# T4D: COMPREHENDING THE DYNAMIC WORLD WITH MLLMS

**Anonymous authors**
Paper under double-blind review

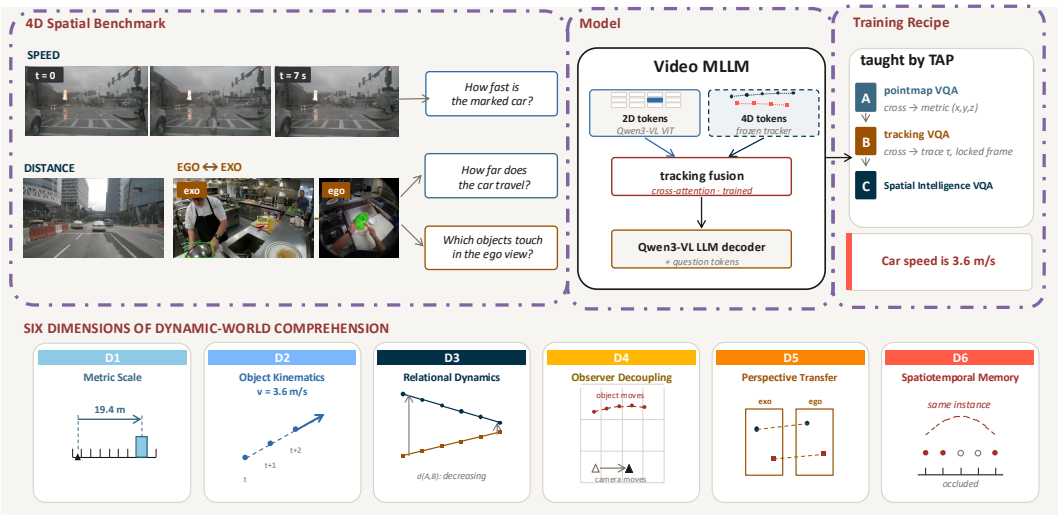
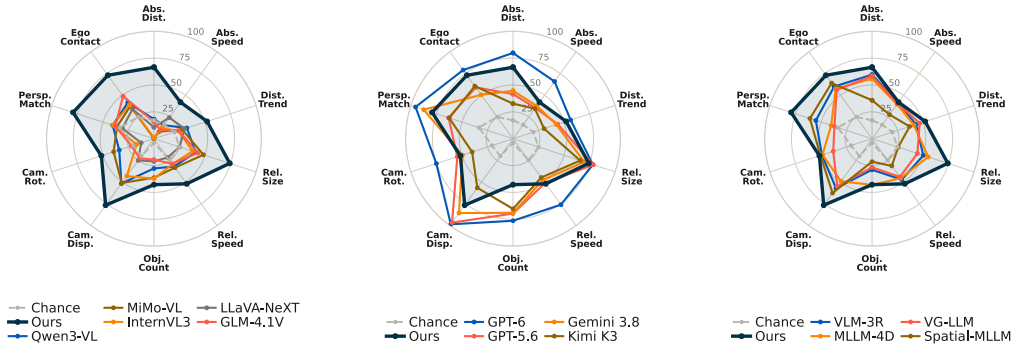



**Figure 1: Overview of T4D.** *4D Spatial Benchmark*: 10 tasks and 3,250 expert-verified questions on real indoor and outdoor video, covering all six dimensions. *Model*: a 4D model is the geometry encoder and supplies 4D trajectories; PAF aligns them with 2D appearance at the same image location. *Recipe*: a training recipe aligns 2D visual information with this 4D evidence and closes the shortcut that would bypass the tracks.

## ABSTRACT

Multimodal large language models (MLLMs) have made rapid progress on 3D spatial reasoning, yet these successes remain confined to a static world. Real scenes evolve, and understanding them requires reasoning jointly over space and time—*4D spatial reasoning*. Existing benchmarks judge spatial change only qualitatively and leave dynamic egocentric–allocentric transformation untested; at comparable scale, open-source and expert models remain weak on dynamic scenes. We present **T4D**, a systematic study of dynamic spatial intelligence. We revisit spatial intelligence and, for the first time, construct the taxonomy of this field from a dynamic perspective. This yields six dimensions, instantiated as **T4D-BENCH**: 10 tasks and 3,250 expert-verified QA pairs. Two difficulties stand in the way of scale: the clips do not share one 4D annotation—outdoor video supplies boxes, tracks, and ego poses, while egocentric video needs another geometric readout—and a language reference is ambiguous among similar objects. Our automated pipeline computes each answer from the native measurements and marks the queried entity on the frame, producing the benchmark and about 600K training QA pairs. On the algorithm side, **(1)** we propose the T4D network. It fuses explicit 3D point trajectories of instances with vision, and it attains state-of-the-art on T4D-BENCH among MLLMs of comparable scale. **(2)** Trajectory-aligned training (*TAP*), a training recipe, supervises those instance trajectories in a fixed reference frame, aligning 2D vision tokens with the trajectory tokens and improving dynamic metric reasoning.

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.




**(a)** Open-source general models. **(b)** Proprietary general models. **(c)** Spatial MLLMs.

**Figure 2: Performance of different VLMs on T4D-BENCH.**

## 1 INTRODUCTION

Recent advances in pre-training and multimodal alignment have enabled multimodal large language models (MLLMs) to achieve remarkable progress on 3D spatial reasoning (Gemini Team et al., 2024; Bai et al., 2025; Hong et al., 2025; Yang et al., 2025a). These successes, however, are confined to a static world. Real scenes evolve: objects move, observers act, and 3D spatial relations shift continuously, so understanding them demands reasoning jointly over space and time. This capability is fundamental: the physical world is inherently four-dimensional, and applications that must act within it—from embodied AI to VR/AR—all depend on it.

