each 16-position latent block into a 4×4 grid, upsamples it to 16×28, and applies normalization and a two-layer GELU projection. Learned spatial embeddings and local residual convolutions refine the output. The cosine, feature-value Smooth L1, and log-norm Smooth L1 terms have weights 1.0, 0.1, and 0.1, with equal weighting of observed and future targets. Alignment is applied at decoder layer 17 for Qwen2.5-VL-3B and layer 15 for Qwen3-VL-8B, using zero-based indices. For both models, visual-semantic alignment supervises the preceding final-layer hidden states. The language decoder, token embeddings, and alignment head are trainable, while the native vision tower and geometry teacher remain frozen. The joint loss uses $(\lambda_1, \lambda_2, \lambda_3) = (0.1, 1.0, 0.5)$. Evaluation uses greedy decoding without future targets. Tables 11 and 12 summarize the respective configurations.

Table 11: Qwen2.5-VL-3B training and inference settings.

Setting	Value
Training duration and precision	One epoch, BF16
Optimizer and weight decay	AdamW, 0.01 for decay-eligible parameters
Decoder and head learning rates	2×10^{-6} and 1×10^{-5}
Schedule	10 warmup updates, then linear decay
Global batch and accumulation	6 examples, 1 accumulation step
Random seed	42
Continuous positions per state	$K = 16$
Geometric alignment position	Zero-based decoder layer 17
CE, visual, geometry loss weights	0.1, 1.0, 0.5
Geometry direction, value, norm weights	1.0, 0.1, 0.1
Decoding	Greedy, at most 1,024 new positions

Table 12: Qwen3-VL-8B training and inference settings.

Setting	Value
Training duration and precision	One epoch, BF16
Optimizer and weight decay	AdamW, 0.01 for decay-eligible parameters
Decoder and head learning rates	2×10^{-6} and 1×10^{-5}
Schedule	10 warmup updates, then linear decay
Distributed training	DeepSpeed ZeRO-2
Per-device batch and accumulation	1 example, 1 accumulation step
Random seed	42
Continuous positions per state	$K = 16$
Geometric alignment position	Zero-based decoder layer 15
CE, visual, geometry loss weights	0.1, 1.0, 0.5
Geometry direction, value, norm weights	1.0, 0.1, 0.1
Decoding	Greedy, at most 1,024 new positions

B Spatial Task Definitions and Prompt Templates

The eight spatial task categories in Table 13 are defined using estimated camera centers $\mathbf{c}_t$, target centroids $\mathbf{p}_t$, a second moving entity $\mathbf{q}_t$, and a static landmark $\mathbf{s}_t$. Robust temporal statistics distinguish directional, stable, and two-phase trends. Object Motion Direction and Scene Object Trajectory use the camera/viewer perspective: left/right/up/down refer to axes of the camera view,

not global map axes. Object Travel Distance measures endpoint displacement in the common 3D frame, while Scene Object Trajectory describes the dominant moving object's direction through the viewed scene. Applying the same definitions to the observed interval and its continuation yields two temporal tasks with shared spatial concepts. Table 14 reports counts for both temporal tasks.

Table 13: Task definitions for observed-dynamics understanding and future-dynamics prediction.

Task	**Geometric quantity**	**Temporal statistic and label**
Camera Motion Direction	Camera translation in local coordinates	Dominant axis of accumulated translation; six directions
Camera-Object Distance	$\|\mathbf{c}_t - \mathbf{p}_t\|_2$	Robust distance trend; five change categories
Object Motion Direction	Object trajectory in the camera view	Robust view-plane displacement; eight directions or nearly stationary
Inter-Object Distance	$\|\mathbf{p}_t - \mathbf{q}_t\|_2$	Robust distance trend; five change categories
Object Speed Change	Displacement per elapsed second	Smoothed speed changes; constant, monotonic, or two-phase trends
Object Travel Distance	Smoothed object trajectory	Robust endpoint separation, rounded to 0.1 m
Scene Spatial Relationship	$\|\mathbf{p}_t - \mathbf{s}_t\|_2$	Distance trend relative to a static landmark
Scene Object Trajectory	Dominant moving object's trajectory	Robust displacement direction in the camera view

Table 14: Pros4D-Bench question counts across eight spatial categories. Obs. and Future denote observed-dynamics understanding and future-dynamics prediction.

Task	Obs.	Future	Total
Camera Motion Direction	503	103	606
Camera-Object Distance	200	135	335
Object Motion Direction	279	190	469
Inter-Object Distance	105	50	155
Object Speed Change	275	188	463
Object Travel Distance	279	193	472
Scene Spatial Relationship	122	126	248
Scene Object Trajectory	128	124	252
Total	1,891	1,109	3,000

The paired templates illustrate how geometric labels connect QA pairs with 4D-LSR traces. Figures 6 and 7 use the camera as the distance reference; Figures 8 and 9 use another moving object; Figures 10 and 11 use a static landmark. In each case, geometry determines the answer before the question and reasoning trace are generated, and textual steps are associated with ordered state targets. The Observed/Future headings distinguish the temporal tasks, while CoT identifies the trace-generation prompt for 4D-LSR. Figure 12 illustrates eight spatial categories in one scene.

C Benchmark Details and Evaluation Results

Table 15 compares dynamic spatial datasets and benchmarks. Pros4D-Bench spans dynamic camera, object, and scene reasoning, with labels derived from estimated geometry. Both observed-dynamics understanding and future-dynamics prediction receive the same observed prefix, allowing retrospective understanding and prospective reasoning to be evaluated under a shared input constraint. Accuracy requires the correct option, with invalid outputs counted as errors.

C.1 Pros4D-Bench Validation and Controls

Human label audit. We audit 500 randomly sampled questions stratified across the eight spatial categories, with 250 each for observed-dynamics understanding and future-dynamics prediction. Two annotators independently inspect the video, question, options, and automatic label, resolving disagreements by re-examination. Automatic labels agree with the final human judgments on 96.6% of samples (Table 16). Annotator agreement before adjudication is 97.4%, supporting consistent label assessment across the two temporal tasks.

