# LMM-TRACK4D: ELICITING 4D DYNAMIC REASONING IN LMMS VIA TRAJECTORY-GROUNDED DIALOGUE

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

## ABSTRACT

Large multimodal models (LMMs) describe dynamic scenes, but grounding successive questions in an entity's changing 3D position requires a representation that connects dialogue references to physical time. We formulate trajectory-grounded multi-turn spatiotemporal dialogue, where a model answers questions and returns metric positions over the requested interval. We introduce Track4D-Bench to associate these dialogue references with calibrated observations and timestamped trajectories across 526 clips, 23.5k frames, and 7.5k entity tracks. To retain the queried entity throughout a conversation, we propose LMM-Track4D with a geometry-conditioned target state. This state preserves entity context across turns and selects spatial and temporal evidence for recovering the entity's trajectory. The resulting representation separates dialogue order from the physical times being queried, allowing later questions to revisit earlier intervals. On Track4D-Bench, LMM-Track4D improves CIDEr by 0.247 over the strongest same-data supervised fine-tuning baseline, alongside gains of 10.2 and 10.5 points in BLEU-4 and METEOR. Geometric evaluation shows a 19.4 percentage-point gain in trajectory-step accuracy over direct coordinate regression. External evaluations further assess dynamic-scene question answering and language-conditioned grounding. These results support geometry-conditioned target states for connecting persistent entity references with metric trajectory reasoning in LMMs. Our code and dataset will be publicly available.

## 1 INTRODUCTION

Large multimodal models (LMMs) have advanced image and video understanding (Jin et al., 2026b;a; 2025a; 2026c;d; Wen et al., 2025; Wang et al., 2025c; Wen et al., 2026a; Ke et al., 2026; Wen et al., 2026b), with recent work extending their representations to 3D position and time (Wang et al., 2025b; Zhou et al., 2025; Zhou & Lee, 2025a; 2026; Huang et al., 2026b). These representations support **4D dynamic reasoning** about entities moving through physical scenes. In a conversation, reasoning about an entity also requires preserving its reference as questions shift between events, positions, and time intervals. A question about a pass, for example, can be followed by a request for the ball's trajectory during that event.

Such follow-up questions connect two temporal orders. Dialogue advances turn by turn, while the requested positions follow physical time and can revisit an earlier interval. The entity reference connects these orders: the model must retain which entity is being discussed and recover its positions at the times specified by the current question. Video-language models encode observed content (Li et al., 2023; 2025b; Zhang et al., 2024a; Wang et al., 2025a), and dynamic-scene models represent motion and temporal relations (Huang et al., 2026b; Yan et al., 2025; Yin et al., 2026). Our focus is how this visual and temporal evidence supports a persistent entity reference and explicit metric answers across dialogue turns.

To study this relationship, we formulate **trajectory-grounded multi-turn spatiotemporal dialogue**. Each turn asks about a dynamic scene and, when required, requests structured 3D positions over a complete short clip or a local interval of a longer clip. The task grounds verbal answers and

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.