Despite its importance, 4D spatial intelligence remains underdeveloped. Existing benchmarks probe only a narrow slice of it. Several ask a metric distance, and a few also ask a speed, but paired egocentric–allocentric transformation still has no full test (Zhou et al., 2025a; Li et al., 2025b; Yang et al., 2025a; Zhou et al., 2026). Existing work on models focuses largely on static spatial understanding: at comparable scales, both open-source and expert models perform poorly in dynamic scenes (Figure 2). In short, 4D spatial intelligence still lacks a comprehensive and systematic exploration.

To fill this void, we introduce T4D, a systematic study of dynamic spatial intelligence. We revisit spatial intelligence and, for the first time, construct the taxonomy of this field from a dynamic perspective, yielding six dimensions—metric scale, object kinematics, relational dynamics, observer decoupling, perspective transfer, and spatiotemporal memory (Figure 4). The taxonomy immediately exposes the evaluation gap: measured against it, no existing 4D benchmark covers all six dimensions, and none fully covers perspective transfer (Figure 3). DSR-Bench's partial mark on that dimension is a viewpoint change inside one clip; a full mark requires the same fact across an egocentric video and an exocentric video. T4D-BENCH is built to close this gap (Section 3), with **10 tasks** and **3,250** QA pairs whose open numerical and structured answers directly probe every dimension. Two obstacles stand in the way of building it at scale. The clips do not share one 4D annotation: outdoor video supplies boxes, tracks, and ego poses, while egocentric video needs another geometric readout, so each answer has to be computed from whichever measurement the footage actually carries. A language reference is also ambiguous among similar objects, so we place a visual mark on the frame to remove that ambiguity and let the MLLM focus on reasoning about spatial relations. The same automated pipeline produces T4D-BENCH and a filtered training set of about **600K** QA pairs; every benchmark item is then screened by professional inspectors.

Dynamic metric reasoning needs explicit 3D point trajectories of instances, not only a reconstructed scene field. Leveraging the capability of existing 4D geometry models, we extract from them sufficient 3D trajectories. The T4D network fuses them with vision. A 2D patch and the track at the same image location no longer share an address, so plain cross-attention would treat the tracks as an anonymous bag; position-addressed 4D fusion (*PAF*) writes each track's image-plane position into the match. Trajectory-aligned training (*TAP*) then supervises the instance trajectory in a fixed reference frame, so the fused tracks are used for the metric answer (Section 4).

To sum up, our contributions are:

| Dimension Space | VSI-Bench | VLM4D | DSI-Bench | STI-Bench | Spatial4D | MLLM4D | DSR | Ours |
|---|---|---|---|---|---|---|---|---|
| D1: Metric Scale | ✓ | × | × | ✓ | ✓ | ⌄ | × | ✓ |
| D2: Object Kinematics | × | ⌄ | ✓ | ✓ | × | ⌄ | ✓ | ✓ |
| D3: Relational Dynamics | ⌄ | × | ⌄ | ⌄ | × | ⌄ | ✓ | ✓ |
| D4: Observer Decoupling | ⌄ | ⌄ | ✓ | ✓ | ⌄ | ✓ | ⌄ | ✓ |
| D5: Perspective Transfer | ⌄ | ⌄ | × | × | ⌄ | × | ⌄ | ✓ |
| D6: Spatiotemporal Memory | ⌄ | ⌄ | ⌄ | × | ✓ | ⌄ | ⌄ | ✓ |

**Figure 3: 4D spatial capability matrix**. ✓ denotes full support, ⌄ partial support, and ✗ not evaluated.

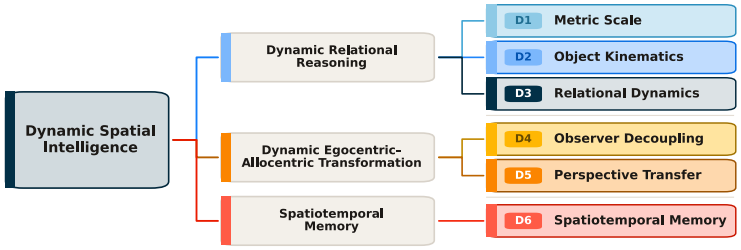



**Figure 4: A taxonomy of dynamic spatial intelligence.** We revisit spatial intelligence and, for the first time, construct this taxonomy from a dynamic perspective, in six dimensions (D1–D6).

**(1) T4D-BENCH.** The first benchmark to evaluate dynamic spatial intelligence along all six dimensions, comprising 10 tasks and 3,250 expert-verified QA pairs with open numerical and structured answers.

**(2) Pipeline and large-scale data.** An automated pipeline over real indoor and outdoor video with native 4D metadata, and the large-scale dataset derived from it: physically checkable VQA comprising T4D-BENCH and about 600K training pairs.

**(3) Model.** We propose the T4D network. It fuses explicit 3D point trajectories of instances with vision, and it attains state-of-the-art on T4D-BENCH among MLLMs of comparable scale. We will open-source the model, advancing future work in this field.

**(4) Training recipe.** Trajectory-aligned training (*TAP*), a training recipe, supervises those instance trajectories in a fixed reference frame, aligning 2D vision tokens with the trajectory tokens and improving dynamic metric reasoning.

## 2 RELATED WORK

T4D sits at the intersection of general-purpose multimodal models, spatial-intelligence evaluation, and geometry-conditioned architectures. We review each strand in turn, covering both static 3D and dynamic 4D settings.