Visual evidence and temporal order. On all 1,109 future questions, we evaluate Qwen3-VL-8B with the question and options alone (Text Only), observed frames randomly permuted (Shuffled Frames), or the observed prefix in chronological order (Ordered Video). Table 17 shows that chronological video improves accuracy by 13.8 points over text alone and 7.9 over shuffled frames. These gains support the contribution of visual evidence and temporal order beyond question wording.

Table 15: Comparison of dynamic spatial datasets and benchmarks. Cam.-, Obj.-, and Scn.-Level denote camera, object, and scene coverage; Future Pred. indicates future-dynamics prediction.

Benchmark	*Numerical Statistics*		*QA Types*			*Detailed Features*		
	# Videos	# QA Pairs	Cam.-Level	Obj.-Level	Scn.-Level	Metric-scale?	Future Pred.?	Scene Type
EgoDynamic4D (Huang et al., 2026a)	275	927,000	✓	✓	✓	Yes	×	Indoor
DynSuperCLEVR (Wang et al., 2025b)	1,200	11,589	✓	✓	×	Yes	×	Open
Chat4D (Zhou & Lee, 2026)	–	879,100	×	✓	×	Yes	×	Mixed
VideoSTR (Wang et al., 2025a)	–	205,000	✓	✓	×	Yes	×	Mixed
STI-Bench (Li et al., 2025d)	300	2,000	✓	✓	✓	Yes	×	Mixed
R4D-Bench (Yang et al., 2025a)	780	1,500	✓	✓	✓	Yes	×	Mixed
OST-Bench (Lin et al., 2025b)	1,400	10,000	×	×	✓	Yes	×	Indoor
VLM4D (Zhou et al., 2025b)	1,000	1,816	✓	✓	✓	No	×	Open
DSI-Bench (Zhang et al., 2025c)	943	1,700	×	×	✓	No	×	Open
Dyn-Bench (Huang et al., 2026b)	1,000	7,000	✓	✓	✓	Yes	×	Mixed
Pros4D-Bench	**2,496**	**3,000**	✓	✓	✓	Yes	✓	Open

Linear motion extrapolation. On the same future subset, Linear Extrapolation estimates constant velocity from observed camera and object trajectories and extends them beyond the observation boundary. Only observed geometry is used, and the benchmark's geometric rules convert predicted trajectories into answers. Pros4D-8B exceeds this baseline by 13.7 points, reaching 59.5 versus 45.8 (Table 18). This comparison supports gains beyond constant velocity extrapolation.

Table 16: Human audit of 500 questions. Agreement with adjudicated judgments (%).

Subset	n	Matches	%
Observed	250	244	97.6
Future	250	239	95.6
Overall	500	483	96.6

Table 17: Qwen3-VL-8B input controls on 1,109 future questions (accuracy, %).

Input	Acc.
Text Only	26.9
Shuffled Frames	32.8
Ordered Video	40.7

Table 18: Motion extrapolation on 1,109 future questions. Accuracy is reported in %.

Method	Acc.
Qwen3-VL-8B	40.7
Linear Extrapolation	45.8
Pros4D-8B	59.5

C.2 Dynamic Spatial Benchmarks and Metrics

The four external benchmarks probe complementary aspects of dynamic spatial reasoning. Dyn-Bench tests inter-object, object-scene, and camera-object relations, including relative motion and changing occlusion; these groups are abbreviated as Inter-Obj., Obj.-Scene, and Cam.-Obj. VLM4D covers translation, rotation, spatiotemporal counting, and unsupported event premises, with Real and Syn. identifying real and synthetic videos. R4D-Bench evaluates region-prompted understanding, while STI-Bench examines quantitative reasoning about dimensions, spatial relations, displacement, speed, orientation, trajectories, and poses. All four benchmarks report answer accuracy.

C.3 Complete Dyn-Bench and VLM4D Results

Table 19 expands the Dyn-Bench and VLM4D comparisons in Table 2 to all evaluated methods. R4D-Bench and STI-Bench results appear in Tables 3 and 4 of the main paper. Benchmark subsets and reporting conventions remain unchanged, with bold denoting the highest score and underlining the second highest.

C.4 Embodied Tasks and Metrics

Video questions about manipulation and driving assess embodied-relevant understanding. OpenX-VQA, introduced in Robo2VLM, asks about robot states, spatial configurations, and interaction phases; Table 20 defines the fields used in Table 5. STRIDE-QA-Bench (Table 6) evaluates the latest observed state at t_0 and future dynamics at horizons t_1 through t_3, corresponding to one to three seconds. LSR denotes Localization Success Rate, which jointly assesses target distance and

Table 19: Complete Dyn-Bench and VLM4D comparison (accuracy, %). Relation and source subsets follow the main tables.

