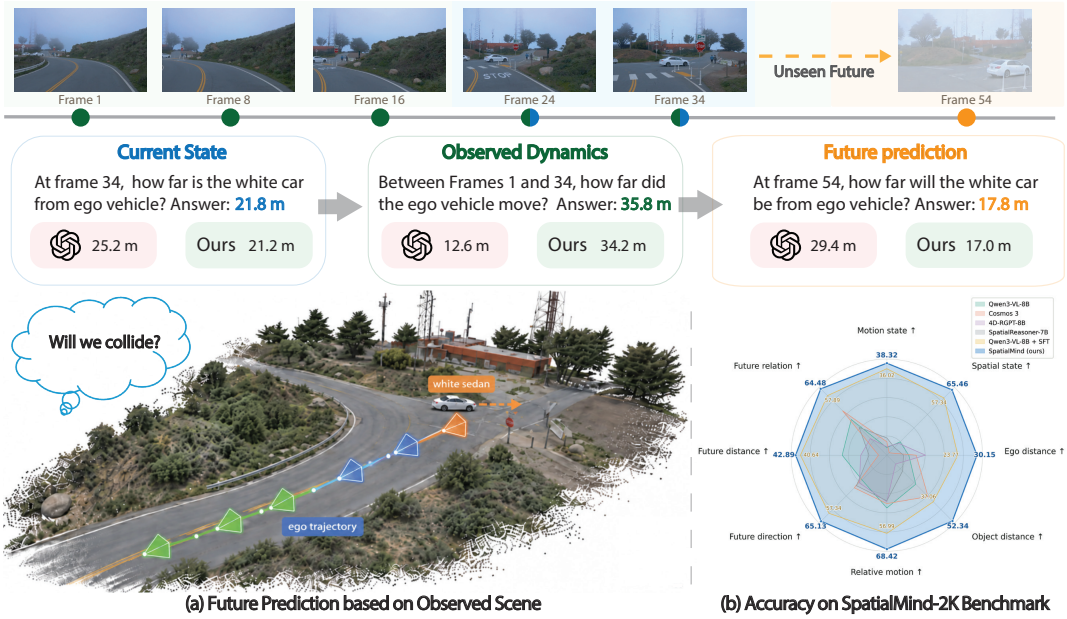
FROM SIGHT TO FORESIGHT: PREDICTIVE SPATIAL REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review



(a) Future Prediction based on Observed Scene (b) Accuracy on SpatialMind-2K Benchmark

Figure 1: SpatialMind enables future spatial prediction by reasoning progressively from observed evidence, which can help identify collision risks and support timely action to avoid hazards. By grounding prediction in current geometry and observed dynamics, SpatialMind yields more accurate estimates than a proprietary VLM. (a) The 3D scene illustrates the spatial context for anticipating future interactions between the ego vehicle and the white sedan. (b) Benchmark results show consistent improvements over the compared models across all eight spatial reasoning tasks.

ABSTRACT

Predicting future spatial states supports collision avoidance and timely decision-making in dynamic environments. However, existing vision-language models (VLMs) and benchmarks for spatial reasoning primarily focus on observed scenes, leaving predictive spatial reasoning beyond the observed interval underexplored. To this end, we introduce SpatialMind, a metric-scale VLM for spatial reasoning and future prediction. Its metric depth adapter anchors spatial reasoning to real-world scale, while its progressive state chain establishes current spatial states and observed dynamics as the foundation for future prediction. Given a video prefix, SpatialMind predicts distances, motion directions, and spatial relations in both observed and unseen future frames. For training and evaluation, we build a scalable data engine that grounds entity descriptions in metric geometry to generate question-answer pairs and state supervision. Using this engine, we construct the SpatialMind-30K dataset and the SpatialMind-2K benchmark, both covering driving and everyday egocentric scenes. The benchmark spans eight tasks across three levels: current-state understanding, observed-dynamics understanding, and future prediction. Experiments show that SpatialMind substantially outperforms both general and spatially specialized models on our benchmark while achieving competitive zero-shot performance on VSI-Bench, OSI-Bench, and VLM4D.

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

Humans anticipate future changes in distances and spatial relations, enabling navigation and interaction with the physical world. In dynamic environments, future prediction is important for assessing whether an action can be completed safely. In Figure 1, a driver can instinctively anticipate the narrowing gap to the white car and assess whether sufficient clearance will remain during braking. Cognitive studies associate this capacity with *prospection*: using observed experience to reason beyond immediate perception and anticipate future changes (Craik, 1967; Buckner & Carroll, 2007; Schacter et al., 2007). While current VLMs have made substantial progress in understanding observed visual content, extending this understanding to physically grounded future spatial prediction remains challenging (Ishihara et al., 2026; Zhou et al., 2025; Fan et al., 2026). This raises a fundamental question: what spatial understanding does a VLM need for reliable future prediction?

Reliable future prediction depends on understanding the current scene and its observed evolution. To understand the current scene, a model first needs to ground entities in metric space and establish their spatial relations, which anchors future prediction in metric geometry. To determine how the scene has changed over the observed interval, it then needs to estimate motion and track changes in spatial relations, which reveals temporal trends that indicate how the scene may evolve. To support VLMs in extending their spatial reasoning from understanding observations to predicting future states, we organize spatial reasoning into three progressive levels: (i) *current state understanding*, which establishes metric geometry and spatial relations; (ii) *observed dynamics understanding*, which captures how the scene evolves throughout the observed interval; and (iii) *future prediction*, which extrapolates this evolution to anticipate subsequent spatial changes.

Existing VLMs and spatial datasets cover only parts of this progression, with future prediction receiving far less attention than reasoning over observed scenes. Recent models have advanced spatial reasoning (Cheng et al., 2024; 2026; Fan et al., 2026; Yang et al., 2026a; Yin et al., 2026), yet many focus on static geometry, explicitly prompted regions, or dynamics within the observed sequence, leaving prediction beyond the observations less explored. Training datasets that provide consistent supervision across the three levels for the same entities also remain limited. Existing benchmarks cover complementary aspects of metric geometry, observed motion, and short-horizon prediction (Li et al., 2025; Zhou et al., 2025; Wu et al., 2026c; Ishihara et al., 2026; Jia et al., 2026), but provide limited evaluation of the full progression from current-scene understanding through observed dynamics to future prediction. It therefore remains unclear whether models can use their understanding of current states and observed changes to anticipate future spatial conditions.

To enable and systematically evaluate future prediction in VLMs, we introduce **SpatialMind**, a spatial reasoning VLM designed to build and update a coherent spatial mind state from observation to prediction. To ground spatial reasoning at real-world scale, SpatialMind augments visual representations with metric depth and motion cues from a pretrained spatial expert. To model the spatial changes, we introduce a *progressive state chain* to structure the model's reasoning output across the three levels: grounding the referred entity and establishing its current spatial state, updating this state from observed dynamics, and extrapolating to the near future when required. This design connects metric grounding, temporal updating, and future reasoning within a single evolving state. To support training and evaluation, we develop a scalable data engine that grounds entity descriptions in metric geometry to generate QA pairs from both driving and everyday egocentric scenes. Using this engine, we construct SpatialMind-30K Dataset with 30,677 training QA pairs and SpatialMind-2K Benchmark with 2,000 evaluation questions. Both cover eight question categories across the same three levels: metric spatial states, observed ego and object dynamics, and future spatial relations.

Experiments on SpatialMind-2K Benchmark reveal persistent limitations in metric reasoning and future prediction among existing VLMs. SpatialMind substantially outperforms these models across all three reasoning levels, with consistent gains over the fine-tuned baseline model on the same training data. Our competitive zero-shot performance on VSI-Bench, OSI-Bench, and VLM4D further demonstrates its robust generalization capability across diverse spatial reasoning tasks.

Our main contributions are summarized as follows:

- We introduce SpatialMind, a method that significantly enhances metric distance estimation and spatiotemporal relation understanding, using observed dynamics to infer future spatial states and showing strong spatial reasoning and future prediction capabilities.

- We build a scalable data curation pipeline that generates geometrically grounded spatiotemporal QA pairs from both driving and everyday scenes, totaling 30,677 QA pairs with corresponding reasoning chains. We further introduce SpatialMind-2K Benchmark to evaluate VLM capabilities in metric-scale estimation, temporal reasoning, and future prediction.
- Experiments demonstrate that our SpatialMind model gains substantial improvement on SpatialMind-2K and competitive performance on VSI-Bench, OSI-Bench, and VLM4D.

2 Related Work

Spatial Reasoning in VLMs. Existing approaches to spatial reasoning in VLMs mainly incorporate external geometric priors through depth supervision, RGB-D inputs, or pretrained 3D features (Chen et al., 2024; Cai et al., 2025; Cheng et al., 2024; Fan et al., 2026; Zheng et al., 2026). Although these representations provide explicit spatial cues, their metric precision remains constrained by the scale accuracy of the underlying geometric estimates. Complementary work enables future prediction through future-event supervision, temporal event chains, or latent visual reasoning (Wang et al., 2026a; Su et al., 2026; Jiang et al., 2026). These methods primarily target future events, with less emphasis on explicitly estimating spatial quantities such as future distances in meters. SpatialMind integrates an absolute metric adapter with a progressive state chain to support spatial reasoning from current state, observed dynamics to future spatial prediction at real-world scale.

Spatial Reasoning Datasets and Benchmarks. Spatial reasoning benchmarks have expanded from static spatial relations to metric and spatiotemporal understanding (Ma et al., 2025; Jia et al., 2026; Yang et al., 2026b; Wu et al., 2026a). Existing evaluations emphasize spatial measurements, viewpoint-dependent relations, and observed motion (Yang et al., 2025; Zhang et al., 2026; Li et al., 2025; Zhou et al., 2025; Fan et al., 2026; Wu et al., 2026c). These benchmarks mainly assess geometry and dynamics within observed scenes, with limited coverage of spatial states beyond the observed interval. OmniSpatial and SCP-Bench extend evaluation to motion prediction and inference about unseen spatial states, respectively (Jia et al., 2026; Zhao et al., 2026). STRIDE-QA (Ishihara et al., 2026) further evaluates numerical predictions of future distances and velocities, but focuses on short-horizon prediction in urban driving. However, future spatial prediction at physical scale lacks systematic evaluation and training data. To fill this gap, SpatialMind offers a dataset and benchmark spanning eight question categories, from observed dynamics to future prediction.

3 SpatialMind Data Engine

SpatialMind supports metric-scale spatial reasoning and future prediction from video. Our data engine constructs metric geometry, establishes entity correspondences, and generates QA pairs with spatial metadata. Using this engine, we build a dataset and benchmark covering three reasoning levels and eight tasks. We further introduce the SpatialMind model in Sec. 4, which we train on this dataset and evaluate on the benchmark to assess its predictive spatial reasoning capabilities.

3.1 Target Selection and Unique Description

Constructing spatial questions involves selecting target objects and providing descriptions that identify each one uniquely. To this end, we follow three steps: target selection, description generation, and verification (Figure 2). To track objects over time, we apply a video instance segmentation model that segments individual objects and associates their regions across frames. To keep appearance details recognizable, we select targets that remain visible across consecutive frames and occupy sufficient image area. To capture the richest visual detail for each target, we take the frame in which it occupies the largest image area as the reference frame, crop the target region, and send the crop to a VLM, which describes the object's category, appearance, and distinctive attributes.

To verify that each description identifies its intended target uniquely, we perform reverse localization. Given the description and an unmarked frame from the original video, a VLM predicts a bounding box for the described object and checks whether any other object also matches the description. We retain a description only when the predicted box sufficiently overlaps the target's box and the VLM finds no other matching object. Together, the selected targets and their verified descriptions form the semantic basis for constructing spatial metadata in the next stage.