### 2.1 MULTIMODAL LARGE LANGUAGE MODELS

Video MLLMs now dominate open-ended visual instruction following, yet their visual tokens remain 2D appearance features (Liu et al., 2023; Bai et al., 2025; Yang et al., 2025a). They can name objects and narrate events, and they have begun to support static 3D spatial reasoning, but they are not constructed to report metric distances, world-frame speeds, or correspondences across moving cameras. T4D therefore keeps a strong video MLLM as the language interface and asks what additional representation—and what evaluation protocol—are required before that interface can perform 4D spatial reasoning.

### 2.2 Benchmarks for Spatial Intelligence

**Static 3D and image-level spatial QA.** ScanQA (Azuma et al., 2022), SQA3D (Ma et al., 2023), and OpenEQA (Majumdar et al., 2024) test indoor understanding from reconstructions or embodied queries; SpatialVLM (Chen et al., 2024a), SpatialRGPT-Bench (Cheng et al., 2024), and CV-Bench (Tong et al., 2024) probe layout, distance, and size on stills. VSI-Bench (Yang et al., 2025a) is the most influential *video* spatial benchmark (∼5,000 QA from ScanNet-family egocentric video); MindCube (Yin et al., 2025) stresses perspective taking from sparse views. These resources established an emerging 3D sense in MLLMs, but remain largely *static*: a snapshot or a slowly explored room is usually enough, and they do not require world-frame object kinematics under a moving observer.

**Dynamic 4D and spatiotemporal benchmarks.** A newer wave targets motion, each covering a slice, and most judge spatial change only qualitatively. VLM4D (Zhou et al., 2025a) and DSI-Bench (Zhang et al., 2025) ask observer-versus-object questions without open metric answers; STI-Bench (Li et al., 2025b) is the closest quantitative predecessor on camera pose, displacement, and some speed items. Spatial4D-Bench (Wang et al., 2026b) scales static metric layout and long-horizon memory; MLLM4D-Bench (Yin et al., 2026) decomposes object versus camera motion as binned multiple choice. DSR-Bench (Zhou et al., 2026) is the closest contemporaneous suite: in-the-wild video, reconstructed relative-scale geometry, viewpoint transform, and procedural qualitative answers for distance, direction, orientation, and speed—all multiple choice, with no open SI-unit targets. Figure 3 marks a cell full when a suite has a primary task for that dimension: metric distance in VSI-Bench, STI-Bench, and Spatial4D-Bench; object motion in DSI-Bench and object speed in STI-Bench; distance change and speed comparison in DSR-Bench; observer-versus-object motion in DSI-Bench, STI-Bench, and MLLM4D-Bench; spatial memory in Spatial4D-Bench. DSR-Bench's perspective cell stays partial, because its viewpoint changes inside one clip and the questions do not match one fact across an egocentric video and an exocentric video. No existing suite covers all six dimensions. That incomplete examination is the evaluation contract of T4D-Bench.

### 2.3 Algorithms for Spatial Intelligence

One family injects reconstructed geometry into the language model—3D-LLM (Hong et al., 2023), LLaVA-3D (Zhu et al., 2025), Spatial-MLLM (Wu et al., 2025) (VGGT (Wang et al., 2025a)), and VLM-3R (Fan et al., 2025) (CUT3R (Wang et al., 2025b)). These reconstructors recover scene geometry and camera ego-motion well, and they help static metric questions, but they treat the scene as quasi-static: moving objects are left without persistent, instance-centric trajectories in a declared world frame. GAP-MLLM (Zhang et al., 2026) further shows that a reconstruction encoder is unused unless a text loss forces the language head to read it; its pointmap VQA still does not ask for a world-frame trace. Specialist 4D systems already recover those traces: SpatialTrackerV2 (Xiao et al., 2025) estimates long-range 3D trajectories and camera pose from ordinary video. Video MLLMs that "think in dynamics"—VLM4D (Zhou et al., 2025a), LAST (Wang et al., 2025c), MLLM-4D (Yin et al., 2026)—remain early and do not treat instance-centric 4D trajectories as a first-class token stream.

The added representation is therefore not a generic 4D geometry code. DSR injects a reconstructed relative-scale field; GAP-MLLM activates a geometry stream with static pointmap supervision. T4D supervises explicit 3D point trajectories of instances in a fixed reference frame, and fuses those trajectories with vision by writing each track's image-plane position into the match. The gain we claim is on dynamic metric reasoning.

## 3 T4D Data

This section describes how T4D turns real video into data. A single pipeline converts indoor and outdoor clips with native 4D metadata into physically checkable QA pairs, producing T4D-Bench—**10 tasks** and **3,250** QA pairs, every item screened by professional inspectors—together with a filtered training set of about **600K** QA pairs. Section 3.1 defines the task suite, Section 3.2 describes the pipeline behind it, and Section 3.3 reports statistics and positions the benchmark against prior suites.

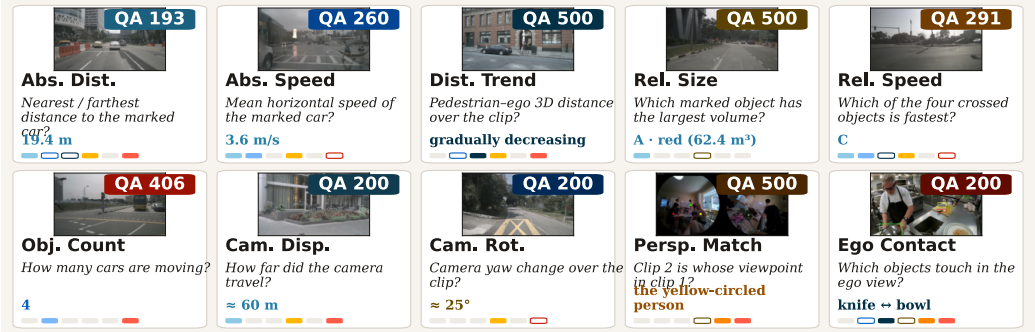



**Figure 5: Examples of the ten T4D-BENCH tasks.** The QA tag is the question count. Chips along the bottom are D1–D6: filled means a primary probe, outlined means also required, empty means not examined.