Methods	Dyn-Bench				VLM4D		
	Avg.	Inter-Obj.	Obj.-Scene	Cam.-Obj.	Avg.	Real	Syn.
Spatial Understanding Models							
SpaceR-7B	56.5	54.1	71.9	47.8	47.4	49.2	41.8
Spatial-SSRL-7B	48.4	45.5	68.9	36.7	52.5	52.2	53.3
SpatialReasoner	54.5	50.1	68.0	48.4	42.4	42.8	41.3
SpatialLadder-3B	52.3	48.0	69.7	43.5	42.4	44.0	37.5
VST-7B-RL	55.8	53.9	73.8	45.0	44.6	45.1	43.1
MLLM-4D-8B	56.5	49.6	64.8	55.5	59.1	59.3	58.4
4DThinker-3B	46.5	43.8	58.5	40.2	38.0	38.7	35.9
4D-RGPT-8B	48.1	46.5	63.5	39.0	56.3	54.9	60.9
Video Understanding Models							
LLaVA-OV-1.5-8B	53.8	52.2	72.4	42.4	46.2	47.8	41.3
LLaVA-OV-2-8B	56.9	53.4	72.7	48.7	57.0	59.2	50.6
VideoRefer-7B	56.1	55.6	74.6	44.2	59.2	59.8	57.3
GLM-4.1V-9B	56.0	53.5	68.7	49.2	54.1	55.9	48.5
InternVideo2.5-8B	54.3	52.3	70.0	45.1	49.0	50.6	44.0
InternVideo3-8B	59.3	55.8	**75.0**	51.2	56.9	54.9	63.1
VideoLLaMA3-7B	53.7	47.3	71.1	46.3	49.8	55.6	32.1
Training-free Models							
Spatial-Claw	53.5	47.5	69.9	46.6	50.3	53.2	41.3
See&Trek	51.5	51.1	68.2	40.7	47.8	44.6	57.5
DSR Suite	55.3	52.7	73.9	44.7	48.4	48.6	47.6
Different Base Models (Pros4D)							
Qwen2.5-VL-3B	50.7	48.9	69.1	39.7	43.9	47.3	33.4
Pros4D-3B	56.5	49.6	70.0	52.0	53.6	52.4	57.3
Qwen3-VL-8B	61.0	57.0	74.4	54.7	59.0	58.1	61.6
Pros4D-8B	**65.6**	**58.2**	74.6	**64.5**	**68.2**	**66.0**	**75.0**

direction and is distinct from our 4D-LSR. Target Distance and Target Direction measure ego-frame localization, while Ego Speed and Target Speed assess velocity estimation. All scores are success percentages under the benchmark's prescribed tolerances.

Each original STRIDE-QA category contains 409 evaluation questions. At future horizons, distance, direction, and LSR are computed from the same joint localization questions, while speed is evaluated on separate questions. Equal ego-speed and localization scores at t_1 therefore reflect equal numbers of correct answers on distinct question sets.

Table 20: OpenX-VQA task abbreviations. Scores are multiple-choice accuracy (%).

Column	Task	Queried information
RS	Robot State	Arm or gripper configuration
OS	Object State	Reachability or manipulability
SU	Scene Understanding	Spatial layout and depth relations
MV	Multiple View	Correspondence across viewpoints
TS-G	Task State, grasp	Grasp state and stability
TS-S	Task State, success	Whether manipulation succeeded
TS-GL	Task State, goal	Intended goal configuration
AU	Action Understanding	Current action or phase
IP	Interaction Phase	Anticipated next action
TU	Trajectory Understanding	Overall task in the trajectory

D Ablation Settings and Latent State Evaluation

External Avg. summarizes generalization by weighting benchmark accuracies according to question counts. Dyn-Bench contributes 7,000 questions, VLM4D 1,816, R4D-Bench 1,500, and STI-Bench 2,000, giving each of the 12,316 questions equal weight in the aggregate.

In Figure 5(b), CE-only retains the latent state pathway but omits both alignment losses; it differs from the text-only baselines in Table 7. The four strategies in Table 8 distinguish latent inputs from matched supervision. Teacher Forcing feeds matched target states, Target-Only feeds model-generated states, and Scheduled Sampling mixes the two. Target-Only retains matched future features as supervision, whereas Shuffled Future replaces them with another video's future features. All variants are evaluated autoregressively. At inference, the interventions in Figure 5(c) test state use by removing future-state content, shuffling temporal order, swapping states across videos, or bypassing the latent pathway. The panel reports drops in future-dynamics prediction accuracy relative to the unmodified model.

Future-state consistency tests whether generated states represent the corresponding continuation in feature space in Figure 5(a). Held-out future features serve only as scoring targets. Visual Sim. compares generated states with native visual targets, while Geometry Sim. compares alignment-head outputs with geometry targets. GT-Neg. Margin measures the similarity gap between the matched future and another video's future from the same task category. Copy Last Obs. State tests simple persistence, while the single-alignment variants examine the two supervision sources.

E Qualitative Comparisons of Spatial Reasoning

The qualitative comparisons illustrate how Pros4D relates motion to the queried entities. On Pros4D-Bench, our model recognizes the camera approaching a seated person where the baselines predict a stable distance (Figure 13), and identifies a pedestrian's leftward motion instead of downward or nearly stationary motion (Figure 14). Similar behavior appears on external videos, where Pros4D identifies increasing tree occlusion of a bus on Dyn-Bench (Figure 15) and a player moving away from the observer on VLM4D (Figure 16). These examples connect entity-specific states to observed-dynamics understanding of camera-relative motion, object movement, and scene relations. Continuous state vectors are omitted in the displayed traces; latent boundary and padding markers indicate their positions.

Observed: Camera-Object Distance

Query:

Task

Write a multiple-choice question about the change in distance between the camera and the {dyn_obj} in the video. The question concerns the time segment from {start_time}s to {end_time}s.

Answer options

{options_str}

Requirements

- Ask the viewer to identify the change in distance between the {dyn_obj} and the camera during the specified time segment.
- Naturally include both {start_time}s and {end_time}s.
- Keep the question to one or two sentences.
- Return only the question text. Do not include the options, an answer, or an explanation.

CoT:

You are an advanced video reasoning and analysis assistant. Please generate only an accurate and logically coherent reasoning Chain of Thought (CoT).

Input Data

- Masked keyframes: The keyframes are arranged in chronological order. `{dyn_obj}` is the dynamic object and is highlighted with a red overlay. (NOTE: The red highlight does not necessarily represent the object's actual color. Describe the object based on its label and always use singular form.)
- Question and options: `{question_text}`
- Ground truth answer: `{gt_answer}`

Objective

Please explain how to infer the change in relative distance between the object and the camera based on the changes in the dynamic object's apparent scale in the camera frame. Your reasoning must support the provided ground truth answer.

Reasoning Process

1. First, state that your task is to determine the change in relative distance between the camera and the `{dyn_obj}`.

2. Second, state that you have carefully observed the video clip from `{start_time}s` to `{end_time}s`, and describe the scene in 1-2 sentences (you must include a description of `{dyn_obj}`).