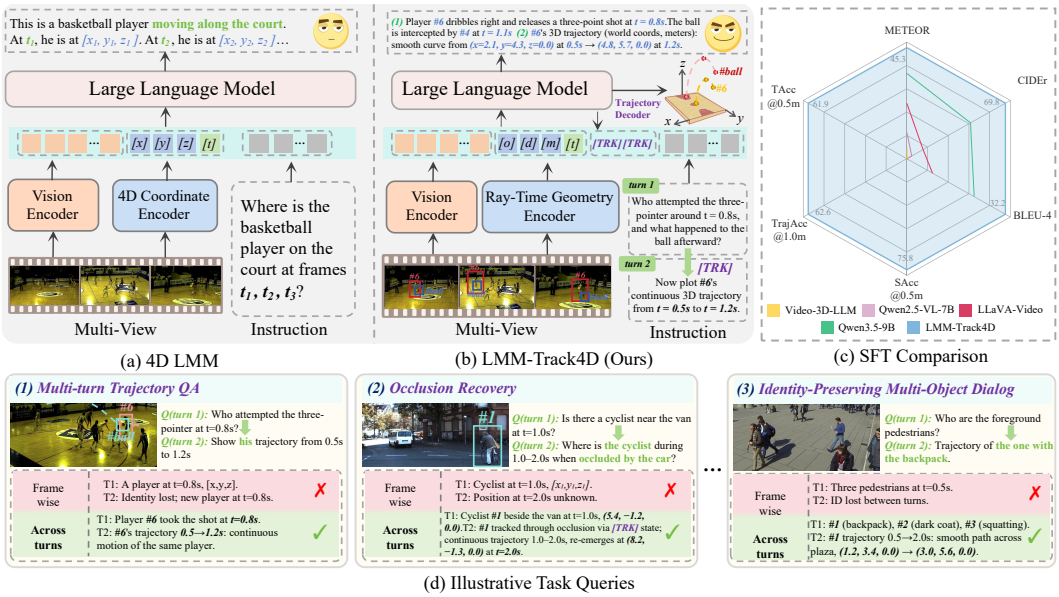



Figure 1: Trajectory-grounded dialogue and LMM-Track4D. Panels (a), (b), and (d) illustrate the task; check and cross marks contrast frame-wise answers with the requested cross-turn behavior. The Ray–Time Geometry Encoding (RTGE) icons denote camera origin $o$, ray direction $d$, moment $m$, and time $t$. Panel (c) compares the same-data fine-tuning controls in Table 1. Radial distances are normalized to LMM-Track4D separately for each metric; annotations show its language scores multiplied by 100 and geometric accuracies in percent. Unreported baseline geometric scores are omitted.

trajectories in the same entity and temporal context, as illustrated in Figure 1. It therefore makes the positions associated with a dialogue reference directly evaluable.

We construct **Track4D-Bench** by associating dialogue references with timestamped entity trajectories and calibrated visual observations from APIDIS, KITTI, and WildTrack (Castanedo et al., 2011; Geiger et al., 2013; Chavdarova et al., 2018). The benchmark contains 526 clips spanning 23.5k frames and 7.5k entity tracks. Its Short regime pairs complete-event trajectories with qualitative questions, while its Long regime requests local positions and trajectory segments within broader temporal context. Both regimes evaluate verbal responses and the metric geometry requested by the dialogue.

We propose **LMM-Track4D** to represent the queried entity through a geometry-conditioned target state. Ray–Time Geometry Encoding (RTGE) embeds calibrated rays and observation times into visual features. A persistent TRK state token retains the resulting entity context across dialogue turns and updates it for the current request. The updated state selects observations for the Object-Slot Kinematic, Residual-Anchor (OSK-RA) decoder, which estimates a geometric anchor and refines it into each requested position. Position supervision shapes this evidence selection, connecting the retained target state to metric trajectory recovery.

On Track4D-Bench, LMM-Track4D improves BLEU-4 and METEOR by 10.2 and 10.5 points over Qwen3.5-9B fine-tuned on the same data. Component ablations examine how geometric encoding, target memory, and residual-anchor decoding contribute to these language and trajectory outputs. External evaluation on Dyn-Bench also shows gains over the strongest listed models in each interaction category, reaching 2.06 percentage points for question answering and 4.83 for language-conditioned object grounding. Our contributions are as follows:

- We formulate **trajectory-grounded multi-turn spatiotemporal dialogue**, connecting persistent entity references to verbal answers and metric trajectories at requested physical times.

- We construct **Track4D-Bench** with linked entity identities, calibrated observations, trajectories, and dialogue supervision across complementary temporal scopes.
- We propose **LMM-Track4D**, which retains geometry-conditioned target states across turns and uses them to select visual evidence for metric trajectory decoding.
- We demonstrate experimentally that this design improves dialogue answering and trajectory recovery, with component analyses and external evaluations of dynamic-scene reasoning and grounding.

## 2 Related Work

**Geometry and dynamics in LMMs.** Video-language models learn temporal representations for visual dialogue (Zhang et al., 2024a; Lin et al., 2024). Geometric representations extend this visual context to physical space: Video-3D LLM and LLaVA-3D associate visual features with metric coordinates (Zheng et al., 2025; Zhu et al., 2025), while LLaVA-4D incorporates position and time through spatiotemporal prompts (Zhou & Lee, 2026). Motion-aware models further encode dynamic-scene information (Huang et al., 2026b; Yin et al., 2026). LMM-Track4D uses geometric and temporal context to form a target state that supports successive requests for an entity's positions.

**Entity references and trajectories.** Ground3D-LMM combines metric grounding with multi-turn spatial conversation (Harsh et al., 2026). MASS, Motion-o, and DynTrace make motion evidence available to language reasoning (Wu et al., 2025; Galoaa et al., 2026; Gao et al., 2026); Mono3DVLT follows a language-specified target in 3D (Wei et al., 2025), and ObjectStream retains object histories across video frames (Dong et al., 2026). In trajectory-grounded dialogue, each turn can change the requested interval while retaining the entity reference. LMM-Track4D updates a geometry-conditioned target state in dialogue order and decodes positions in the requested physical-time order. Track4D-Bench links these references to metric supervision. Appendix G discusses the broader grounding and dynamic-scene literature.