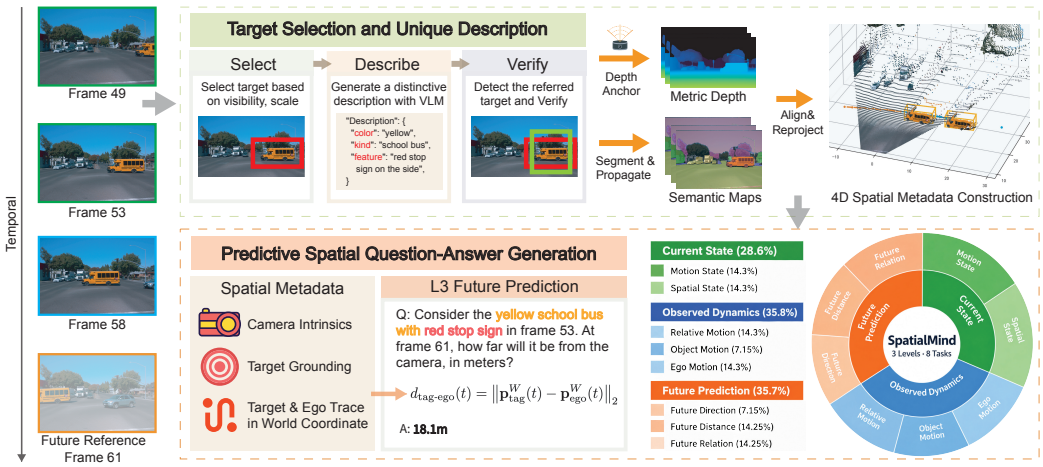


Figure 2: **SpatialMind Data Engine.** We assign each target a distinctive natural-language description through a Select–Describe–Verify process, then combine metric geometry, semantic propagation, and temporal alignment to construct spatial metadata, which supports automatic QA generation across three levels and eight spatial reasoning tasks, with the data distribution shown on the right.

3.2 4D Spatial Metadata Construction

Constructing spatial reasoning data depends on scene geometry at real-world scale. Sparse depth measurements from source data provide physical scale but have limited coverage on individual targets, which can introduce errors into the spatial measurements used for Question-Answer pairs generation. To obtain dense depth while preserving physical scale, we first predict dense depth with a feed-forward 3D model. We then use the sparse source measurements as metric references to estimate a scale and offset, calibrating the predicted depth to real-world scale. The calibrated depth provides dense geometry of visible surfaces, from which we measure entity positions and distances. To cover diverse scenes and spatial scales, we apply this procedure to outdoor driving videos from Waymo (Sun et al., 2020) and indoor egocentric videos from Ego-Exo4D (Grauman et al., 2024).

Predictive spatial reasoning questions need the states of targets and the observer over time, both for generating the questions and computing their reference answers. To localize each target from Sec. 3.1 in 3D, we unproject the calibrated depth within its mask into a shared world coordinate system using the camera intrinsics and poses. To recover each target's 4D trajectory, we propagate its mask across the video and link the resulting 3D localizations over time. To measure relative directions, we derive the observer's trajectory from the camera poses, taking the camera center at each frame as the observer's position and the camera orientation as its heading. The resulting spatial metadata comprises the targets' 4D trajectories, the observer's trajectory, and the verified referring descriptions that uniquely identify each target. This metadata serves as the computational basis for constructing spatial reasoning and future prediction questions.

3.3 Predictive Spatial QA Generation

We generate spatial QA pairs for eight tasks across three progressive levels: current-state understanding, observed-dynamics understanding, and future prediction. Current states describe entity positions and local motion, while observed changes provide evidence for future prediction. We compute reference answers from spatial metadata using each entity description, spatial reference, and temporal scope. Mathematically, we use t_s and t_o to denote the start and end of the observed interval, and $\mathbf{p}^W(t)$ to denote an entity's position in world coordinates at frame t.

Level 1: Current-state understanding describes where entities are and how they are moving at a given time. We design *spatial-state* questions about distances between entities and *motion-state* questions about speed and turning angle. At frame t, we compute the target–observer distance as

$$d(t) = \left\| \mathbf{p}^W_{\text{tag}}(t) - \mathbf{p}^W_{\text{ego}}(t) \right\|_2, \qquad t_s \le t \le t_o. \tag{1}$$

Here, $d(t)$ denotes the target–observer distance in meters, and $\|\cdot\|_2$ denotes the Euclidean norm. These questions jointly assess the model's understanding of current spatial relations and motion, which provides a foundation for reasoning about observed dynamics and future prediction.

Level 2: Observed-dynamics understanding measures how far entities travel and how spatial relations change over an observed interval. We design *ego-motion distance*, *target-motion distance*, and *relative-motion* questions. For the observer or target, the total distance traveled over $[t_s, t_o]$ is

$$s_u(t_s, t_o) = \sum_{t=t_s}^{t_o-1} \left\| \mathbf{p}_u^W(t+1) - \mathbf{p}_u^W(t) \right\|_2, \qquad u \in \{\text{ego}, \text{tag}\}. \tag{2}$$

Here, $s_u(t_s, t_o)$ denotes the path length of entity u over the observed interval, measured in meters. Ego-motion distance, target-motion distance, and relative-motion questions jointly assess the model's ability to track entity movement and changing spatial relations across frames while accounting for observer motion. Observed dynamics provide motion evidence for future spatial prediction. Appendix H illustrates a Level 2 relative-motion example, linking the observed video, metric geometry, and progressive state-chain supervision.

Level 3: Future prediction asks where entities will be and how entities will move beyond the observed interval. We split each source video into an observed prefix over $[t_s, t_o]$ and a withheld future segment. Using only the prefix as visual input, the model answers *ego future direction*, *target future distance*, and *target future relation* questions about the future segment. We compute reference answers from annotations of the future segment; for a distance query at t_f, the reference answer is

$$d(t_f) = \left\| \mathbf{p}_{\text{tag}}^W(t_f) - \mathbf{p}_{\text{ego}}^W(t_f) \right\|_2, \qquad t_f > t_o. \tag{3}$$

Here, $d(t_f)$ denotes the future target–observer distance. These questions extend spatial reasoning from observation to prediction, assessing the model's ability to integrate the spatial and temporal understanding developed in the preceding levels into coherent predictions of future spatial states.

Dataset and Benchmark. We construct SpatialMind-30K Dataset with 30,677 training QA pairs and spatial metadata, and SpatialMind-2K Benchmark with 2,000 questions. Figure 2 shows the benchmark's task distribution. We review benchmark questions and answers through a human-in-the-loop process, using geometric and temporal computation records to correct errors. Across three progressive levels and eight task categories, the dataset and benchmark support systematic training and evaluation of predictive spatial reasoning at real-world scale. Please refer to Appendix A for further details on the data engine and Appendix B for the mathematical introduction of all tasks.

4 SPATIALMIND MODEL

Building on the data set and the benchmark constructed in Sect. 3, we introduce SpatialMind, a VLM with an absolute metric depth adapter for real-world scale reasoning and a progressive state chain that guides the model to establish current spatial states, characterize observed dynamics, and predict future spatial conditions, grounding the final answer in a sequence of explicit state estimates.

4.1 METRIC DEPTH ADAPTATION AND SPATIOTEMPORAL FUSION

To reason about distances and changes in motion, a VLM needs to estimate scene geometry at real-world scale from visual observations. However, scale errors in pretrained geometric predictions can affect distance estimates and subsequent reasoning. We introduce an absolute metric depth adapter to refine metric depth, followed by spatiotemporal fusion to integrate geometry over time.

Absolute Metric Depth Adapter. As shown in Figure 3, the adapter combines initial metric depth D_0 from a pretrained depth estimator with multiframe features, depth maps, and confidence maps from a pretrained multiview geometry model. It predicts a global scale correction δ_{g} for each frame and a local depth correction map Δ_{l} for individual pixels. Applying both corrections to D_0 in logarithmic depth space yields the refined metric depth prediction D_{m}:

$$D_{\text{m}} = D_0 \odot \exp\left(\gamma_{\text{g}} \tanh \delta_{\text{g}} + \mathcal{U}\left(\gamma_{\text{l}} \tanh \Delta_{\text{l}}\right)\right), \tag{4}$$

where $\mathcal{U}$ upsamples the local correction to the depth resolution, and $\odot$ denotes multiplication at corresponding pixels. The coefficients γ_{g} and γ_{l} bound the corrections to limit deviations from

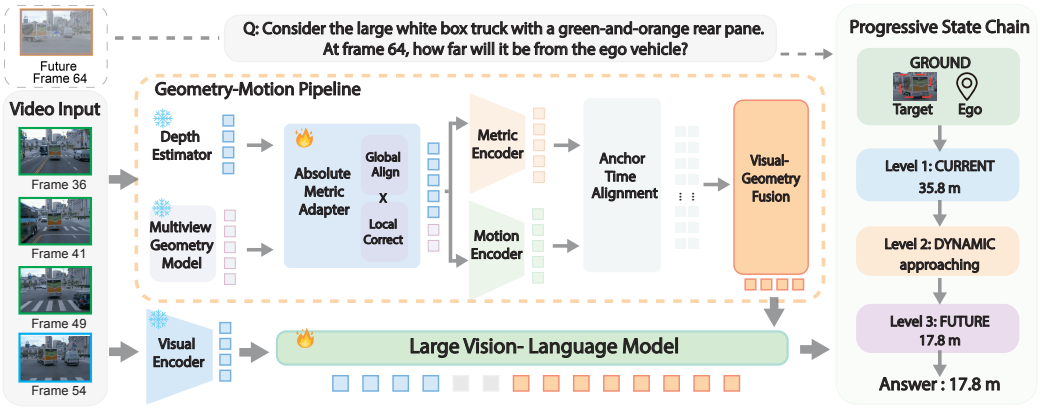


Figure 3: **Model Architecture.** Given an observed RGB video and a spatial question, SpatialMind refines metric depth with multiview geometry through global and local corrections. We separately encode metric geometry and camera motion, align their features temporally, and integrate them into visual tokens through cross-attention. The VLM then generates a progressive state chain that grounds the target, establishes its current spatial state, and extends to observed dynamics and future prediction according to the question level before producing the final answer.

D_0. We train the adapter with dense metric depth supervision while keeping both pretrained models frozen, then freeze it during VLM training. Further details are provided in Appendix C.

Spatiotemporal Fusion. The Metric Encoder processes geometry derived from D_{m}, while the Motion Encoder processes relative camera changes from the multiview model. To align these cues with video content, we extract them from the same sampled RGB frames and use their frame times for Anchor Time Alignment, as shown in Figure 3. Cross attention then fuses the encoded geometry and motion into visual tokens, providing spatial and motion cues for the progressive state chain.

4.2 Progressive State Chain

To predict future spatial states, a VLM needs to relate current states to observed motion. However, final answer supervision does not explicitly constrain these intermediate estimates. We introduce a progressive state chain to supervise these estimates according to each question's temporal scope.

As shown in Figure 3, we train the same VLM to generate chains of different lengths for the three task levels. Level 1 includes `GROUND` to identify the target and spatial reference, and `CURRENT` to describe spatial and motion states. Level 2 adds `DYNAMICS` for observed changes, while Level 3 adds `FUTURE` for future predictions. Each chain ends with `ANSWER` to provide the final response. The model generates each state section conditioned on the video, question, and preceding sections.

We derive supervision for each state section from spatial metadata in Sec. 3. Current states and observed dynamics use annotations from the observed interval. For future questions, annotations from the withheld interval provide future state targets, while the video input is the observed prefix.

Training and Inference. We train the VLM with LoRA on state chain and direct answer examples from the same QA pairs. Let X denote the fused video representation, q the question, a the answer, and $S_{1:m}$ the state sections before `ANSWER`, with $m = 2, 3, 4$ for Levels 1, 2, and 3. The losses are

$$\mathcal{L}_{\mathrm{S}} = \frac{1}{m}\sum_{k=1}^{m} \ell_\theta(S_k \mid X, q, S_{<k}), \qquad \mathcal{L}_{\mathrm{A}} = \ell_\theta(a \mid X, q, S_{1:m}). \tag{5}$$

Here, ℓ_θ averages token cross entropy over each target sequence, with θ denoting VLM parameters and $S_{<k}$ the preceding state sections. Averaging over sections gives them equal weight regardless of length. The direct answer loss $\mathcal{L}_{\mathrm{D}}$ uses the same averaging without a state chain. The total loss is

$$\mathcal{L} = \rho\mathcal{L}_{\mathrm{D}} + (1-\rho)\left[(1-\lambda)\mathcal{L}_{\mathrm{A}} + \lambda\mathcal{L}_{\mathrm{S}}\right]. \tag{6}$$

Here, ρ balances direct answer and chain supervision, while λ balances state and answer losses within chain examples. At inference, SpatialMind receives only the observed RGB video and question, then generates the state sections for the question's task level, followed by the final answer.

Table 1: **Results on SpatialMind-Benchmark (%).** Tasks are grouped by reasoning level. Among evaluations within each group, **bold** marks the best result and underline marks the second best.