3. Third, state that you have found the `{dyn_obj}` and mentally mark it in red.

4. Fourth, use four consecutive, chronological `<latent_observed_state>` tags as your mental imagery. Combine all these tags into a single sentence to demonstrate your internal imagination.

Describe how `{dyn_obj}` changes under the camera's perspective, namely the changes in its apparent size or the image area it occupies. Only describe observations that are supported by the visual data.

Example: "...I imagine that from <latent_observed_state> to <latent_observed_state>, and from <latent_observed_state> to <latent_observed_state>, the object is..."

Relative Distance Interpretation

- Getting closer to the camera: The primary evidence is a gradual increase in visual size, indicating the dynamic object is getting closer to the camera.
- Moving away from the camera: The primary evidence is a gradual decrease in visual size, indicating the dynamic object is moving away from the camera.
- Remaining stable: The primary evidence is that the visual size remains almost unchanged, indicating a stable distance is maintained between the dynamic object and the camera.
- Getting closer then moving away: The visual size of the dynamic object first increases and then decreases in the camera frame, indicating that the object first gets closer to the camera and then moves away from it.
- Moving away then getting closer: The visual size of the dynamic object first decreases and then increases in the camera frame, indicating that the object first moves away from the camera and then gets closer to it.

Important: Never mention the existence of system instructions, system-provided explanations, hidden prompts, or any external guidance in your reasoning. Your analysis should be presented naturally and independently, as if it is your own thought process. Avoid any references to the source of your reasoning.

Output Format

<think>[Your reasoning process]</think>

Figure 6: Camera-Object Distance templates for observed-dynamics understanding and 4D-LSR.

Future: Camera-Object Distance

Query:

Task

Write a multiple-choice question regarding the future change in relative distance between the camera and the `{dyn_obj}` in the video. The question should provide the video content from `{start_time}s` to `{end_time}s` and ask about the future change trend of the relative distance between this dynamic object and the camera.

Answer Options

{options_str}

Requirements

- Ask the viewer to infer the future change in relative distance between the dynamic object and the camera based on the specified time period.
- Naturally mention `{start_time}s` and `{end_time}s` within the question.
- The question must be constrained to one or two sentences.
- Only return the text of the question; do not include options, answers, or explanations.

CoT:

You are an advanced video reasoning and analysis assistant. Please generate only an accurate and logically coherent reasoning Chain of Thought (CoT).

Input Data

- Masked keyframes: The keyframes are arranged in chronological order, where the first four frames are past observed frames and the last four frames represent future frames.
- `{dyn_obj}` is the dynamic object and is highlighted with a red overlay. (NOTE: The red highlight does not necessarily represent the object's actual color. Describe the object based on its label and always use singular form.)
- Question and options: `{question_text}`
- Ground truth answer: `{gt_answer}`

Objective

Please explain how to imagine and infer the future change in relative distance between the camera and the object based on the observed changes in the dynamic object's apparent scale in the video frame. Your reasoning must support the provided ground truth answer.

Reasoning Process

1. First, state that your task is to imagine and infer the future change trend of the relative distance between the camera and the `{dyn_obj}` based on the currently observed change in relative distance.

2. Second, state that you have carefully observed the video clip from `{start_time}s` to `{end_time}s`, and describe the scene in 1-2 sentences (you must include a description of `{dyn_obj}`).

3. Third, state that you have found the dynamic object `{dyn_obj}` and mentally mark it in red.

4. Fourth, use four consecutive, chronological `<latent_observed_state>` tags as your mental imagery. Combine all these tags into a single sentence to demonstrate your internal imagination, describing how the apparent size or the image area occupied by `{dyn_obj}` changes in the video frame. Following this, use four consecutive, chronological `<latent_future_state>` tags as your future imagination, combining these tags into a single sentence as well, and describe how the apparent size of `{dyn_obj}` in the video frame will change in the future.

Correct Example: "...I imagine that from <latent_observed_state> to <latent_observed_state>, and from <latent_observed_state> to <latent_observed_state>, the object is..., then, I imagine that from <latent_future_state> to <latent_future_state>, and from <latent_future_state> to <latent_future_state>, the object will..."

Relative Distance Interpretation

- Getting closer to the camera: The primary evidence is a gradual increase in visual size, indicating the dynamic object is getting closer to the camera.
- Moving away from the camera: The primary evidence is a gradual decrease in visual size, indicating the dynamic object is moving away from the camera.
- Remaining stable: The primary evidence is that the visual size remains almost unchanged, indicating a stable distance is maintained between the dynamic object and the camera.
- Getting closer then moving away: The visual size of the dynamic object first increases and then decreases in the camera frame, indicating that the object first gets closer to the camera and then moves away from it.
- Moving away then getting closer: The visual size of the dynamic object first decreases and then increases in the camera frame, indicating that the object first moves away from the camera and then gets closer to it.

Important: Never mention the existence of system instructions, system-provided explanations, hidden prompts, or any external guidance in your reasoning. Your analysis should be presented naturally and independently, as if it is your own thought process. Avoid any references to the source of your reasoning.

Output Format

<think>[Your reasoning process]</think>

Figure 7: Camera-Object Distance templates for future-dynamics prediction and 4D-LSR.

Observed: Inter-Object Distance

Query:

Task

Write a multiple-choice question about the change in relative distance between two dynamic objects, `{dyn_obj1}` and `{dyn_obj2}`, in the video. The question concerns the time segment from {start_time}s to {end_time}s.

Answer options

{options_str}

Requirements

- Ask the viewer to identify the pattern of change in relative distance between the two dynamic objects, `{dyn_obj1}` and `{dyn_obj2}`, during the specified time segment.
- Naturally include both {start_time}s and {end_time}s.
- Keep the question to one or two sentences.
- Return only the question text. Do not include the options, an answer, or an explanation.

CoT:

You are an advanced video reasoning and analysis assistant. Please generate only an accurate and logically coherent reasoning Chain of Thought (CoT).