### 3.1 TASK SUITE

T4D-BENCH is organized around the six dimensions of Figure 4. *Metric scale*, *object kinematics*, and *relational dynamics* (D1–D3) extend relational reasoning to a moving world: they ask for the physical quantities—metric distance and speed, evolving 3D relations—on which embodied agents plan and act, and for which linguistic proxies such as "near" or "fast" are no substitute. *Observer decoupling* and *perspective transfer* (D4–D5) extend egocentric–allocentric transformation to moving observers: an embodied agent's camera moves with its body, and assistance from a second viewpoint requires the same physical fact to hold across first- and third-person views. *Spatiotemporal memory* (D6) underlies both, keeping instance identity and events aligned across frames and occlusions, as any long-horizon embodied task presumes. Fuller per-dimension definitions are in Appendix A.1.

The task–dimension mapping and the question count of each task are on the cards in Figure 5.

### 3.2 BENCHMARK CONSTRUCTION

All ten tasks come from one pipeline (Figure 6); only the annotation source and the geometric readout differ. The same stages produce the training questions from footage that already carries metric 4D annotations. Those questions pass the automatic filters and are kept strictly separate and isolated from the benchmark data. The procedure for each task is given in Appendix A.2.

**Data filtering.** A clip is kept only when that task's own criterion is met, so the answer is computed from measurements the footage already carries. Outdoor clips from nuScenes (Caesar et al., 2020) supply 3D boxes, tracks, and ego poses, at a keyframe interval of about 0.5 s. Metric-motion tracks must stay visible for at least four keyframes and must not be near-static; a distance comparison needs a co-visible pair; counting uses instance identities of a named motion state; camera tasks use a 10-frame window with stride 2 and score the summed path, so a loop is not standing still. Relative size and relative speed concatenate outdoor nuScenes clips into one video: the four marked objects are not taken from a single shot. `perspective_matching` uses EgoLife (Yang et al., 2025b), which supplies no boxes: a person is kept only when present for at least 70% of a five-second window, and exactly three people are retained. `egocentric_contact` keeps a 15-second EgoExo4D window (Grauman et al., 2024) with three objects.

**Annotation.** Each source gets its own automated pipeline for extracting annotations, matched to the measurements that source actually carries. Following Cai et al. (2025); Zhang et al. (2026), the queried entity is marked on the frame rather than named in the prompt: an adaptive cross on the projected 3D centre, a distinct color when several objects are compared, a colored circle on each person for perspective matching, or three SAM2 boxes on the exocentric view for contact. Counting draws no mark, because the question covers the full field. Contact is read from the aligned egocentric view as apparent touch or occlusion, not a 3D collision. SpatialVID (Wang et al., 2026a) and OmniWorld (Zhou et al., 2025b) do not carry metric object tracks, so their pipeline estimates a pseudo-trajectory instead. That estimate supports an ordering or a trend rather than a metre, and therefore only multiple-choice questions: 35,000 relative-speed and 35,000 relative-distance items from SpatialVID, and 30,000 of each from OmniWorld. This pool stays out of the official split.

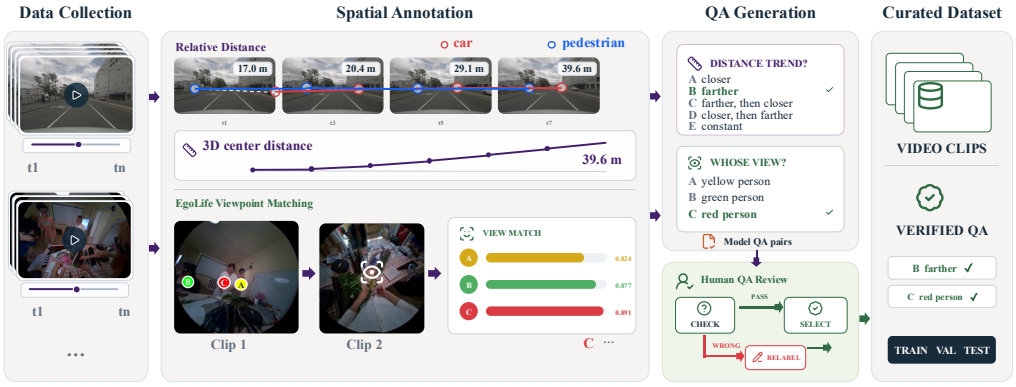



**Figure 6: Data construction pipeline.** An automated pipeline that scales up training data for dynamic spatial intelligence.

**Question-answer pair generation.** Each procedure computes one checkable answer. Where several wordings are valid, the template varies the queried extreme, the entities, or the bins, while the measurement stays fixed. Open answers are metres, metres per second, or a count; the remaining tasks are multiple choice, including five distance-trend templates, magnitude or degree bins, and color rankings. Windows and formulas are in Appendix A.2.

**Human-in-the-loop review.** An automatic filter runs first: several MLLMs each answer the question and assign a confidence score. Only items with low confidence, or on which the models disagree, are passed to a person. Benchmark candidates are drawn with a stricter setting of this filter, and every candidate is then reviewed by hand; the items that pass form the official split. We hired **12 professional annotators** from a specialist data-annotation company to screen these candidates. These annotators have extensive experience in data annotation and quality inspection. The annotators are mixed in gender, age, and background, so that judgements of contact, identity, and marker placement are not left to one person. They work in a professional data-inspection system (interface in Appendix A.3) and follow a strict written protocol. The review checks three aspects. *Data quality*: video clarity, the fraction of the frame occupied by the main object, its visibility, and how long it remains in view. *Answer quality*: whether the speed, size, pose, and spatial relations stated in the answer agree with the video and with physical common sense. *QA-template quality*: whether the generated question and answer are fluent and unambiguous.