		L1: Current state		L2: Observed dynamics			L3: Future prediction		
Model	Overall	Motion	Spatial	Ego	Object	Relative	Direc.	Dist.	Relat.
Proprietary Models (API)									
GPT-5.6 Luna	22.60	20.28	14.34	12.59	25.87	44.76	27.27	14.04	25.61
Claude Sonnet 5	27.80	16.43	**23.08**	9.79	26.57	49.30	30.07	23.51	44.21
Gemini 3.8 Flash	30.19	21.62	18.66	19.39	49.65	52.85	20.28	**41.34**	22.91
GPT-6-Astra (medium)	**39.00**	**38.46**	13.99	**19.58**	**51.75**	**73.78**	**39.86**	29.12	**52.28**
General-purpose VLMs									
Qwen2.5-VL-3B (Bai et al., 2025b)	14.40	5.95	5.25	**10.85**	2.80	32.20	28.70	17.85	13.00
Qwen3.5-9B (Qwen Team, 2026)	10.30	**10.85**	9.45	5.60	13.30	9.45	4.20	20.35	8.05
LLaVA-NeXT-7B (Zhang et al., 2024)	14.40	5.25	6.65	5.95	7.00	30.10	31.50	15.05	18.60
Qwen2.5-VL-7B (Bai et al., 2025b)	17.60	3.85	**17.15**	2.45	30.80	33.60	25.90	10.15	27.70
Cosmos3-8B (Agarwal et al., 2026)	18.95	5.60	2.40	9.80	**32.90**	35.30	**32.20**	3.90	**43.20**
Qwen3-VL-8B (Bai et al., 2025a)	**20.20**	3.15	13.30	7.35	23.10	**38.50**	24.50	**21.00**	34.70
Spatially specialized VLMs									
SR-3D-8B (Cheng et al., 2026)	9.60	2.45	6.30	12.60	2.10	24.85	14.70	8.75	3.50
SpatialRGPT-8B (Cheng et al., 2024)	9.70	6.65	5.25	5.25	2.80	17.85	**34.30**	3.85	10.20
SpatialBot-3B (Cai et al., 2025)	11.80	2.80	7.70	1.05	**34.30**	23.45	28.70	2.80	13.00
VLM-3R-7B (Fan et al., 2026)	14.00	**7.90**	1.05	4.20	7.70	31.50	29.40	2.45	**32.65**
SpaceR-7B (Ouyang et al., 2025)	14.70	5.60	**14.70**	5.25	14.70	33.25	21.70	3.50	22.45
ViLaSR-7B (Wu et al., 2026b)	14.80	7.70	9.10	7.00	7.00	32.20	16.80	6.00	30.15
SR-ReaL-8B (Ji et al., 2026)	14.90	5.95	11.55	11.90	8.40	31.50	32.20	7.05	16.15
4D-RGPT-8B (Yang et al., 2026a)	15.40	3.50	11.90	**12.95**	7.00	**33.95**	27.30	**12.25**	16.15
SpatialReasoner-7B (Ma et al., 2026)	**15.80**	7.20	7.70	5.60	12.60	33.25	30.10	8.40	27.00
SFT on SpatialMind training data									
Qwen3-VL-8B + SFT	45.70	36.02	57.34	23.77	37.06	56.99	57.34	40.64	57.89
Ours	**52.65**	**38.32**	**65.46**	**30.15**	**52.34**	**68.42**	**65.13**	**42.89**	**64.48**

5 EXPERIMENTS

Baselines and Setup. To evaluate predictive spatial reasoning, we compare SpatialMind with proprietary VLMs, including GPT, Claude, and Gemini models; general-purpose models such as Qwen3-VL (Bai et al., 2025a) and Cosmos 3 (Agarwal et al., 2026); and spatially specialized models such as SpatialRGPT (Cheng et al., 2024) and VLM-3R (Fan et al., 2026). We also fine-tune Qwen3-VL-8B on SpatialMind-30K as an SFT baseline using the same backbone and training data. For future-prediction tasks, all models receive only the observed video prefix and question.

Evaluation Metrics. SpatialMind-2K Benchmark contains three multiple-choice categories and five numerical fill-in-the-blank categories. We evaluate multiple-choice questions on relative motion, ego future direction, and target future relation using accuracy (ACC). The numerical categories cover spatial state, motion state, ego-motion distance, target-motion distance, and target future distance. We evaluate numerical answers, such as absolute distances, using mean relative accuracy (MRA), following prior spatial reasoning evaluations (Yang et al., 2025; Fan et al., 2026).

5.1 COMPARISON ON SPATIALMIND-BENCHMARK

We compare SpatialMind with models in four categories: proprietary models, general-purpose open-source VLMs, spatially specialized VLMs, and a baseline fine-tuned on SpatialMind-Dataset. SpatialMind-Benchmark evaluates these models across three progressive levels, from understanding current spatial states to reasoning about observed dynamics and predicting future states.

Proprietary Models. We first evaluate proprietary models through their APIs, including GPT-5.6 Luna, Claude Sonnet 5, Gemini 3.8 Flash, and GPT-6-Astra. As shown in Table 1, these models achieve higher overall accuracy than the evaluated general-purpose and spatially specialized VLMs. All four achieve their highest task accuracy on observed relative motion, while current-state distance

and ego-motion distance estimation remain less accurate. These results suggest that stronger overall performance can coexist with difficulties in estimating distances at real-world scale.

General-Purpose VLMs. We evaluate representative general-purpose VLMs, including Qwen3-VL (Bai et al., 2025a), LLaVA (Zhang et al., 2024), and Cosmos 3 (Agarwal et al., 2026), which support broad vision-language understanding. As shown in Table 1, these models tend to achieve higher accuracy on categorical questions involving directions or spatial relations than on tasks requiring direct estimation of metric quantities such as distance, displacement, and speed. Metric estimation remains challenging in Level 1, when establishing the current spatial state, and Level 2, particularly when measuring ego-motion distance. We further observe that some models retain relatively stronger performance on categorical future-prediction questions despite less accurate metric estimation at the preceding levels. The results suggest that stronger performance on categorical future questions does not necessarily imply accurate metric understanding of the observed scene.

Spatially Specialized VLMs. We further evaluate VLMs specialized for spatial reasoning, many of which incorporate depth or 3D geometric priors to enhance spatial perception. As shown in Table 1, these models remain limited on metric tasks, particularly Level 1, ego-motion distance in Level 2, and future-distance prediction in Level 3. SpatialMind-Benchmark spans outdoor driving scenes and everyday egocentric environments, requiring models to estimate distances and motion across different spatial scales. The results suggest that incorporating geometric priors alone does not ensure accurate metric reasoning across the diverse scenes covered by our benchmark.

SpatialMind. We first fine-tune Qwen3-VL-8B on SpatialMind-30K dataset to establish an SFT baseline. Compared with zero-shot Qwen, this baseline improves performance across all eight tasks, with substantial gains in metric estimation and motion reasoning, suggesting the benefit of supervision grounded in linguistic entities, metric geometry, and temporal computations. We then train our SpatialMind model on the same dataset and observe further improvements across all tasks, increasing overall accuracy from 45.70% to 52.65%. These additional gains support the combined benefit of incorporating metric geometry and structuring inference through the progressive state chain, which establishes the current spatial state, summarizes observed dynamics, and predicts future states.

5.2 Generalization to Spatial Benchmarks

To examine the generalization of SpatialMind beyond our benchmark, we conduct zero-shot evaluation on three complementary spatial reasoning benchmarks. VSI-Bench (Yang et al., 2025) evaluates video-based spatial understanding, with emphasis on metric measurement and spatiotemporal reasoning; OSI-Bench (Wu et al., 2026c) focuses on open-world spatial inference in both static and dynamic metric settings; and VLM4D (Zhou et al., 2025) evaluates 4D spatial-temporal understanding over real and synthetic videos. Together, these benchmarks cover spatial measurement, dynamic reasoning, and motion understanding across diverse scenes and evaluation protocols.

VSI-Bench. As shown in Table 2, SpatialMind achieves the highest measurement and spatiotemporal scores among the compared models, reaching 52.1 and 52.4. These exceed the strongest competing scores by 4.4 and 3.8 points, respectively, while the overall score of 46.4 remains competitive. The results support the transfer of metric estimation and spatiotemporal reasoning to indoor scenes.

OSI-Bench. SpatialMind achieves the highest overall score of 35.0 among the compared models, surpassing Qwen3-VL-32B by 2.8 points. It leads in static metric reasoning with 30.2 and obtains a competitive dynamic metric score of 33.4, compared with 34.9 for Qwen3-VL-32B. These results support the transfer of static and dynamic metric reasoning to open-world environments.

VLM4D. SpatialMind achieves an overall score of 53.4, exceeding 4D-RGPT by 1.0 points, and obtains the highest scores among the compared models on both real and synthetic videos, reaching 53.5 and 52.9, respectively. The gains across both splits provide further evidence that SpatialMind generalizes to motion and perspective understanding across different visual environments.

5.3 Ablation Studies

We examine the metric depth adaptation pipeline and progressive state chain introduced in Sec. 4 to assess their contributions to SpatialMind. We compare geometric inputs and reasoning formats across the three reasoning levels to evaluate their effects on predictive spatial reasoning.

Table 2: **Zero-shot evaluation on external benchmarks.** VSI-Bench reports measurement (Meas.) and spatiotemporal (S-T) reasoning; OSI-Bench reports static and dynamic reasoning; VLM4D reports the real and synthetic splits. Our method achieves competitive results across three benchmarks.

VSI-Bench			
Model	Avg.	Meas.	S-T
GPT-4o	34.0	29.1	28.5
Gemini-1.5-Pro	45.4	46.2	34.6
InternVL2-8B	37.5	40.7	46.4
LLaVA-NeXT-Video-72B	40.9	38.5	48.6
LLaVA-OneVision-72B	40.2	39.7	44.6
Spatial-MLLM-4B	**48.4**	47.7	46.3
Ours	46.4	**52.1**	**52.4**

OSI-Bench			
Model	Avg.	Static	Dyn.
GPT-5	29.7	25.7	25.0
Claude-3.7-Sonnet	26.5	28.4	13.4
InternVL3.5-8B	28.5	22.6	24.0
Ovis2-16B	28.1	22.6	24.4
Qwen3-VL-8B	31.2	23.1	31.1
Qwen3-VL-32B	32.2	22.3	**34.9**
Ours	**35.0**	**30.2**	33.4

VLM4D			
Model	Overall	Real	Synth.
Qwen3-VL-8B	52.2	52.1	52.4
Claude-Sonnet-4	51.3	52.2	48.3
InternVL2.5-8B	42.0	42.4	40.9
VG-LLM-7B	46.5	49.5	37.3
VLM-3R-7B	33.9	36.9	24.7
4D-RGPT-8B	52.4	53.1	50.6
Ours	**53.4**	**53.5**	**52.9**

Table 3: **Geometric ablation** with the progressive state chain used in all variants.

Variant	Overall	Level 1	Level 2	Level 3
RGB only	50.42	48.96	46.31	54.42
+ DA3	51.30	49.48	48.13	55.59
+ Ω	51.20	50.17	46.62	55.71
Full model	**52.65**	**51.89**	**50.30**	**57.47**

Table 4: **Progressive state-chain ablation** with absolute metric adapter retained in all variants.

Variant	Overall	Level 1	Level 2	Level 3
Direct answer	49.95	48.43	48.25	55.22
Free-form CoT	34.15	33.05	35.57	33.73
w/o grounding	51.55	49.48	46.74	**57.58**
Full state chain	**52.65**	**51.89**	**50.30**	57.47

Geometric Inputs and Metric Adaptation. We evaluate the metric depth adaptation pipeline in Sec. 4.1 by comparing different geometric inputs and their effects on spatial reasoning accuracy. The *RGB-only* variant uses visual features without geometric conditioning. The *DA3* variant replaces our metric depth adaptation pipeline with depth predicted directly by DA3 (Lin et al., 2025). The *VGGT-Ω* variant independently replaces the same pipeline with depth predicted by VGGT-Ω (Wang et al., 2026b). The full model uses the adapted metric depth produced by our pipeline, and all variants retain the progressive state chain. As shown in Table 3, both pretrained-depth alternatives provide modest gains over RGB-only input, while the full model achieves the highest accuracy overall and across all three reasoning levels. The largest improvement over RGB-only input occurs in Level 2 observed-dynamics reasoning, which requires estimating spatial changes across frames. These results support the effectiveness of our complete metric depth adaptation pipeline.