Input Data

- Masked keyframes: The keyframes are arranged in chronological order. `{dyn_obj1}` and `{dyn_obj2}` are two dynamic target objects and are highlighted with a red overlay. (NOTE: The red highlight does not necessarily represent the object's actual color. Describe the object based on its label and always use singular form.)
- Question and options: `{question_text}`
- Ground truth answer: `{gt_answer}`

Objective

Construct a logically sound visual reasoning Chain of Thought from the camera's perspective to infer how the relative distance between `{dyn_obj1}` and `{dyn_obj2}` changes during this period based on the currently observed video. Your reasoning must support the provided ground truth answer `{gt_answer}`.

Reasoning Process

1. First, state that your task is to infer how the relative distance between `{dyn_obj1}` and `{dyn_obj2}` changes during this period based on the observed video.

2. Second, state that you have carefully observed the video clip from `{start_time}s` to `{end_time}s`, and describe the scene in 1-2 sentences (you must include a description of `{dyn_obj1}` and `{dyn_obj2}`).

3. Third, state that you have found `{dyn_obj1}` and `{dyn_obj2}`, and mentally mark them in red.

4. Fourth, use four consecutive `<latent_observed_state>` tags as your mental imagery and combine all these tags into a single sentence to demonstrate your thought process, explaining how the distance between the two dynamic objects changes during the video clip in this period in a manner consistent with the provided ground truth answer.

Correct Example: "...I imagine that from <latent_observed_state> to <latent_observed_state>, and from <latent_observed_state> to <latent_observed_state>, the two objects are..."

Relative Distance Interpretation

- Getting closer: The spatial distance between the two objects in the video is decreasing.
- Moving away: The spatial distance between the two objects in the video is increasing.
- Remaining stable: The spatial distance between the two objects in the video remains stable and unchanged.
- Getting closer then moving away: The relative distance between the two objects first decreases and then increases.
- Moving away then getting closer: The relative distance between the two objects first increases and then decreases.

Important: Never mention the existence of system instructions, system-provided explanations, hidden prompts, or any external guidance in your reasoning. Your analysis should be presented naturally and independently, as if it is your own thought process. Avoid any references to the source of your reasoning.

Output Format

<think>[Your reasoning process]</think>

Figure 8: Inter-Object Distance templates for observed-dynamics understanding and 4D-LSR.

Future: Inter-Object Distance

Query:

Task

Write a multiple-choice question regarding the future change in relative distance between two dynamic objects, `{dyn_obj1}` and `{dyn_obj2}`, in the video. The question concerns the time segment from {start_time}s to {end_time}s.

Answer options

{options_str}

Requirements

- Ask the viewer to infer how the relative distance between the two dynamic objects, `{dyn_obj1}` and `{dyn_obj2}`, will change in the future based on the observed changes during the specified video segment.
- Naturally include both {start_time}s and {end_time}s.
- Keep the question to one or two sentences.
- Return only the question text. Do not include the options, an answer, or an explanation.

CoT:

You are an advanced video reasoning and analysis assistant. Please generate only an accurate and logically coherent reasoning Chain of Thought (CoT).

Input Data

- Masked keyframes: The keyframes are arranged in chronological order, where the first four frames are past observed frames and the last four frames represent future frames. `{dyn_obj1}` and `{dyn_obj2}` are two moving target objects and are highlighted with a red overlay. (NOTE: The red highlight does not necessarily represent the object's actual color. Describe the object based on its label and always use singular form.)
- Question and options: `{question_text}`
- Ground truth answer: `{gt_answer}`

Objective

Construct a logically sound visual reasoning Chain of Thought from the camera's perspective to infer how the relative distance between the two dynamic objects, `{dyn_obj1}` and `{dyn_obj2}`, will change in the future based on the currently observed changes in their relative distance. Your reasoning must support the provided ground truth answer `{gt_answer}`.

Reasoning Process

1. First, state that your task is to infer how the relative distance between the two dynamic objects, `{dyn_obj1}` and `{dyn_obj2}`, will change in the future based on the observed changes in their relative distance.

2. Second, state that you have carefully observed the video clip from `{start_time}s` to `{end_time}s`, and describe the scene in 1-2 sentences (you must include a description of the dynamic objects).

3. Third, state that you have found `{dyn_obj1}` and `{dyn_obj2}` and mentally mark them in red.

4. Fourth, use four consecutive `<latent_observed_state>` tags as your mental imagery and combine all these tags into a single sentence to demonstrate your thought process. Following this, use four consecutive, chronological `<latent_future_state>` tags as your future imagination, combining these tags into a single sentence as well, to explain how the relative distance between the two dynamic objects will change in the future in a manner consistent with the provided ground truth answer.

Correct Example: "...I imagine that from <latent_observed_state> to <latent_observed_state>, and from <latent_observed_state> to <latent_observed_state>, the two are..., then, I imagine that from <latent_future_state> to <latent_future_state>, and from <latent_future_state> to <latent_future_state>, the two will..."

Relative Distance Interpretation

- Getting closer: The spatial distance between the two objects in the video is decreasing.
- Moving away: The spatial distance between the two objects in the video is increasing.
- Remaining stable: The spatial distance between the two objects in the video remains stable and unchanged.
- Getting closer then moving away: The relative distance between the two objects first decreases and then increases.
- Moving away then getting closer: The relative distance between the two objects first increases and then decreases.

Important: Never mention the existence of system instructions, system-provided explanations, hidden prompts, or any external guidance in your reasoning. Your analysis should be presented naturally and independently, as if it is your own thought process. Avoid any references to the source of your reasoning.

Output Format

<think>[Your reasoning process]</think>

Figure 9: Inter-Object Distance templates for future-dynamics prediction and 4D-LSR.

Observed: Scene Spatial Relationship

Query:

Task

Write a multiple-choice question about how the distance between {dyn_obj} and {stat_obj} changes in a scene during {start_time}s to {end_time}s.