### 3.3 Benchmark Statistics

The official split contains 2,944 clips and 3,250 questions (Table 5): 2,298 outdoor clips (2,604 questions) and 646 indoor clips (646 questions), as in Figure 11a. Outdoor questions outnumber outdoor clips because object counting asks several questions of one clip. Clips last 1.4–72 s (mean 15.8 s), and evaluation keeps at most 128 frames (Figure 12). Of the 3,250 answers, 859 are open numbers—metres, metres per second, or counts—and the remaining 2,391 are multiple choice. Per-task share and question wording are in Figure 11.

## 4 Framework

This section introduces the T4D network and the training recipe TAP. Section 4.1 specifies the network, Section 4.2 the position-addressed fusion of tracks with vision (PAF), and Section 4.3 the trajectory-aligned training recipe (TAP).

### 4.1 Model Architecture

The T4D network uses a frozen SpatialTrackerV2 as its 3D trajectory expert and Qwen3-VL as the video backbone, with a small trainable fusion stack between them (Figure 7). On the appearance side, we take the vision tokens of Qwen3-VL *before* its multimodal merger, $V \in \mathbb{R}^{T\times H\times W\times d}$; the merger

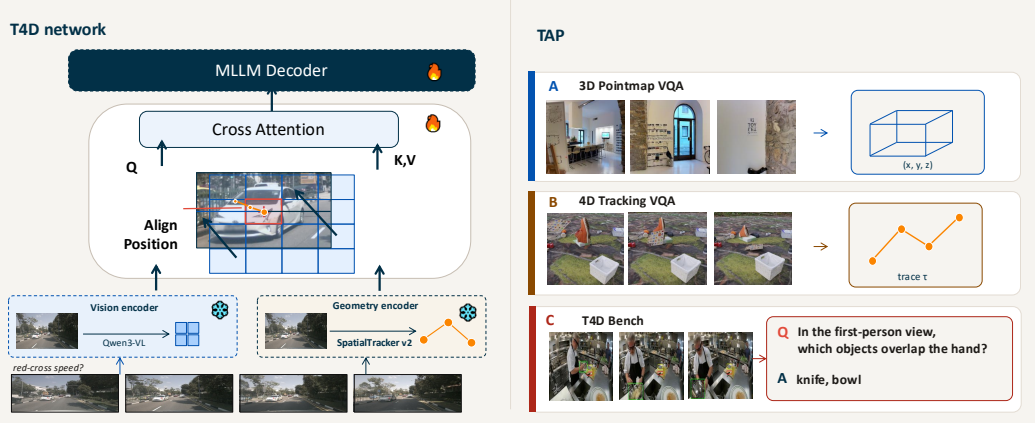



**Figure 7: The T4D network and TAP.** A frozen vision encoder and a frozen 4D tracker are fused on the image plane, and the language model reads the result. TAP aligns this model to pointmaps, tracks, and spatial questions.

and the language decoder sit after fusion. On the geometry side, SpatialTrackerV2 is used off the shelf and kept frozen: we run it once per clip and cache its per-frame sync tokens, each the concatenation of a 2D and a 3D feature, $G_{\text{raw}} \in \mathbb{R}^{T\times L\times 768}$. A learned projection $\text{token_proj} : \mathbb{R}^{768} \to \mathbb{R}^{d}$ maps them to the vision width ($d$=1024 on the 4B backbone), giving $G \in \mathbb{R}^{T\times L\times d}$. The bank is the image-plane grid of tracking queries $\mathcal{Q}$ (as in SpatialTrackerV2), followed by the visual-odometry (VO) support tracks used for camera motion, which are not bound to that grid. How these two streams are matched and written back is the subject of Section 4.2. Only the projection, the fusion block, and a LoRA adapter on the LLM ($r$=64, $\alpha$=128) receive gradients, so the 4D evidence available to the model is fixed by the tracker rather than relearned.

### 4.2 Position-Addressed 4D Fusion

A vision token carries its image position; the co-located track token does not, so a query that has grounded the object still cannot index that track, and plain cross-attention matches the two only by content. PAF writes the missing position into the track bank, making the correspondence available to the attention rather than guaranteeing that every query selects its co-located track.

$$\text{PE}_\ell = \begin{cases} \text{MLP}(u_\ell, v_\ell), & \ell \in \mathcal{Q}, \\ e^{\text{vo}} + E^{\text{idx}}_\ell, & \ell \in \text{VO}, \end{cases} \qquad \hat{G} = G + \text{PE}, \tag{1}$$

where $(u_\ell, v_\ell)$ are the normalized image coordinates of grid query $\ell \in \mathcal{Q}$. A VO support track is not bound to the query grid $\mathcal{Q}$—SpatialTrackerV2 samples these points to stabilize camera motion rather than to mark a fixed image location—so it receives a learned type vector $e^{\text{vo}}$ and an index embedding $E^{\text{idx}}$, and enters the match as global motion context rather than as co-located local evidence. Vision tokens then attend to this position-tagged bank (Figure 7),

$$\tilde{V} = V + \text{MHA}\big(\text{LN}(V),\, \text{LN}(\hat{G}),\, \text{LN}(\hat{G})\big), \tag{2}$$

with appearance as queries and tracks as keys and values. The residual writes the retrieved geometry back onto the vision tokens, so the language model keeps reading the appearance stream with the attended traces mixed in. Two Transformer blocks then mix the fused grid, and Qwen3-VL's spatial merger lifts it into the LLM width.