Progressive State-Chain Reasoning. We evaluate the progressive state chain by comparing different output formats and their effects on spatial reasoning accuracy. The *direct-answer* variant trains the model to output only the final answer, without an intermediate state chain. For *free-form CoT*, we prompt the original Qwen3-VL model to follow a `detect--think--answer` format without task-specific fine-tuning. The *w/o grounding* variant removes the initial `GROUND` step, which interprets the question by identifying the target, spatial reference, and quantity to estimate. The direct-answer, w/o grounding, and full-chain variants retain the same geometric pipeline. As shown in Table 4, the full chain improves accuracy over direct-answer prediction across all three reasoning levels and also exceeds the prompt-based CoT reference. Removing `GROUND` lowers overall accuracy and performance on Levels 1 and 2, while yielding a slightly higher Level 3 score. These results support the benefit of state-chain supervision for predictive spatial reasoning.

6 Conclusion

We introduced SpatialMind, a spatial reasoning VLM that extends scene understanding to physically grounded future prediction. We formulate spatial reasoning as a progressive process that establishes current spatial states, captures observed dynamics, and uses both to anticipate future states beyond the observed interval. To support this progression, we develop a scalable data engine and construct SpatialMind-30K Dataset and SpatialMind-2K Benchmark with metric geometric and temporal supervision. SpatialMind combines an absolute metric depth adapter with a progressive state chain to ground reasoning at real-world scale and structure inference from observation to prediction. Experiments demonstrate consistent gains across all three reasoning levels and competitive zero-shot performance on external benchmarks, while ablations support the contributions of both components. These findings highlight the value of metric grounding and observed-dynamics understanding for enabling VLMs to anticipate future spatial changes in dynamic environments.

AI USE STATEMENT

Generative AI assistants were used only to polish the manuscript for clarity, grammar, concision, and presentation. Proprietary generative AI models were evaluated solely as baselines on our benchmark and were not used as components of our method. All AI-assisted edits were reviewed by the authors, who take responsibility for the final content of the paper.

ETHICS STATEMENT

This work uses the Waymo Open Dataset and Ego-Exo4D under their applicable access and usage terms. Our dataset and benchmark may inherit biases and privacy limitations from the source data. SpatialMind is intended for research purposes only, supporting the study of spatial reasoning and future prediction in vision-language models.

REPRODUCIBILITY STATEMENT

We provide detailed information required to reproduce our method and experiments in the appendix, including the data construction pipeline, task definitions and geometric computations, model implementation and training details, dataset composition, and evaluation protocol. Additional worked examples illustrate the construction of reference answers and progressive state-chain supervision. We will release the code, dataset, and evaluation tools to facilitate reproduction and further research.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Randy L. Buckner and Daniel C. Carroll. Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2):49–57, 2007. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2006.11.004. URL https://www.sciencedirect.com/science/article/pii/S1364661306003275.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9490–9498. IEEE, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, 2024. URL https://arxiv.org/abs/2401.12168.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

An-Chieh Cheng, Yang Fu, Yukang Chen, Zhijian Liu, Xiaolong Li, Subhashree Radhakrishnan, Song Han, Yao Lu, Jan Kautz, Pavlo Molchanov, et al. 3d aware region prompted vision language

model. In *International Conference on Learning Representations*, volume 2026, pp. 123286–123308, 2026.

Kenneth James Williams Craik. *The nature of explanation*, volume 445. CUP Archive, 1967.

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.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

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.

Yatai Ji, An-Chieh Cheng, Yang Fu, Yukang Chen, Han Zhang, Zhaojing Yang, Wei Huang, Ka Chun Cheung, Song Han, Vidya Nariyambut Murali, et al. Reinforcing dual-path reasoning in spatial vision language models. *arXiv preprint arXiv:2606.17539*, 2026.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *International Conference on Learning Representations*, volume 2026, pp. 35634–35670, 2026.

Tianxiang Jiang, Linquan Wu, Sheng Xia, Songze Li, Ziang Yan, Haoyu Yang, Yu Qiao, and Yi Wang. Imagine before you predict: Interleaved latent visual reasoning for video event prediction. *arXiv preprint arXiv:2606.05769*, 2026.

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, 2025.

Haotong Lin, Sili Chen, Junhao 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*, 2025.

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, 2025.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *Advances in Neural Information Processing Systems*, 38:140751–140774, 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. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Daniel L Schacter, Donna Rose Addis, and Randy L Buckner. Remembering the past to imagine the future: the prospective brain. *Nature reviews neuroscience*, 8(9):657–661, 2007.

Qile Su, Jing Tang, Rui Chen, Lei Sun, and Xiangxiang Chu. Video-coe: Reinforcing video event prediction via chain of events. *arXiv preprint arXiv:2603.14935*, 2026.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451. IEEE, 2020.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction. In *International Conference on Learning Representations*, volume 2026, pp. 31524–31570, 2026a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026b.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. Spatialscore: Towards comprehensive evaluation for spatial intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31029–31041, 2026a.

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. *Advances in Neural Information Processing Systems*, 38:143297–143330, 2026b.

Mingrui Wu, Zhaozhi Wang, Fangjinhua Wang, Jiaolong Yang, Marc Pollefeys, and Tong Zhang. From indoor to open world: Revealing the spatial reasoning gap in mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16789–16799, 2026c.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31042–31053, 2026a.

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, 2025.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. In *International Conference on Learning Representations*, volume 2026, pp. 157051–157088, 2026b.

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.

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026.

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yanguang Zhao, Jie Yang, Shengqiong Wu, Shutong Hu, Hongbo Qiu, Yu Wang, Guijia Zhang, Tan Kai Ze, Hao Fei, Chia-Wen Lin, et al. Scp: Spatial causal prediction in video. *arXiv preprint arXiv:2603.03944*, 2026.

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 2026.

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, 2025.

Supplementary Material

This supplementary material provides further details on the SpatialMind data engine, geometric computations for the eight task families, and the metric depth adapter and temporal fusion. We also document the dataset composition and evaluation metrics, analyze the persistence of current states in future-prediction questions, and discuss related spatiotemporal tasks and limitations. Finally, worked examples and visualizations illustrate how video observations and geometric annotations are used to construct reference answers and progressive state-chain supervision.

Contents

A Data Engine Details 13
B Task Definitions and Geometric Computations 14
C Metric Geometry and Temporal Fusion 16
D Dataset Composition and Evaluation 17
E Additional Analysis 18
F Relationship to Other Spatiotemporal Tasks 19
G Limitations and Future Work 20
H Visualizations 20

A Data Engine Details

A.1 Source Geometry and Correspondences

Waymo (Sun et al., 2020) supplies sparse LiDAR measurements, calibrated cameras, observer poses, and tracked 3D boxes. For Ego-Exo4D (Grauman et al., 2024), Aria semidense landmarks and their image observations provide sparse metric references, together with the device trajectory and camera calibration. The references are aligned with the RGB images by composing the relevant sensor transformations. Source track identifiers establish correspondences when available; otherwise, SAM3 (Carion et al., 2026) video segmentation propagates instance identities. Following Section 3.2, calibrated depth within each selected target mask is back-projected using camera intrinsics and poses into the shared world frame W. The resulting per-frame 3D localizations are associated over time to obtain target trajectories. The observer's position is the camera centre, and its heading is given by the camera orientation.

A.2 Dense-depth Calibration

DA3-Large (Lin et al., 2025), the Large variant of Depth Anything 3 (DA3), predicts dense depth, and sparse metric references determine its scale and offset. At frame t, a reference pixel $\mathbf{x}_i$ pairs the predicted depth $\widehat{D}_t(\mathbf{x}_i)$ with the metric measurement $z_{t,i}$. We fit the affine relation $z_{t,i} \simeq a_t \widehat{D}_t(\mathbf{x}_i) + b_t$ using valid, positive pairs, reducing the influence of inconsistent references through robust weighting. The fitted coefficients are stabilized across frames. Denoting the resulting coefficients by $\bar{a}_t$ and $\bar{b}_t$, the calibrated depth is

$$D_t(\mathbf{x}) = \max\{0, \bar{a}_t \widehat{D}_t(\mathbf{x}) + \bar{b}_t\}. \tag{7}$$

Here D_t and the offset $\bar{b}_t$ are expressed in metres. Valid calibrated depths provide the visible-surface geometry used with instance masks, while sparse references support calibration checks. This procedure supplies the metric geometry described in Section 3.2.

A.3 Referring-expression Construction

Following Section 3.1, we select sufficiently visible targets and use Qwen3-VL-32B (Bai et al., 2025a) to describe their category, appearance, and distinguishing attributes. Image location may disambiguate visually similar instances. Explicit metric distances, motion states, and future relations are excluded from the descriptions. Reverse localization then uses the description and an unmarked frame to identify the target and check alternative matches. Incorrect or ambiguous descriptions are

revised or discarded. Accepted descriptions remain linked to persistent entity identifiers, connecting the question's referent to its geometry throughout the video.

B Task Definitions and Geometric Computations

The eight task families follow Section 3.3. We retain its notation: $\mathbf{p}_u^W(t)$ is the position of entity u in metres in the shared world frame W, with $u = \text{ego}$ for the observer and $u = \text{tag}$ for the target. The variables t, t_s, t_o, t_f denote integer frame indices, not elapsed seconds. A question may specify a subinterval $[t_1, t_2] \subseteq [t_s, t_o]$ of the observed video. We write τ_t for the corresponding timestamp in seconds and require increasing timestamps over a queried motion interval. Future queries satisfy $t_f > t_o$. Turning-angle answers are reported in degrees; angular calculations below state explicitly when radians are used.

B.1 Level 1: Current-state Understanding

Spatial state. This task asks about metric separation at an observed frame. Target–observer distance is $d(t)$ in Equation 1. For two queried entities u and v, the corresponding point-to-point distance is

$$d_{u,v}(t) = \|\mathbf{p}_u^W(t) - \mathbf{p}_v^W(t)\|_2, \qquad t_s \le t \le t_o. \tag{8}$$

Both positions refer to the same frame and world coordinate system, and the distance is measured in metres. For target–observer queries, $\mathbf{p}_{\text{ego}}^W(t)$ is the camera centre defined in Section 3.2. Applying the same rigid coordinate transformation to both positions preserves this distance. Figure 9 illustrates an explicitly requested visible-surface quantity: the median camera-to-surface range over valid target-mask samples. This scalar surface statistic is distinct from the point-to-point distance defined above.

Motion state. This task asks about speed or turning angle over a specified local interval. With $t_1 < t_2$ and $\tau_{t_2} > \tau_{t_1}$, trajectory-based mean speed is

$$\bar{v}_u(t_1, t_2) = \frac{s_u(t_1, t_2)}{\tau_{t_2} - \tau_{t_1}}, \qquad u \in \{\text{ego}, \text{tag}\}, \tag{9}$$

where s_u is the accumulated path length defined in Equation 2; thus $\bar{v}_u$ is in metres per second. Frame counts alone do not determine elapsed time when timestamps are nonuniform. For questions defined from sampled camera speeds, the answer is the arithmetic mean of the supplied speed samples over the queried interval (Figure 8); the serialized measurement convention specifies which construction applies. For view-turn questions, let $\mathbf{f}_t^W$ be the unit optical forward direction expressed in the same world frame at every time. The unsigned angle between the endpoint viewing directions is

$$\theta = \frac{180^\circ}{\pi} \arccos\Big(\text{clip}_{[-1,1]}\big((\mathbf{f}_{t_1}^W)^\top \mathbf{f}_{t_2}^W\big)\Big). \tag{10}$$

The arccosine returns radians, so the conversion gives $\theta \in [0^\circ, 180^\circ]$. This endpoint viewing-angle change does not encode a left/right sign or cumulative rotation.

B.2 Level 2: Observed-dynamics Understanding

Ego-motion distance. The answer is $s_{\text{ego}}(t_1, t_2)$ from Equation 2, evaluated over the queried observation interval. It accumulates the observer's travel along the trajectory; an interval with identical endpoints can still have nonzero path length.

Target-motion distance. The answer is $s_{\text{tag}}(t_1, t_2)$ from the same equation, using the same target identity throughout the interval. Measuring positions in W separates target motion from image changes caused by camera motion. Both motion-distance tasks return a numerical answer in metres. The ground-route example in Figure 10 applies the segment-length sum to its recorded ground-plane positions; this planar measurement is distinct from a general three-dimensional trajectory length.