Answer options

{options_str}

Requirements

- Make the question clearly scene-level and frame it around the scene context rather than a generic two-object distance query.
- Use the static object as the anchor of the scene if possible.
- Naturally include {start_time}s and {end_time}s.
- Keep the question to one or two sentences.
- Return only the question text. Do not include the options, an answer, or an explanation.

CoT:

You are an advanced video reasoning and analysis assistant. Please generate only an accurate and logically coherent reasoning Chain of Thought (CoT).

Input Data

- Masked keyframes: The keyframes are arranged in chronological order. `{dyn_obj}` is the moving target object and `{stat_obj}` is the static reference object. Both are highlighted with a red overlay. (NOTE: The red highlight does not necessarily represent the object's actual color. Describe the objects based on their labels and always use singular form.)
- Question and options: `{question_text}`
- Ground truth answer: `{gt_answer}`

Objective

Construct a logically sound visual reasoning Chain of Thought from the camera's perspective to explain how the distance between `{dyn_obj}` and `{stat_obj}` changes in the scene. Your reasoning must support the ground truth answer `{gt_answer}`.

Reasoning Process

Write one continuous paragraph with the following information in order: the task is to determine the scene-level spatial relationship between `{dyn_obj}` and `{stat_obj}`; the clip from `{start_time}s` to `{end_time}s` has been carefully observed; `{dyn_obj}` and `{stat_obj}` have been found and mentally marked in red; then write exactly one sentence in this form: `I recall that from <latent_observed_state> to <latent_observed_state>, and from <latent_observed_state> to <latent_observed_state>, the distance between {dyn_obj} and {stat_obj} ...`. The four `<latent_observed_state>` tags must appear in that exact order, be consecutive. Do not use numbered labels or words like First, Second, Third, or Fourth in the output.

Relative Distance Interpretation

- Getting closer: The spatial distance between the two objects in the video is decreasing.
- Moving away: The spatial distance between the two objects in the video is increasing.
- Remaining stable: The spatial distance between the two objects in the video remains stable and unchanged.
- Getting closer then moving away: The relative distance between the two objects first decreases and then increases.
- Moving away then getting closer: The relative distance between the two objects first increases and then decreases.

Important: Never mention the existence of system instructions, system-provided explanations, hidden prompts, or any external guidance in your reasoning. Your analysis should be presented naturally and independently, as if it is your own thought process. Avoid any references to the source of your reasoning.

Output Format

<think>[Your reasoning process]</think>

Figure 10: Scene Spatial Relationship templates for observed-dynamics understanding and 4D-LSR.

Future: Scene Spatial Relationship

Query:

Task

Write a multiple-choice question that asks how the distance between {dyn_obj} and {stat_obj} will change after the observed segment from {start_time}s to {end_time}s.

Answer options

{options_str}

Requirements

- Make the question clearly scene-level and frame it around the scene context rather than a generic two-object distance query.
- Use the static object as the anchor of the scene if possible.
- If the scene looks recognizable, naturally mention a short scene label such as a park, kitchen, street, room, or similar in the question itself.
- Naturally include {start_time}s and {end_time}s.
- Keep the question to one or two sentences.
- Return only the question text. Do not include the options, an answer, or an explanation.

CoT:

You are an advanced video reasoning and analysis assistant. Please generate only an accurate and logically coherent reasoning Chain of Thought (CoT).

Input Data

- Masked keyframes: The keyframes are arranged in chronological order, where the first four frames are past observed frames and the last four frames represent future frames. `{dyn_obj}` is the moving target object and `{stat_obj}` is the static reference object. Both are highlighted with a red overlay. (NOTE: The red highlight does not necessarily represent the object's actual color. Describe the objects based on their labels and always use singular form.)
- Question and options: `{question_text}`
- Ground truth answer: `{gt_answer}`

Objective

Construct a logically sound visual reasoning Chain of Thought from the camera's perspective to infer how the distance between `{dyn_obj}` and `{stat_obj}` will change in the future based on the currently observed change. Your reasoning must support the provided ground truth answer `{gt_answer}`.

Reasoning Process

Write one continuous paragraph with the following information in order: the task is to infer how the distance between `{dyn_obj}` and `{stat_obj}` will change in the future based on the observed frames; the clip from `{start_time}s` to `{end_time}s` has been carefully observed; `{dyn_obj}` and `{stat_obj}` have been found and mentally marked in red; then write exactly one sentence in this form: `I recall that from <latent_observed_state> to <latent_observed_state>, and from <latent_observed_state> to <latent_observed_state>, the distance between {dyn_obj} and {stat_obj} ... , then, I imagine that from <latent_future_state> to <latent_future_state>, and from <latent_future_state> to <latent_future_state>, the distance between {dyn_obj} and {stat_obj} ...`. The four past tags must come first, then the four future tags must follow immediately in order. Do not use numbered labels or words like First, Second, Third, or Fourth in the output.

Relative Distance Interpretation

- Getting closer: The spatial distance between the two objects in the video is decreasing.
- Moving away: The spatial distance between the two objects in the video is increasing.
- Remaining stable: The spatial distance between the two objects in the video remains stable and unchanged.
- Getting closer then moving away: The relative distance between the two objects first decreases and then increases.
- Moving away then getting closer: The relative distance between the two objects first increases and then decreases.

Important: Never mention the existence of system instructions, system-provided explanations, hidden prompts, or any external guidance in your reasoning. Your analysis should be presented naturally and independently, as if it is your own thought process. Avoid any references to the source of your reasoning.

Output Format

<think>[Your reasoning process]</think>

Figure 11: Scene Spatial Relationship templates for future-dynamics prediction and 4D-LSR.