### 4.3 Trajectory-Aligned Training

Even with PAF, the residual path can still answer from appearance and leave the tracks unread. GAP-MLLM (Zhang et al., 2026) closes this shortcut on static scenes with a visual cross and pointmap VQA. In dynamic scenes, however, pointmap VQA does not align the vision tokens with the trajectory tokens: the language head can still produce the metric answer from 2D appearance alone. TAP therefore adds trajectory supervision on top of that recipe (Figure 7). The stage-to-axis map is in Table 3.

**Table 1: Test results on T4D-BENCH (%).** Open-weight and proprietary general models are zero-shot. Qwen3-VL-4B and the specialist spatial models are finetuned on the same data as T4D. **Bold** marks the highest score within each block. △ means the parameter count is not publicly disclosed.

| Model | Params | Abs. Dist. | Abs. Speed | Dist. Trend | Rel. Size | Rel. Speed | Obj. Count | Cam. Disp. | Cam. Rot. | Persp. Match | Ego Contact | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Chance* | – | 17.1 | 16.0 | 20.0 | 25.0 | 24.7 | 4.0 | 20.0 | 20.5 | 33.4 | 25.0 | 20.6 |
| *Zero-shot open models (official split)* | | | | | | | | | | | | |
| Qwen3-VL (Bai et al., 2025) | 8B | **18.1** | 15.5 | **32.2** | 40.8 | 32.0 | 27.7 | **51.5** | 34.0 | **40.4** | 41.5 | **33.4** |
| MiMo-VL (LLM-Core Team Xiaomi, 2025) | 7B | 0.3 | 8.0 | 27.6 | **48.4** | **33.3** | 36.3 | 51.0 | **39.5** | 36.2 | 37.5 | 31.8 |
| InternVL3 (Chen et al., 2024b) | 8B | 0.9 | 12.1 | 23.8 | 37.8 | 29.2 | **37.2** | 43.0 | 17.0 | **40.4** | 39.0 | 28.0 |
| LLaVA-NeXT (Zhang et al., 2024) | 7B | 10.7 | **24.3** | 30.0 | 25.0 | 21.6 | 21.1 | 25.0 | 12.5 | 31.0 | 34.0 | 23.5 |
| GLM-4.1V-Thinking (Hong et al., 2025) | 9B | 15.8 | 9.9 | 24.2 | 43.0 | 28.9 | 19.5 | 22.5 | 22.0 | 37.8 | **49.0** | 27.3 |
| *Zero-shot proprietary models* | | | | | | | | | | | | |
| GPT-6 Astra | △ | **79.8** | **65.8** | **56.4** | 78.2 | **75.9** | **76.4** | **98.5** | **75.0** | **95.4** | **79.0** | **78.0** |
| GPT-5.6 Sol | △ | 41.5 | 35.4 | 43.8 | **78.6** | 50.9 | 69.7 | 96.5 | 54.5 | 63.0 | 59.0 | 59.3 |
| Gemini 3.8 Flash | △ | 44.6 | 37.7 | 42.4 | 69.8 | 47.4 | 69.0 | 85.5 | 49.5 | 87.4 | 50.5 | 58.4 |
| Kimi K3 | 2800B | 32.6 | 34.6 | 30.4 | 65.8 | 44.7 | 65.3 | 56.5 | 40.0 | 62.4 | 60.5 | 49.3 |
| *Finetuned on the same data as T4D* | | | | | | | | | | | | |
| Qwen3-VL-4B (Bai et al., 2025) | 4B | 49.1 | 28.9 | 42.4 | **58.4** | 38.8 | 40.1 | 58.5 | 47.0 | 37.0 | **65.0** | 46.5 |
| VLM-3R (Fan et al., 2025) | 7B | **59.6** | **41.6** | 43.0 | 50.2 | **47.1** | 28.9 | 57.0 | 49.0 | 55.0 | 60.0 | **49.1** |
| MLLM-4D (Yin et al., 2026) | 7B | 55.8 | 39.6 | 41.0 | 55.0 | 45.4 | **43.0** | 49.0 | 48.5 | 40.0 | 57.5 | 47.5 |
| VG-LLM (Zheng et al., 2025) | 8B | 58.5 | 40.5 | **46.6** | 44.4 | 44.0 | 26.6 | 55.0 | 38.0 | 36.8 | 56.0 | 44.6 |
| Spatial-MLLM (Wu et al., 2025) | 3B | 35.8 | 27.7 | 36.8 | 25.6 | 31.3 | 21.4 | **62.5** | **50.5** | **60.8** | 64.0 | 41.6 |
| **Ours (T4D)** | 4B | **66.5** | **42.0** | **52.0** | **74.2** | **51.9** | **42.8** | **76.5** | **51.5** | **79.2** | **73.0** | **61.0** |

**Stage A** is only a scale warm-up: 10k pointmap QA pairs from 3RScan (Wald et al., 2019), ARKitScenes (Baruch et al., 2021), and Matterport3D (Chang et al., 2017), with a GAP-style metric $(x, y, z)$ in a locked first-frame camera, so fusion must use SpatialTrackerV2's metric structure rather than 2D scale heuristics. **Stage B** trains 4D tracking on 15k trajectory QA pairs from PointOdyssey (Zheng et al., 2023), nuScenes (Caesar et al., 2020), Kubric (Greff et al., 2022), Waymo (Sun et al., 2020), and Dynamic Replica (Karaev et al., 2023). It uses the same cross as an instance query: on a $T$-frame clip we sample a mark index $K$, draw the cross only on frame $K$, lock a right–down–forward camera at $K$, and ask for the marked object's 3D centre from $K$ through $T{-}1$,$\tau = \big[(x_K, y_K, z_K), \ldots, (x_{T-1}, y_{T-1}, z_{T-1})\big]$