Relative motion. This task asks whether the target becomes closer, farther, leftward, rightward, or approximately unchanged relative to the observer. The signed change in metric separation is

$$\Delta d = d(t_2) - d(t_1), \tag{11}$$

with $d(t)$ defined by Equation 1. Negative and positive values indicate decreasing and increasing separation, respectively. In the observer's optical frame, whose axes point right, down, and forward, a target at (X_t, Y_t, Z_t) with $Z_t > 0$ has right-positive bearing $\alpha(t) = \operatorname{atan2}(X_t, Z_t)$ in radians. For these forward-facing observations, the bearing lies in $(-\pi/2, \pi/2)$, and

$$\Delta\alpha = \alpha(t_2) - \alpha(t_1) \tag{12}$$

is negative for leftward change and positive for rightward change. The categorical question identifies the dominant stable change, or approximately unchanged relations when both changes are small. Visibility, noise, and trend consistency checks exclude ambiguous candidates. For driving relative-motion questions, the range is normalized using every sampled frame in the queried interval. With $N = t_2 - t_1 + 1$, define

$$\begin{aligned} \bar{d} &= \frac{1}{N}\sum_{t=t_1}^{t_2} d(t), \\ R &= \frac{|\Delta d|}{\max(\bar{d}, 10^{-6}\,\mathrm{m})}, \qquad L = |\Delta\alpha|, \qquad T = \sqrt{R^2 + L^2}. \end{aligned} \tag{13}$$

Here the angular difference is in radians, so both normalized terms are dimensionless. The driving rule first assigns approximately unchanged motion when $T < 0.08$. Otherwise, after checking trend consistency, $R \geq 1.5L$ selects the radial class and $L \geq 1.5R$ selects the lateral class; the sign of the corresponding change determines farther/closer or right/left. Candidates without a dominant stable change are excluded. Figure 6 illustrates this computation.

B.3 Level 3: Future Prediction

Only the observed prefix is supplied as video input. Future annotations provide the reference quantities for prediction tasks.

Ego future direction. This task classifies the observer's subsequent manoeuvre over a designated interval $[t_a, t_f]$, where $t_o \leq t_a < t_f$. Let ψ_t be the observer's heading in radians, with positive change corresponding to a left turn. The accumulated signed turn, expressed in degrees, is

$$\Theta = \frac{180^\circ}{\pi}\sum_{t=t_a}^{t_f-1} \mathrm{wrap}_{[-\pi,\pi)}(\psi_{t+1} - \psi_t), \tag{14}$$

where $\mathrm{wrap}_{[-\pi,\pi)}$ chooses the equivalent angle in $[-\pi, \pi)$. For the driving questions, terminal speed uses the final five sampled frame intervals within the future interval:

$$v_{\mathrm{end}} = \frac{s_{\mathrm{ego}}(t_f - 5, t_f)}{\tau_{t_f} - \tau_{t_f-5}}, \qquad t_f - t_a \geq 5. \tag{15}$$

Here the same path-length sum is applied to annotated future positions. The timestamps in the denominator are in seconds and strictly increase. Using $v_0 = 0.5\,\mathrm{m/s}$, the mutually exclusive label rule is

$$y = \begin{cases} \text{come to a stop}, & v_{\mathrm{end}} < v_0, \\ \text{turn left}, & v_{\mathrm{end}} \geq v_0,\ \Theta \geq 12^\circ, \\ \text{turn right}, & v_{\mathrm{end}} \geq v_0,\ \Theta \leq -12^\circ, \\ \text{go straight}, & v_{\mathrm{end}} \geq v_0,\ |\Theta| < 6^\circ. \end{cases} \tag{16}$$

Stopping takes priority. Non-stopping candidates with $6^\circ \leq |\Theta| < 12^\circ$ are excluded. The left-positive heading convention here is distinct from the right-positive bearing convention for target-relative motion.

Target future distance. The answer is $d(t_f)$ from Equation 3, with both the target and observer positions evaluated at the same future frame $t_f > t_o$. The model predicts this distance in metres from the observed prefix.

Target future relation. This task predicts the target's horizontal relation to the observer at t_f. Let $\mathbf{c}^W(t_f) = \mathbf{p}^W_{\mathrm{ego}}(t_f)$ be the future optical camera centre and let $R_{W\to C}(t_f)$ rotate world vectors into

the optical camera frame, whose axes point right, down, and forward. Camera poses and calibrated sensor extrinsics determine these quantities. Project the target's future position as

$$(X_f, Y_f, Z_f)^\top = R_{W\to C}(t_f)\big(\mathbf{p}^W_{\mathrm{tag}}(t_f) - \mathbf{c}^W(t_f)\big),$$
$$\xi_f = \frac{f_x X_f/Z_f + c_x}{w_{\mathrm{img}}}, \qquad Z_f > 0, \tag{17}$$

where f_x and c_x are the horizontal focal length and principal point in pixels, and w_{img} is the image width in pixels. The coordinate ξ_f is not clipped to the image. For a valid forward projection, the horizontal-band label is

$$B(\xi_f) = \begin{cases} \text{out of view, left}, & \xi_f < 0, \\ \text{front left}, & 0 \le \xi_f < 0.4, \\ \text{straight ahead}, & 0.4 \le \xi_f < 0.6, \\ \text{front right}, & 0.6 \le \xi_f \le 1, \\ \text{out of view, right}, & \xi_f > 1. \end{cases} \tag{18}$$

The relation uses the future camera orientation and target geometry. Candidates close to the band boundaries are excluded, and the remaining label is mapped to the options in the question. This horizontal-band rule does not assign a label to a point behind the camera or an undefined projection.

C Metric Geometry and Temporal Fusion

We supplement the absolute metric depth adapter and spatiotemporal fusion in Section 4.1, using the same depth variables and observation boundary. The Metric Encoder processes geometry derived from the refined depth D_{m}, while the Motion Encoder processes relative camera changes from the multiview geometry model. Their features are aligned in time and fused into visual tokens for the progressive state chain. The Ω entry in Table 3 denotes VGGT-Ω (Wang et al., 2026b). The VLM fine-tuning described in Section 4.2 uses LoRA (Hu et al., 2021).

Global and local depth corrections. For A observed geometric anchors, let $D_0, D_{\mathrm{m}} \in \mathbb{R}^{A\times H\times W}_{>0}$ denote the initial and refined metric depth maps. The global head predicts $\delta_{\mathrm{g}} \in \mathbb{R}^{A\times 1\times 1}$, which is broadcast across each frame, and the local head predicts $\Delta_{\mathrm{l}} \in \mathbb{R}^{A\times h\times w}$. Both heads use multiframe geometric features. The global head aggregates spatial features to estimate a frame-wide correction; the local head retains spatial features to estimate nonuniform corrections. The operation $\mathcal{U}$ in Equation 4 bilinearly upsamples each local map to $H \times W$. Thus the nonnegative coefficients γ_{g} and γ_{l} bound dimensionless log-depth updates, while D_0 and D_{m} remain in metres. For the update defined in Equation 4, these bounds imply

$$e^{-(\gamma_{\mathrm{g}}+\gamma_{\mathrm{l}})} \le \frac{D_{\mathrm{m}}(a,\mathbf{r})}{D_0(a,\mathbf{r})} \le e^{\gamma_{\mathrm{g}}+\gamma_{\mathrm{l}}}, \tag{19}$$

where $\mathbf{r}$ is an image location. The multiplicative form preserves positive depth and makes the correction proportional to the initial estimate. Its accuracy consequently depends on the quality of that estimate and of the geometric features used to refine it.

Metric geometry and camera motion. The refined depth represents distance along the camera's optical axis. At an anchor with intrinsics $\mathbf{K}_a$, a pixel (x, y) is lifted to the local camera frame as $\mathbf{x}^C_a = D_{\mathrm{m}}(a, x, y)\mathbf{K}_a^{-1}[x, y, 1]^\top$. Metric-surface features combine the depth, camera rays, lifted surface coordinates, and confidence and validity information. Camera-motion features from the multiview geometry model describe relative rotations and camera centres in the first valid anchor's coordinate system. The translation component is normalized by the accumulated camera path length, with a positive numerical floor, to accommodate the scale convention of the multiview predictor. Separate spatial encoders therefore process the metric surface and normalized camera-motion features. Their outputs at a shared anchor and spatial cell are concatenated and projected to a common latent width d. This preserves the distinct geometric roles of the two inputs while allowing them to contribute jointly to attention.

Anchor-time alignment and fusion. RGB frames and geometric anchors retain their source-frame indices within the observed interval $[i_s, i_o]$, whose endpoints correspond to the observation boundaries t_s and t_o. Their shared normalized temporal coordinate is $\eta(i) = (i - i_s)/\max(1, i_o - i_s)$; repeated RGB frames retain the same coordinate, and padded anchors carry an invalid mask. This coordinate represents position within the input window. Query-time embeddings are added to projected RGB tokens, and anchor-time embeddings are added to the geometry bank.

For attention head h, let $\mathbf{q}_{hq}$ be the visual query at RGB token q, and $\mathbf{k}_{hk}, \mathbf{v}_{hk}$ the key and value for geometry-bank token k. Each head has width d_h. Token k belongs to anchor $a(k)$, giving $\Delta\eta_{qk} = \eta(i_q) - \eta(i_{a(k)})$. A learned temporal bias $b_h(\Delta) = \beta_{h1}\Delta + \beta_{h2}|\Delta| + \beta_{h3}\Delta^2 + \beta_{h0}$ augments the content score:

$$\mathbf{o}_{hq} = \sum_k \mathrm{softmax}_k\left(\frac{\mathbf{q}_{hq}^\top \mathbf{k}_{hk}}{\sqrt{d_h}} + b_h(\Delta\eta_{qk}) + M_{qk}\right)\mathbf{v}_{hk}. \tag{20}$$

Here $M_{qk} = 0$ for valid geometry tokens and $M_{qk} = -\infty$ for invalid spatial cells or padded anchors; this validity mask is shared across queries. After concatenating the head outputs, an output projection and normalization map the result to the RGB token width:

$$\widetilde{\mathbf{x}}_q = \mathbf{x}_q + \alpha\,\mathcal{P}(\mathrm{Concat}_h\,\mathbf{o}_{hq})\,, \tag{21}$$

where α is a learned bounded scalar and $\mathcal{P}$ includes this projection and normalization. The resulting visual representation is the fused video representation X used in Sec. 4.2. Both the RGB queries and geometry bank are formed from the observed prefix ending at t_o; the withheld future interval supplies supervision targets rather than additional visual inputs.

D Dataset Composition and Evaluation

D.1 Training and Test Composition

We use SpatialMind-30K for VLM fine-tuning; it contains 30,677 question–answer pairs comprising 11,456 egocentric and 19,221 driving questions. We evaluate on the SpatialMind-2K benchmark, which contains 2,000 test questions comprising 858 egocentric and 1,142 driving questions. Table 5 gives the distribution across the eight task families. The test questions cover 357 source clips from 172 recording groups: 272 egocentric clips from 87 groups and 85 driving segments. A question may refer to a temporal crop, so these counts do not imply one independent recording per question.

Table 5: Question distribution by source domain and task family. Training statistics describe SpatialMind-30K, and test statistics describe SpatialMind-2K. A zero denotes an absent source-domain and task-family combination.

	Training: SpatialMind-30K		Test: SpatialMind-2K	
Task family	Ego	Driving	Ego	Driving
L1: Motion state	2,889	2,125	143	143
L1: Spatial state	1,968	2,312	143	143
L2: Ego-motion distance	3,758	7,073	143	143
L2: Target-motion distance	0	1,682	0	143
L2: Relative motion	308	1,734	143	143
L3: Ego future direction	0	1,007	0	143
L3: Target future distance	1,334	1,393	143	142
L3: Target future relation	1,199	1,895	143	142
Total	11,456	19,221	858	1,142

D.2 Evaluation Metrics

We use mean relative accuracy (MRA) for the five numerical task families and accuracy (ACC) for the three multiple-choice task families, following the main-paper protocol. MRA credits numerical predictions at multiple tolerance levels and averages the resulting correctness indicators. It therefore

measures closeness to the reference value instead of requiring exact numerical equality. Multiple-choice accuracy compares the selected option with the reference option. Relative motion, ego future direction, and target future relation use ACC; the remaining five families use MRA.