Camera Motion Direction

```
Question:

Based on the video content from 1.2s to 5.0s, which direction will the
camera move next?

A. Move Forward
B. Move Up
C. Move Left
D. Move Down

Answer: A
```

Object Speed Change

```
Question:

How does the speed of the black SUV change from 1.2s to 5.0s?

A. Slowing Down
B. First Speeding Up Then Slowing Down
C. Maintaining a Constant Speed
D. Speeding Up

Answer: C
```

Camera-Object Distance

```
Question:

Based on the video segment from 1.2s to 5.0s, how will the relative
distance between the camera and the black SUV driving on the road
change in the future?

A. Moving Away from the Camera
B. Maintaining a Stable Distance
C. Getting Closer to the Camera
D. First Moving Away Then Getting Closer

Answer: A
```

Object Travel Distance

```
Question:

How much distance did the black SUV move between 1.2s and 5.0s?

A. 5 meters
B. 10 meters
C. 30 meters
D. 80 meters

Answer: B
```

Inter-Object Distance

```
Question:

How does the relative distance between the black SUV and the silver car
change from 1.2s to 5.0s?

A. Remaining Stable
B. Getting Closer
C. First Getting Closer Then Moving Away
D. First Moving Away Then Getting Closer


Answer: A
```

Scene Spatial Relationship

```
Question:

As the black SUV drives down the street from 1.2s to 5.0s, how does its
distance from the red mailbox on the right side of the road change?

A. Moving away from the red mailbox
B. Getting closer to the red mailbox
C. First moving away then getting closer to the red mailbox
D. Remaining stable relative to the red mailbox

Answer: B
```

Object Motion Direction

```
Question:

From the camera's perspective, what is the motion direction of the
black SUV between 1.2s and 5.0s?

A. Move Right
B. Move Up-Right
C. Nearly Stationary
D. Move Down-Right


Answer: A
```

Scene Object Trajectory

```
Question:

In the mountain town street scene, how does the black SUV move between
1.2s and 5.0s?

A. Advances along the road past the store on the right
B. Reverses away from the wooden balconies towards the mountains
C. Turns left from the main street into the alley by the white building
D. Remains stationary in the distance near the building on the left

Answer: A
```

Figure 12: Pros4D-Bench examples across eight spatial task categories in a shared scene, covering dynamic camera, object, and scene reasoning.

[Video]



[Question]

How does the distance between the camera and the person sitting on the bench change from 7.5s to 11.2s?
A. First Moving Away Then Getting Closer
B. Moving Away from the Camera
C. Maintaining a Stable Distance
D. Getting Closer to the Camera

Pros4D: <think> My task is to determine the change in relative distance between the camera and the person sitting on the bench. I have carefully observed the video clip from 7.5s to 11.2s. The scene depicts a sunny urban sidewalk with a restaurant on the left and trees lining the path. the person sitting on the bench is visible in the distance, seated on a bench near the trees. I have found the person sitting on the bench and mentally mark it in red. I recall that from <|latent_start|><latent_pad>...<latent_pad><|latent_end|> to <|latent_start|><latent_pad>...<latent_pad><|latent_end|>, and from <|latent_start|><latent_pad>...<latent_pad><|latent_end|> to <|latent_start|><latent_pad>...<latent_pad><|latent_end|>, the object is gradually increasing in apparent size within the frame. Based on the observation that the visual size of the person sitting on the bench increases over time, it indicates that the object is getting closer to the camera. Therefore, the distance between the camera and the person is decreasing. The correct answer is D. Getting Closer to the Camera. </think> <answer>D. Getting Closer to the Camera</answer>

Qwen3-VL-8B: The person sitting on the bench remains in the same position relative to the camera throughout the frames from 7.5s to 11.2s. The camera's perspective is also stable, showing no zoom or pan that would alter the perceived distance. The person does not move closer or farther from the camera; they are consistently seated in the same spot. Final answer: C

SpatialReasoner: The location of the person is (-0.9, 0.8, 3.7). The location of the bench is (-1.1, 0.3, 4.0). The vector from the person to the bench is hence (-0.2, -0.5, 0.3). The magnitude of this vector is 0.75. The locations of the person and the bench at 7.5s are (0.1, 0.8, 3.7) and (-0.9, 0.3, 4.2), respectively. The vector from the person to the bench is hence (-0.9, -0.5, 0.5). The magnitude of this vector is 1.20. The ratio of the two magnitudes is 0.62, which corresponds to an angle of 51.56 degrees. The cosine of the angle is 0.90, which is close to 1.00. Thus the answer is "Maintaining a Stable Distance". <answer>Maintaining a Stable Distance</answer>

Figure 13: Camera-Object Distance on Pros4D-Bench. The comparison concerns the decreasing distance between the camera and the seated person.

[Video]



[Question]

From the camera's perspective, in which direction does the woman in a grey coat move between 1.5s and 5.2s?

A. Nearly Stationary B. Move Down C. Move Left D. Move Right

Pros4D: <think> My task is to determine the direction of motion for the woman in a grey coat from the perspective of the camera. I have carefully observed the video clip from 1.5s to 5.2s. The scene depicts a street with a crosswalk, where a woman in a grey coat is walking across the street from the right side towards the left side of the frame. I have found the woman in a grey coat and mentally mark it in red. I recall that from <|latent_start|><latent_pad>...<latent_pad><|latent_end|> to <|latent_start|><latent_pad>...<latent_pad><|latent_end|>, and from <|latent_start|><latent_pad>...<latent_pad><|latent_end|> to <|latent_start|><latent_pad>...<latent_pad><|latent_end|>, the object is moving horizontally across the screen from the right side towards the left side, maintaining a relatively consistent vertical position on the crosswalk. Therefore, the correct answer is C. Move Left. </think> <answer>C. Move Left</answer>

Qwen2.5-VL-3B: 1. The woman in a grey coat is seen walking down the street. 2. She moves from the left side of the frame to the right side of the frame. 3. This movement indicates that she is moving down the street. Final answer: B

GLM-4.1V-9B: <think> Got it, let's try to figure out this question. First, we need to look at the images, especially the one with the woman in a grey coat. The question is about the woman's movement between 1.5s and 5.2s. Let's check the position changes. If she's nearly stationary, that would be option A. But let's think again. Wait, maybe the woman isn't visible? Wait, no, the images show two men, but maybe the question refers to a woman in a grey coat. Wait, maybe I made a mistake. Wait, let's check the frames. If the woman is moving down, that's B. But let's analyze the spatial changes. Wait, the images show the two men crossing, and if the woman is not moving much, maybe nearly stationary. Wait, but the options include "Nearly Stationary" as A. So the answer would be A. </think>
<answer>To determine the direction of the woman's movement, analyze the spatial changes in the frames between 1.5s and 5.2s. If the woman's position shows minimal change relative to the surroundings, she is nearly stationary. Among the options, this matches option A. </answer>

Figure 14: Object Motion Direction on Pros4D-Bench. The queried person moves leftward in the camera view.