optionally prefixed by a short object name. Emitting $\tau$ binds the cross to an appearance identity, reads a metric track rather than image-plane flow, and holds the declared frame while the ego camera moves. The **Stage B** data pipeline is deferred to Appendix A.4 (Figure 10) for space. The nuScenes and Waymo clips used to build this set are completely disjoint from T4D-BENCH. **Stage C** finetunes on the spatial tasks of Section 3.1 without changing the architecture: the head that learned to speak traces is asked for the derived quantities those traces determine. Motivated by the mismatch test of 3D-RDQA (Li et al., 2025a), *TrajDQA* randomly replaces the 4D features of a clip with those of another video while leaving the RGB unchanged. The model is then asked to output the text `INCONSISTENT` rather than the ground-truth trajectory $\tau$, so the language head must read the supplied 4D features and cannot produce the answer from 2D appearance alone.

# 5 EXPERIMENTS AND ANALYSIS

## 5.1 SETUP

We evaluate on the official ten-task split of T4D-BENCH. Multiple-choice tasks use exact-match accuracy. Numeric tasks (`absolute_distance`, `absolute_speed`, `object_counting`) use mean relative accuracy (MRA). Exact match is a poor fit for a continuous answer: given a prediction $\hat{y}$, a ground truth $y$, and a confidence threshold $\theta$, $\hat{y}$ is counted correct when the relative error $|\hat{y} - y|/|y|$ is at most $1 - \theta$. One threshold only checks a single tolerance, so MRA averages this indicator over $\mathcal{C} = \{0.5, 0.55, \ldots, 0.95\}$: $\text{MRA} = \frac{1}{|\mathcal{C}|} \sum_{\theta \in \mathcal{C}} \mathbf{1}(|\hat{y} - y|/|y| \leq 1 - \theta)$, where $\theta$=0.95 allows at most a 5% relative error and $\theta$=0.5 allows 50%. When $y$=0, MRA is 1 if $\hat{y}$=0 and 0 otherwise. A task score is the mean of per-question MRA, reported in percent. Overall is the mean of the ten task scores. *Chance* is a video-blind guess on the same items. Every model in Table 1, including the proprietary ones, uses this same resolution: frames are rescaled with the aspect ratio preserved so that the longer side is 512, with at most 128 frames per clip.

The T4D network uses Qwen3-VL-4B as the video backbone and frozen SpatialTrackerV2 as the trajectory expert, joined by PAF (Section 4). The grid size is 6 and the VO size is 189. Only

**Table 2: Ablations on T4D-BENCH (%).**

| Variant | Abs. Dist. | Abs. Speed | Dist. Trend | Rel. Size | Rel. Speed | Obj. Count | Cam. Disp. | Cam. Rot. | Persp. Match | Ego Contact | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Backbone, same data* | | | | | | | | | | | |
| Qwen3-VL-4B | 49.1 | 28.9 | 42.4 | 58.4 | 38.8 | 40.1 | 58.5 | 47.0 | 37.0 | 65.0 | 46.5 |
| *Training recipe* | | | | | | | | | | | |
| w/o Stage A | 61.7 | 38.5 | **52.2** | 65.4 | **55.3** | 41.5 | 73.0 | 44.0 | 72.8 | 66.0 | 57.0 |
| w/o Stage B | 59.1 | 26.9 | 50.4 | 70.2 | 47.8 | 42.6 | 74.5 | 41.0 | 76.2 | 64.5 | 55.3 |
| w/o Stage A/B | 44.6 | 29.2 | 47.4 | 62.4 | 49.8 | 43.8 | 72.0 | 37.5 | 77.4 | 60.5 | 52.5 |
| w/o TrajDQA | 61.7 | 41.5 | 48.2 | 72.6 | 54.0 | 40.5 | 72.5 | 46.0 | 63.8 | **73.0** | 57.4 |
| *Fusion, on full TAP* | | | | | | | | | | | |
| w/o PAF | 62.2 | 39.2 | 51.4 | 73.4 | 51.2 | **44.9** | 75.0 | 49.5 | 71.6 | 68.5 | 58.7 |
| **Ours (T4D)** | **66.5** | **42.0** | 52.0 | **74.2** | 51.9 | 42.8 | **76.5** | **51.5** | **79.2** | **73.0** | **61.0** |

token_proj, the fusion block, and a LoRA adapter on the LLM ($r$=64, $\alpha$=128) are trained. We use AdamW with a cosine schedule and 3% warmup on 16 GPUs (ZeRO-2). Alignment stages use a peak learning rate of $5\times10^{-6}$; Final finetune stage(Stage C) uses a peak learning rate of $1\times10^{-5}$.

### 5.2 RESULT

Our 4B model reaches Overall 61.0, against 46.5 for the same-data Qwen3-VL-4B backbone without a trajectory expert (Table 1). It trails GPT-6 Astra (78.0), yet surpasses every other proprietary or open-weight general model in the table—GPT-5.6 Sol (59.3), Gemini 3.8 Flash (58.4), Kimi K3 (49.3), and the best open model (33.4)—despite a far smaller parameter count. The specialists trained on the same data remain behind as well (best 49.1).

The gap is not uniform across tasks. Specialists nearly match us on absolute speed (41.6 against 42.0) and camera rotation (50.5 against 51.5). The largest lead is on perspective matching (79.2 against 60.8). Counting is the one column a same-data specialist leads (MLLM-4D, 43.0 against 42.8).




**Figure 8: Predicted trace against ground truth and SpatialTrackerV2.** Left: ground truth. Middle: SpatialTrackerV2, queried with the point coordinate. Right: the trace parsed from the text generated by T4D.