For a numerical prediction $\widehat{y}_i$ and reference y_i, let $\epsilon_i(c)$ denote the allowed error at tolerance setting c in the evaluation protocol. Then

$$\mathrm{MRA}_i = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathbf{1}\big[|\widehat{y}_i - y_i| \le \epsilon_i(c)\big], \qquad \mathrm{ACC}_i = \mathbf{1}[\widehat{a}_i = a_i]. \tag{22}$$

Here $\mathcal{C}$ is the set of tolerance settings, and a_i is the reference answer option. Numerical answers use the requested physical unit; a missing or unparseable final answer receives zero credit. External-benchmark results follow each benchmark's original evaluation convention.

E Additional Analysis

E.1 Persistence as a Dataset Diagnostic

We examine how much a future label changes relative to the last observed state. This analysis uses annotation-derived current states and is therefore a diagnostic of the question distribution, not an RGB-only predictive baseline. For target future distance, retaining the current annotated distance satisfies the item's numerical tolerance on 117 of 143 egocentric questions (81.8%) and 49 of 142 driving questions (34.5%). This distance diagnostic uses the single primary tolerance specified by each item's benchmark annotation, without averaging over tolerance levels. For target future relation, retaining the current relation agrees with the future category on 92 of 143 egocentric questions (64.3%) and 77 of 142 driving questions (54.2%). The central current label and the future label for straight ahead are treated as the same category. These rates characterize persistence within the selected tasks and tolerances; they neither establish a bound on model performance nor measure access to visual information.

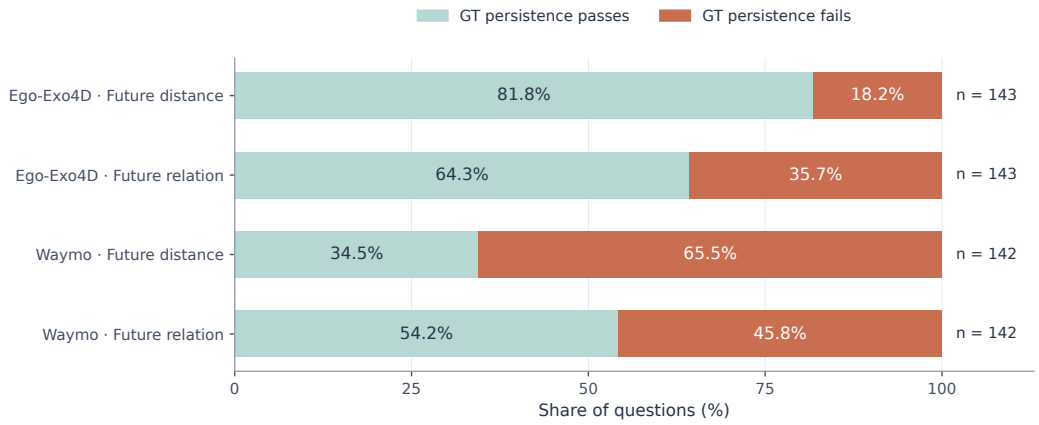


Figure 4: Persistence of annotation-derived current states in future distance and relation questions. A distance passes if its unchanged current value is within the future item's tolerance; a relation passes if the current and future categories match. Counts are 143 per task in the egocentric domain and 142 per task in driving. "GT persistence" uses the annotated current state and is not a model receiving only RGB.

E.2 Numerical Error Composition

Figure 5 summarizes the recorded StateChain, Free-form, and Qwen3-VL-8B predictions on all 1,286 numerical questions in SpatialMind-2K. We normalize each absolute error by the item's primary tolerance and retain unparseable answers as a separate category. This diagnostic reports question-level proportions at a single tolerance per item.

StateChain, our progressive state-chain reasoning setting, places 45.26% of predictions within tolerance, compared with 31.03% for Free-form and 12.52% for Qwen3-VL-8B. The share of errors exceeding four times the tolerance is 13.61%, compared with 26.83% and 57.85%, respectively. Relative to Free-form, StateChain reduces the frequency of these large errors by 49.3%. The distribution shows an advantage in both numerical precision and the avoidance of large deviations.

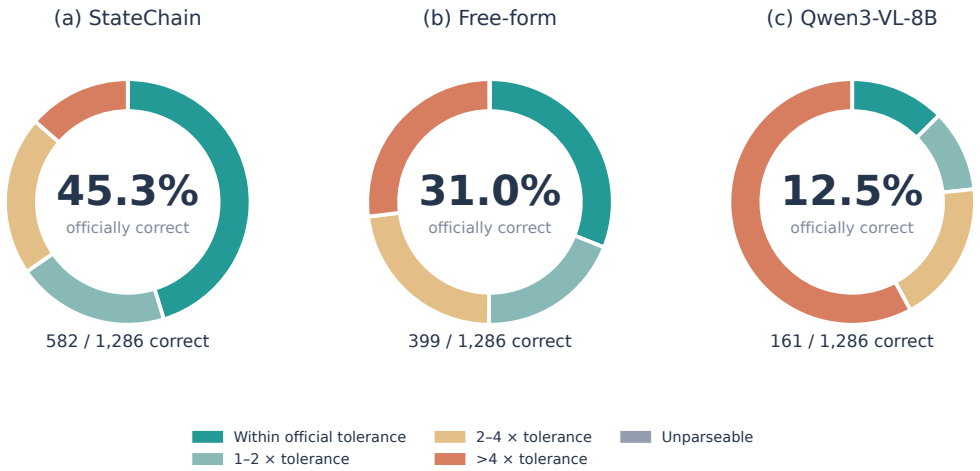


Figure 5: Numerical-error composition on the 1,286 numerical questions in SpatialMind-2K. StateChain denotes our progressive state-chain reasoning setting. Wedges group absolute error divided by the item's primary tolerance into $[0, 1]$, $(1, 2]$, $(2, 4]$, and $(4, \infty)$; unparseable responses remain in the denominator as a separate category. Ring centers show the within-tolerance percentages, with the corresponding counts below.

F Relationship to Other Spatiotemporal Tasks

A connected progression from spatial states to prediction. SpatialMind organizes eight tasks into current-state understanding, observed-dynamics understanding, and future prediction. This progression connects metric spatial estimates to changes over an observed interval and then to conditions beyond its endpoint. Its future tasks explicitly distinguish the observation boundary t_o from the queried time $t_f > t_o$. Future reasoning is also an explicit component of STRIDE-QA (Ishihara et al., 2026), whose evaluation interface includes future distance, bearing, and ego-speed queries. SpatialMind complements this setting with a common three-level task structure and progressive state-chain supervision, allowing current estimates, observed changes, and future targets to be studied within the same framework.

Driving and everyday first-person settings. SpatialMind combines Waymo driving videos and Ego-Exo4D everyday interactions, covering both traffic scenes and first-person activities with substantial changes in viewpoint and spatial context. Questions identify targets through linguistic descriptions linked to persistent entities. STRIDE-QA's released interface instead provides visually marked targets in four observed images at -1.5, -1.0, -0.5, and 0 seconds. The use of linguistic descriptions makes target grounding an explicit part of SpatialMind's spatial reasoning task. SpatialMind-2K's variable observation windows extend to 19.8 seconds, supporting longer temporal evidence than the 1.5-second span of this published STRIDE-QA interface; these durations describe observation spans rather than the number of frames sampled by a model. The two-domain design also supports examining the same reasoning hierarchy under different scene and motion conditions; it does not require treating their annotation sources as identical.

Traceable geometric targets. MLLM-4D (Yin et al., 2026) studies time-evolving object–camera relationships in three-dimensional space, making it a relevant comparison for reasoning about dynamic geometry. SpatialMind emphasizes the connection between this reasoning and explicitly computed supervision: entity identities, coordinate references, measurement intervals, and geomet-

ric operations determine the reference answers and state-chain targets. These records make it possible to check whether a distance, path length, direction, or future relation corresponds to the quantity requested by the question. Together with numerical and categorical evaluation, this provides a traceable basis for analyzing errors across the three levels. Comparisons retain each benchmark's input and scoring conventions; shared spatial terminology alone does not imply identical tasks.

G Limitations and Future Work

SpatialMind depends on estimated geometry, correspondences, and trajectories, whose errors can propagate to answers and state-chain targets. Its two source domains and the prediction horizons evaluated in the benchmark do not establish unrestricted transfer or long-term forecasting. Readable state chains do not establish causal faithfulness, and the reported experiments do not characterize end-to-end latency. Future work will quantify sensitivity to geometric errors and extend evaluation to longer horizons and more diverse motion, alongside matched-protocol comparisons and explicit measurements of inference cost.

H Visualizations

The following examples connect observed video, metric geometry, reference-answer computation, and progressive state-chain supervision. Model inputs contain only the observed video and its question, with no future frames or textual information about future states. Future references shown in prediction examples are used to construct supervision and are withheld from model inputs.

H.1 A Worked Example of Reference-Answer Computation

From a video question to structured supervision. Figure 6 illustrates the reference-answer computation for a Level 2 question, linking the observed video and displayed spatial measurements to the numerical decision and progressive state-chain supervision.

Question header. The silver pickup truck with a black roof rack is the target. The question concerns changes in its distance and direction relative to the observer. Its five options distinguish approaching, moving rightward, remaining approximately unchanged, moving leftward, and receding. This wording determines which geometric relation must be measured throughout the clip.

A: Video clip. The observed window contains 16 consecutive frames sampled at 5 fps, spanning $(16-1)/5 = 3.0$ s from the first to the last frame. The six displayed thumbnails summarize this window while preserving the target's identity. Bounding boxes help readers locate the pickup; they are not part of the model's video input. The target remains visible throughout the observed window used for the calculation.

B: Depth and semantics. At the last observed frame, metric depth describes the scene's distance structure and the semantic map locates the target within the scene. Together, these views connect the appearance-based referring expression to a spatial entity. The relative-motion calculation uses the target's annotated three-dimensional centres, whose camera-relative ranges and bearings are expanded in the subsequent computation figure.

C: Geometry and decision. The spatial view and computation summary show range increasing from 12.603 to 18.165 m, alongside a small rightward bearing change from approximately 0.034° to 0.652°. Applying Equation 13 compares normalized radial change with angular change in radians. Radial motion dominates, and the trend check passes, yielding the receding relation and option E. A small rightward change therefore does not determine the categorical answer by itself.

D: Progressive state chain. The grounding block fixes the target, observation scope, reference frame, and requested quantity. The current-state block records the endpoint measurements. The dynamics block expresses their signed changes and resulting motion relation, and the answer block maps that relation to E. This structure preserves the connection between measured geometry and the final categorical supervision.