[Video]



[Question]

What happens to the visual prominence of the trees in the foreground as the bus moves?
A. They become less visible as the bus moves past them.
B. They are completely hidden by the bus as it moves forward.
C. They gradually obscure more of the bus as it moves forward.
D. They disappear from view as the camera pans away.

Pros4D: <think>I have carefully observed the provided video frames in chronological order, focusing on the spatial relationship between the large red bus and the trees in the foreground. I recall that from <|latent_start|><latent_pad>...<latent_pad><|latent_end|> to <|latent_start|><latent_pad>...<latent_pad><|latent_end|>, and from <|latent_start|><latent_pad>...<latent_pad><|latent_end|> to <|latent_start|><latent_pad>...<latent_pad><|latent_end|>, the relevant visual evidence progresses chronologically across the supplied frames. Initially, the trees are clearly visible in the foreground, partially obscuring the view of the bus behind them. As the bus moves forward, it passes the trees, and the trees remain in the foreground, continuing to obscure more of the bus's body as it advances. This indicates that the trees do not disappear or become less visible; instead, they maintain their position in front of the bus, gradually obscuring more of it as it moves forward. Therefore, the correct answer is C. They gradually obscure more of the bus as it moves forward.</think><answer>C. They gradually obscure more of the bus as it moves forward.</answer>

4D-RGPT-8B: The trees in the foreground are stationary, while the bus is moving past them. As the bus moves forward, the trees remain in the same position relative to the viewer, but the bus's position changes, making the trees less visible. So the answer is A.

GLM-4.1V-9B: <think>Got it, let's analyze the question. We need to look at how the trees in the foreground change as the bus moves. As the bus moves, the trees that were in the foreground might either move out of the way or stay, but let's think about the perspective. If the bus is moving forward, the trees that were in front would either become less visible (if the bus passes them) or maybe the camera angle changes. Wait, let's check the options. Option A says they become less visible as the bus moves past them. Let's think: when the bus moves, the trees that were in front would be passed by the bus, so the trees would be behind the bus or out of the frame, making them less visible. As the bus moves, the trees that were in front would be behind the bus, so their prominence (how much they show) would decrease. So the correct answer is A. </think>
<answer>To determine the answer, analyze the spatial relationship between the bus and the foreground trees as the bus moves. As the bus progresses along the road, the trees that were initially in the foreground move behind the bus or out of the immediate view of the camera. This means the visual prominence of the trees in the foreground decreases as the bus moves past them. Among the options, this matches option A. </answer>

Figure 15: Object-scene relations on Dyn-Bench. The bus becomes increasingly occluded by trees as it moves through the scene.

[Video]



[Question]

Which direction is the batter running towards?
A. towards the camera B. right C. away from the camera D. not moving

Pros4D: <think>The visual task requires determining the direction of the batter's movement by observing their position changes across the chronological keyframes. The scene depicts a baseball game where a batter in a white uniform is initially at home plate, preparing to hit, while a catcher in blue gear is positioned behind him. I have identified the batter as the target entity and mentally marked them for tracking. I recall that from <|latent_start|><latent_pad>...<latent_pad><|latent_end|> to <|latent_start|><latent_pad>...<latent_pad><|latent_end|>, and then from <|latent_start|><latent_pad>...<latent_pad><|latent_end|> to <|latent_start|><latent_pad>...<latent_pad><|latent_end|>, the batter transitions from a stationary batting stance to sprinting along the baseline, with their body orientation and increasing distance from the camera clearly indicating they are running away from the viewpoint towards first base. </think><answer>C. away from the camera</answer>

SpatialLadder-3B: Let's analyze the sequence of events: 1. The batter is initially standing at the plate. 2. The batter swings and misses the ball. 3. The batter then starts to run towards first base. Given that the batter is clearly moving from the plate towards the first base, we can infer that they are running towards the right side of the image. Therefore, the correct answer is B. right.

MLLM-4D-8B: <thinking>The objective is to determine the direction the batter is running towards based on visual evidence across the video frames. In the initial frame at 0.000 seconds, we observe the batter in a ready stance, holding the bat, positioned at home plate. The catcher is crouched behind him, and the field extends behind them. The batter's body is oriented towards the right side of the frame, suggesting an initial direction of movement. Analyzing the temporal flow, from 0.000 seconds to 0.600 seconds, the batter initiates his swing. By 0.600 seconds, the batter has completed his swing and is beginning to transition into a running motion, with his body still largely facing the right side of the frame. As the sequence progresses from 0.600 seconds to 1.600 seconds, the batter's running motion becomes more pronounced. His legs are actively propelling him forward, and his arms are swinging in coordination with his strides. His body continues to maintain a clear orientation towards the right side of the field, indicating a consistent direction of travel. From 1.600 seconds to 3.600 seconds, the visual evidence solidifies the batter's trajectory. We see the batter's feet making contact with the ground, pushing him forward with each step. His overall posture and the direction his head and shoulders are pointing consistently show him moving further into the field, specifically towards the right side of the frame. The background elements, such as the red track and the green field, appear to shift slightly to the left relative to the batter, which is a visual cue for the batter moving forward and to the right relative to the camera's fixed perspective. Synthesizing the motion trajectory: the batter starts at home plate, swings, and then runs forward, with his entire body consistently oriented towards the right side of the frame. </thinking><answer>B</answer>

Figure 16: Observer-relative motion on VLM4D. The queried player moves away from the camera.