### 5.3 ABLATION

Against this backbone, the largest gain is perspective matching (37.0 to 79.2). Dropping Stage A hurts comparisons that can be read from scene scale, most clearly relative size (74.2 to 65.4), while dropping Stage B hurts the questions that follow a marked instance: absolute distance, absolute speed, and distance trend (52.0 to 50.4). Dropping TrajDQA lowers absolute distance (66.5 to 61.7) and absolute speed (42.0 to 41.5). Ablating PAF while keeping the full TAP recipe reduces Overall by 2.3 points (61.0 to 58.7), mostly on perspective matching (79.2 to 71.6) and egocentric contact (73.0 to 68.5). Asking T4D to generate the 3D trajectory yields a comparable track, with ADE 10.45 m versus 8.36 m for SpatialTrackerV2 (Figure 8). This performance shows that the alignment is effective and that the language model already uses nearly the full capability of the 4D encoder.

## 6 CONCLUSION

T4D studies dynamic spatial intelligence. T4D-BENCH is the first benchmark to evaluate it along all six dimensions, with 10 tasks and 3,250 expert-verified QA pairs, and the same automated pipeline yields about 600K training pairs from real indoor and outdoor video. The T4D network fuses explicit 3D point trajectories of instances with vision and attains state-of-the-art on T4D-BENCH among MLLMs of comparable scale. Trajectory-aligned training (*TAP*) supervises those trajectories in a fixed reference frame, aligning 2D vision tokens with the trajectory tokens and improving dynamic metric reasoning.

**Limitation.** The state-of-the-art claim holds among MLLMs of comparable scale and does not cover frontier proprietary models: GPT-6 Astra still leads on every task. The perspective tasks are built only on EgoLife and EgoExo4D, so this part of the benchmark still needs more data.

## AI USE STATEMENT

In this work, we used Grok 4.7 and GPT-5.5 for the data-construction pipeline and to draft the paper and check grammar. Grok 4.7 was also used for data-quality inspection and model construction. Kimi K3, GPT-6, and Gemini 3.8 were used only for benchmark evaluation. GPT-5.5 was also used for benchmark evaluation. We have not used generative AI tools to develop theoretical or conceptual frameworks, design the experimental methodology, or produce the interpretation of results reported in the paper. We have reviewed all AI-assisted work. Professional annotators manually screened every benchmark annotation under a written protocol. We inspected all AI-written code and implementation details. After the draft, the main text was written by human authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## REFERENCES

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xuejing Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Gao, Wenbin Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-D data. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Zhipeng Cai, Ching-Hui Yeh, Hao Xu, Zhuang Liu, Gregory Meyer, Xun Lei, Chen Zhao, Shang Li, Vikas Chandra, and Yangyang Shi. DepthLM: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *International Conference on 3D Vision*, 2017.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Danny Driess, Pete Florence, 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, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

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

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Ye, Peihao Duan, Hongyan Li, Yixing Wang, Zhengzhong Chen, Dejia Wang, Zhen Huang, et al. VLM-3R: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J. Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomeng Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*, 2023.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Haoyuan Li, Yanpeng Zhou, Yufei Gao, Tao Tang, Jianhua Han, Yujie Yuan, Dave Zhenyu Chen, Jiawang Bian, Hang Xu, and Xiaodan Liang. Does your 3D encoder really work? when pretrain-SFT from 2D VLMs meets 3D VLMs. *arXiv preprint arXiv:2506.05318*, 2025a.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, 2023.

LLM-Core Team Xiaomi. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Xiaojian Ma, Silong Yong, Ziyu Zheng, Qing Li, Yitao Liang, Song-Chun Li, and Siyuan Zhu. SQA3D: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Sujith Nair, Dhruv Parikh, et al. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Mishra, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, 2024.

Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. RIO: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. 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.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.

Pan Wang, Yang Liu, Guile Wu, Eduardo R. Corral-Soto, Chengjie Huang, Binbin Xu, Dongfeng Bai, Xu Yan, Yuan Ren, Xingxin Chen, et al. Spatial4D-Bench: A versatile 4D spatial intelligence benchmark. *arXiv preprint arXiv:2601.00092*, 2026b.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Shuai Wang, Daoan Zhang, Tianyi Bai, Shitong Shao, Jiebo Luo, and Jiaheng Wei. LAST: LeArning to think in space and time for generalist vision-language models. *arXiv preprint arXiv:2511.19261*, 2025c.

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

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. SpatialTrackerV2: 3D point tracking made easy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10632–10643, 2025a.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, Lei Yang, Bo Li, and Ziwei Liu. EgoLife: Towards egocentric life assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28885–28900, 2025b.

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. *arXiv preprint arXiv:2506.21458*, 2025.

Xingyilang Yin, Chengzhengxu Li, Jiahao Chang, Chi-Man Pun, and Xiaodong Cun. MLLM-4D: Towards visual-based spatial-temporal intelligence. *arXiv preprint arXiv:2603.00515*, 2026.

Jiaxin Zhang, Junjun Jiang, Haijie Li, Youyu Chen, Kui Jiang, and Dave Zhenyu Chen. GAP-MLLM: Geometry-aligned pre-training for activating 3D spatial perception in multimodal large language models. *arXiv preprint arXiv:2603.16461*, 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. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024.

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*, 2025.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3D world: Enhancing MLLMs with 3D vision geometry priors. In *Advances in Neural Information Processing Systems*, 2025.

Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J. Guibas. PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9637–9646, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. OmniWorld: A multi-domain and multi-modal dataset for 4D world modeling. *arXiv preprint arXiv:2509.12201*, 2025b.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering LMMs with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4295–4305, 2025.