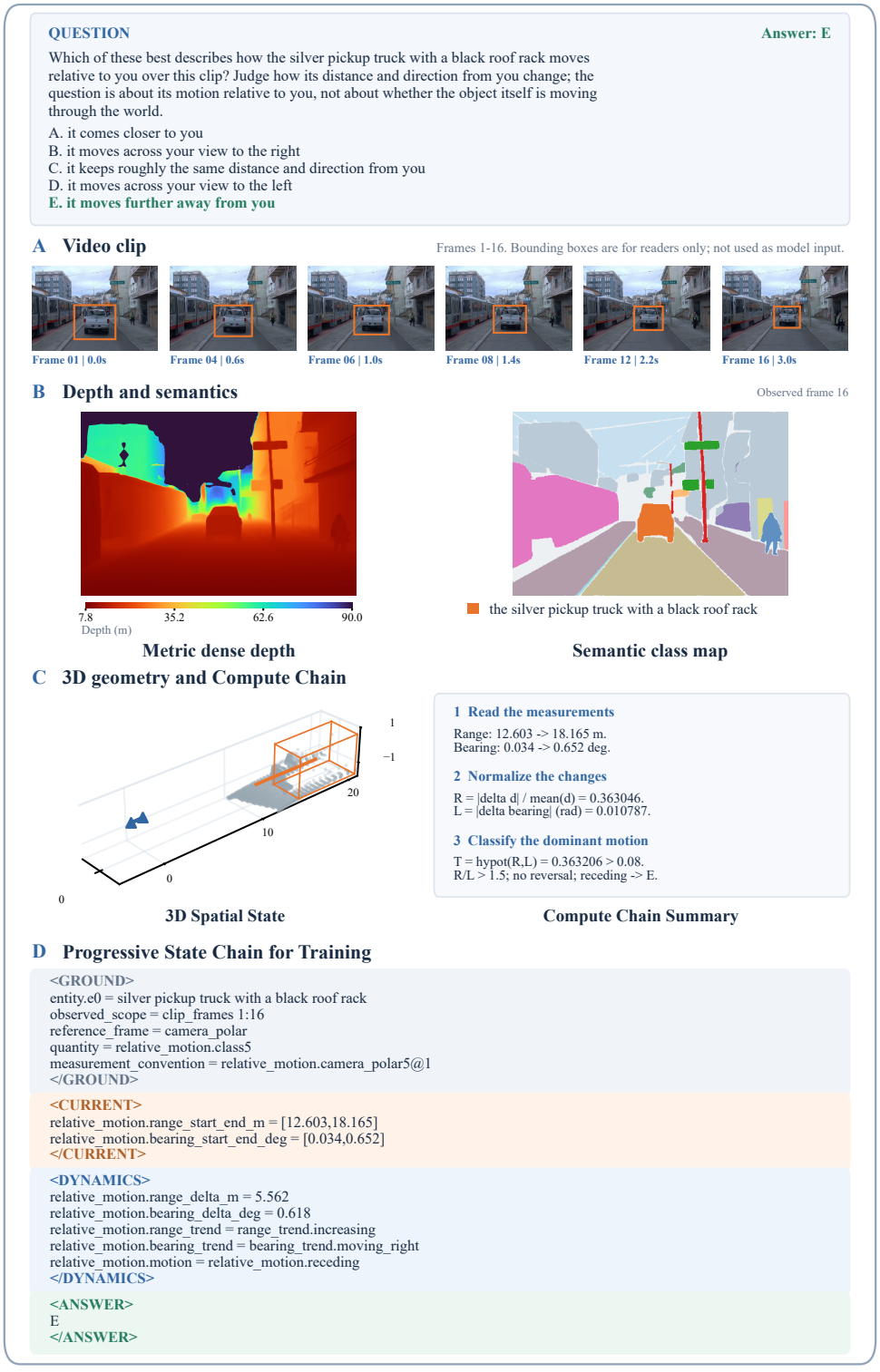


Figure 6: Level 2 relative-motion example, from the question and observed video to structured supervision. The target pickup recedes from the observer, giving option E. The question header and panels A–D show the observed clip, depth and semantics, geometric calculation, and progressive state-chain target.

H.2 The Full Compute Chain for the Same Example

Expanded reference-answer computation. Figure 7 expands the same example into seven steps, exposing the measurements and checks behind its reference label.

01: Target grounding. The referring expression identifies the pickup. Verification on two observed frames gives intersection-over-union values of 0.514 and 0.543, both marked as passing.

02: Observation window. The selected sequence supplies 16 consecutive observations over 3.0 s. All core motion operands come from this window, with the target visible in every frame.

03: Spatial quantities. The table lists optical-frame centres in metres, with axes pointing right, down, and forward. Range is the centre's Euclidean norm, rounded to millimetres. Bearing is the right-positive horizontal angle, displayed as β_k in degrees; thus $\alpha_k = \pi\beta_k/180$ in the preceding equations.

04: Normalization and classification. All 16 ranges give $\bar{d} = 15.320375$ m. Their endpoint difference is 5.562 m, and the bearing change is $0.618024°$. Equation 13 gives $R = 0.363045944$, $L = 0.010786549$, and $T = 0.363206149$. With zero-based midpoint $m = 8$, the products $(d_m - d_0)(d_{15} - d_m) = 7.726905\ \mathrm{m}^2$ and $(\beta_m - \beta_0)(\beta_{15} - \beta_m) \simeq 0.095344\ \mathrm{deg}^2$ are positive: the two halves agree in sign. The ordered classifier rules out unchanged motion because $T \geq 0.08$, passes these reversal checks, and selects receding because $R \geq 1.5L$ and $\Delta d > 0$. A dominant-axis sign reversal or lack of dominance would exclude a candidate.

05: Alternative readings. World displacement measures a different quantity: the object and observer displacements are 6.657 and 1.086 m. Pixel-size estimates and fitted trends provide cross-checks; agreement here does not validate mixing metres with degrees or reversing the bearing convention. The pixel-derived final range, 20.150 m, also differs from the annotated 18.165 m.

06: Answer mapping. Receding maps to E under exact option matching, with 20% uniform chance accuracy. The inset's nine unsuccessful blind rules describe this sample only, not a benchmark-wide evaluation.

07: Executable verification. Recomputation reproduces E, with range/bearing residuals at most 10^{-9}. The optical-to-camera mapping $(x, y, z) \mapsto (z, -x, -y)$, followed by extrinsics and pose, recovers the world centre with maximum residual 7.11×10^{-15} m. These checks establish agreement between the displayed quantities and their geometric source. The residuals verify numerical and coordinate consistency, not absolute measurement accuracy.

Compute Chain

01 Target grounding

the silver pickup truck with a
black roof rack
Track: TCJudzwmoKhRoLoVYgQ4Sw
Describe @ 56; verify @ 72 and 81.
Verification IoU: 0.514 / 0.543 (pass).

02 Observation window

Source 69-84 -> clip frames 1-16.
16 consecutive frames; 5 fps; 3.0 s.
Target visible in every served frame.
Core operands use this window only.
Displayed boxes are reader aids only.

03 Read the original spatial quantities

Optical center C_k (m): x right, y down, z forward. Centers shown to 6 d.p.
d_k=round(||C_k||,3); beta_k=degrees(atan2(x_k,max(z_k,1e-6))).

Clip	Source	x (m)	y (m)	z (m)	d (m)	beta (deg)
1	69	0.007444	1.010024	12.562862	12.603	0.033950
2	70	0.017471	1.010121	12.872551	12.912	0.077764
3	71	0.026719	1.008415	13.202820	13.241	0.115951
4	72	0.038220	1.009484	13.529156	13.567	0.161860
5	73	0.048880	1.010612	13.870314	13.907	0.201916
6	74	0.062681	1.017924	14.265529	14.302	0.251748
7	75	0.072648	1.014081	14.651516	14.687	0.284094
8	76	0.084022	1.019721	15.045019	15.080	0.319975
9	77	0.095624	1.025516	15.433564	15.468	0.354990
10	78	0.109004	1.031250	15.851140	15.885	0.394003
11	79	0.123771	1.060817	16.262672	16.298	0.436056
12	80	0.138618	1.064248	16.647797	16.682	0.477061
13	81	0.155720	1.069473	17.021847	17.056	0.524141
14	82	0.174380	1.065292	17.418589	17.452	0.573577
15	83	0.190388	1.062327	17.788378	17.821	0.613210
16	84	0.206342	1.056625	18.132683	18.165	0.651974

04 Normalize the changes and apply the complete decision rule

mean(d)=15.320375 m; delta_d=5.562 m; delta_beta=0.618024 deg
R=|delta_d|/max(mean(d),1e-6) = 0.363045944
L=|radians(delta_beta)| = 0.010786549
T=hypot(R,L)=0.363206149; R/L=33.657283

1 T < 0.08 -> same	false
2 p_d < 0 and R >= 1.5L -> refuse: distance reversal	false
3 p_b < 0 and L >= 1.5R -> refuse: bearing reversal	false
4 R >= 1.5L -> closer if delta_d < 0, otherwise away	away
5 L >= 1.5R -> left if delta_beta < 0, otherwise right	skipped
6 Otherwise -> refuse: neither axis dominates	skipped

m=8 (clip 9 / source 77); p_d=(d_m-d_0)(d_15-d_m)=+7.726905
p_b=(beta_m-beta_0)(beta_15-beta_m)=+0.095344; no reversal.

05 Alternative readings and cross-checks

World displacement: object 6.657 m; ego 1.086 m (a different quantity).
Pixel-size approximation -> away; full-series least-squares slope -> away.
Raw metre-vs-degree comparison and flipped bearing sign also yield away here.
Pixel estimate at clip end: 20.150 m; original 3D center range: 18.165 m.

06 Answer mapping and scoring

A: closer; B: right; C: same.
D: left; E: away.
Exact option match; chance = 20%.
Recorded blind rules: 0 / 9 correct.
State: relative_motion.receding.

07 Executable verification

Recompute = E; matches source proof.
Range / bearing residuals <= 1e-9.
Optical -> camera: (x,y,z) -> (z,-x,-y).
Then extrinsic + pose -> world center.
Maximum world residual: 7.11e-15 m.

E | it moves further away from you

Figure 7: Reference-answer computation for Figure 6. This is the expanded record behind the compressed calculation in panel C: grounding and temporal scope, per-frame measurements, normalization and decision rules, cross-checks, answer mapping, and numerical verification. The chain documents how the displayed measurements determine the answer and its supervision.

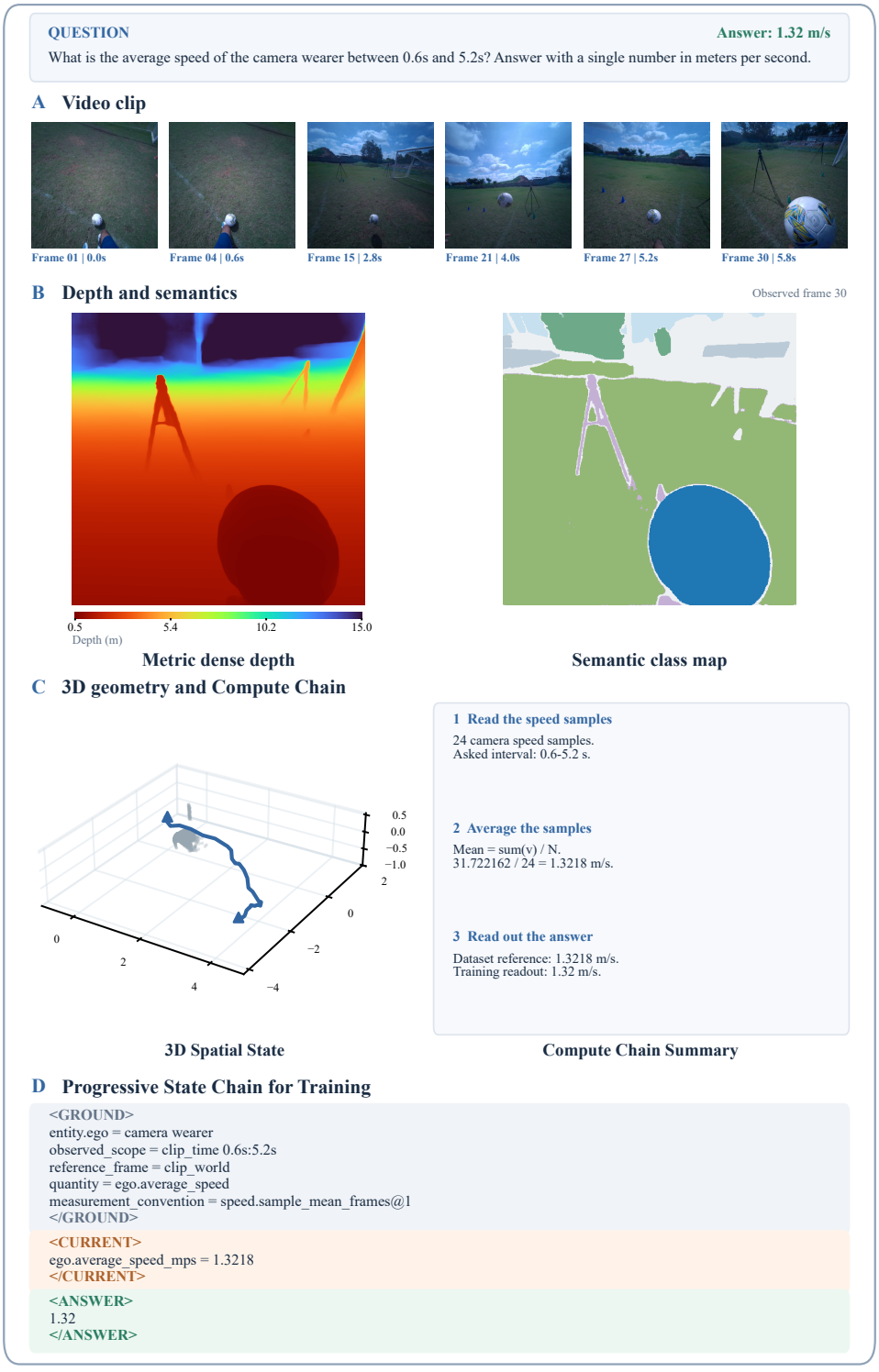


Figure 8: Level 1 motion-state example of egocentric scene: averaging 24 camera-speed samples over 0.6–5.2 s gives 1.3218 m/s, with 1.32 m/s as the training answer. The displayed state chain is a supervision target.

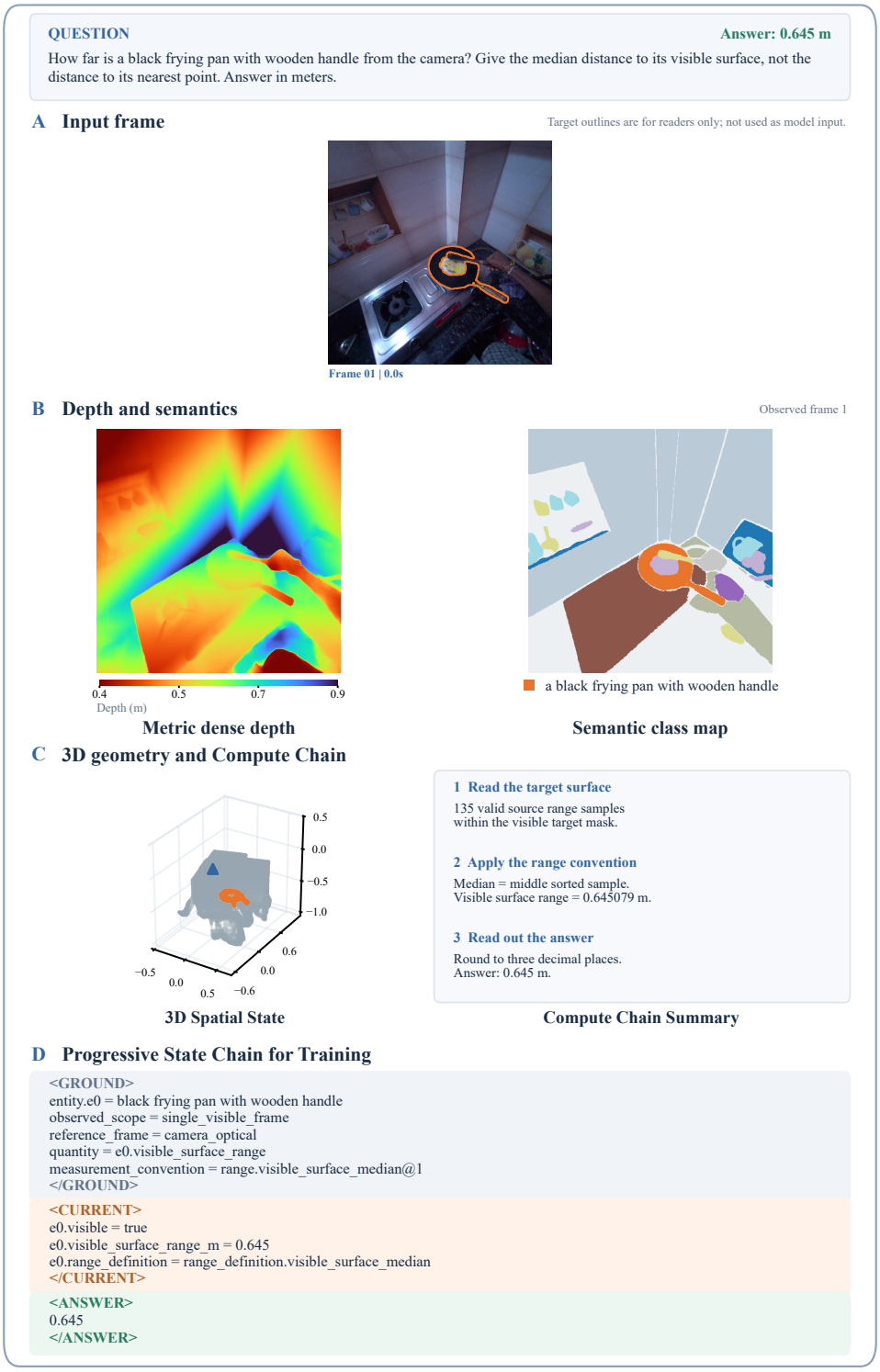


Figure 9: Level 1 spatial-state example of egocentric scene: the median range of 135 valid samples on the frying pan's visible surface is 0.645 m. This question uses a visible-surface statistic rather than an object-centre or nearest-point distance; outlines are reader annotations, and the state chain is a training target.

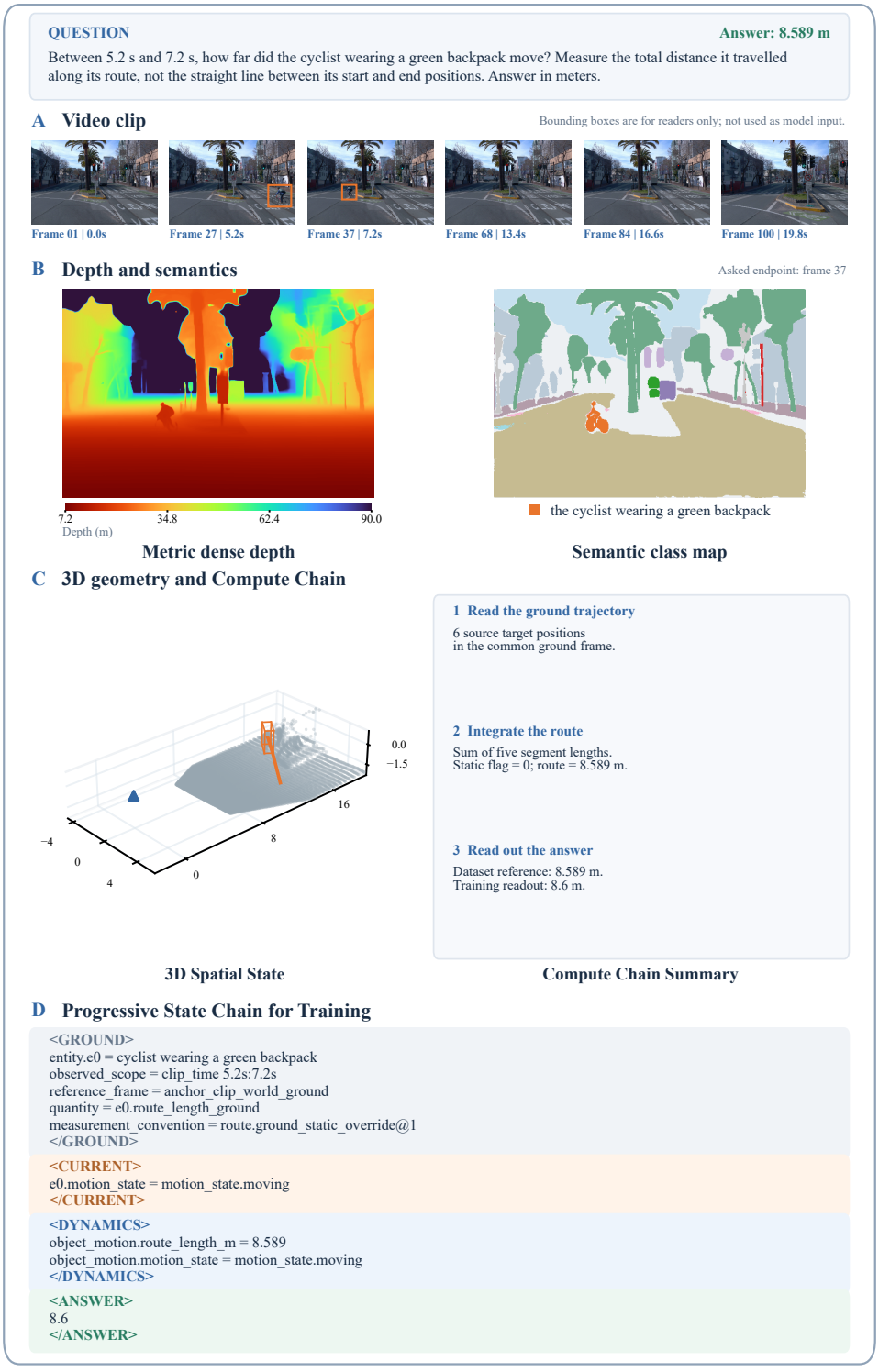


Figure 10: Level 2 target-motion example: summing five ground-trajectory segments over 5.2–7.2 s gives the cyclist a route length of 8.589 m, with 8.6 m as the training answer. The route uses six source positions; boxes are reader annotations, and the chain is a supervision target.

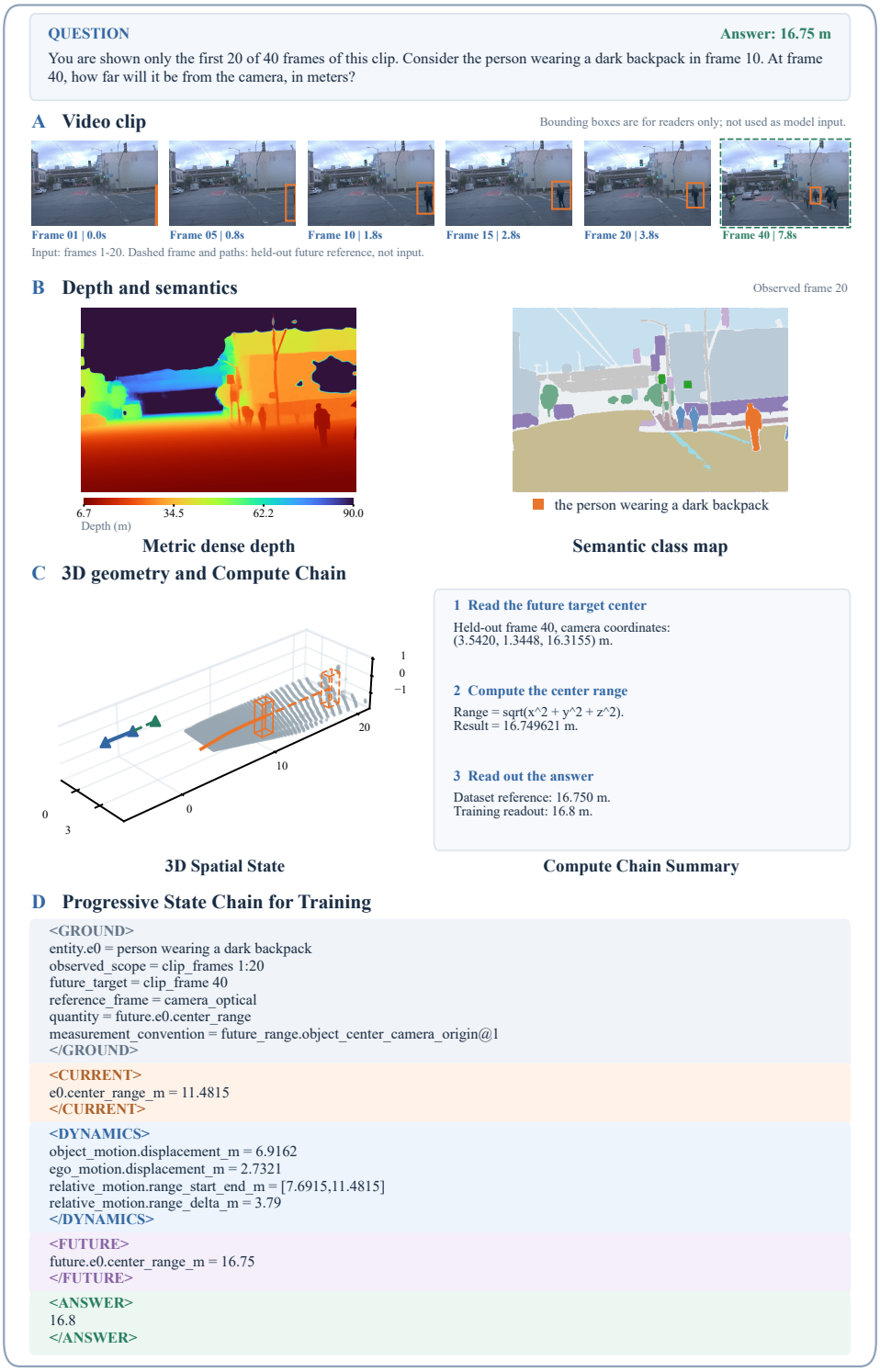


Figure 11: Level 3 target-future-distance example: frames 1–20 are supplied to query the distance from the backpack wearer's centre to the optical camera origin at frame 40. The displayed reference is 16.75 m and the serialized training answer is 16.8 m. Dashed future references and chain targets illustrate supervision; neither future frames nor future-state text is supplied at inference.