## 3 Track4D-Bench

**Task and construction.** Dialogue turns ask about scene content or specific entities. Entity questions identify targets through descriptions, visible identifiers, spatial conditions, or earlier references, and

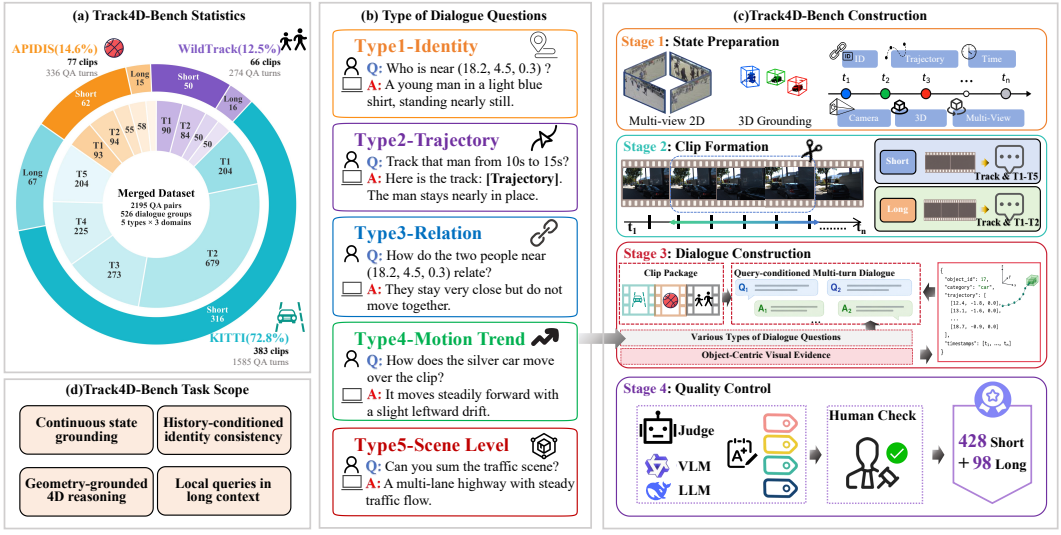



Figure 2: Track4D-Bench overview: source composition, five question types, and state-to-dialogue construction. In panel (a), a dialogue group is one clip and a QA pair is one turn. Panel (d) summarizes task properties, including continuous state grounding and entity references across questions. Section 3 defines the language and geometric metrics.

request verbal answers, positions, or trajectories. Track4D-Bench associates these requests with source identities, timestamps, calibrated observations, and metric positions from APIDIS, KITTI, and WildTrack. The positions use ground locations for APIDIS and WildTrack people, centers for APIDIS basketballs, and box-bottom centers for KITTI objects, all expressed in a fixed clip coordinate system in meters. These state records ground dialogue generation in the source geometry. Record checks, language and vision-language model filtering, and author spot checks validate the resulting dialogues (Figure 2; Appendix B).

The benchmark contains 428 Short and 98 Long clips. Short questions cover complete compact events; Long questions request local positions or intervals within broader context. In both regimes, requested timestamps can fall between observed images. Each entity track is counted once per clip, and each frame is a camera–timestamp image record. Appendix A gives these counts by source.

**Evaluation.** CIDEr, BLEU-4, and METEOR score verbal responses after removing structured coordinates (Vedantam et al., 2015; Papineni et al., 2002; Banerjee & Lavie, 2005). BLEU-4 and METEOR are reported on a 0–100 scale; CIDEr retains its original scale. Geometric eligibility follows the requested output and valid references, independently of semantic question type. SAcc@0.5 measures the fraction of point requests whose mean target error is below 0.5 m; Traj-Acc@1.0 measures the fraction of trajectory requests with mean error below 1.0 m; TAcc@0.5 measures the fraction of valid reference steps with error below 0.5 m. Errors use world-coordinate XYZ. Invalid predictions remain in the denominators as failures. Appendix C defines alignment and micro averaging.

# 4 METHOD

## 4.1 GEOMETRY-CONDITIONED TARGET STATES

The queried entity connects dialogue context to geometric estimation through a retained target state. At turn $k$, the model receives question $q_k$, the preceding questions and generated answers $\mathcal{H}_{k-1}$, and calibrated observations $\mathcal{X} = \{I_{v,\tau}, K_{v,\tau}, E_{v,\tau}\}_{v,\tau}$. Here $I_{v,\tau}$ is an image from camera $v$ at physical time $\tau$, with intrinsics $K_{v,\tau}$ and extrinsics $E_{v,\tau}$. The model generates answer $a_k$, updates the entity's latent target state $\mathbf{s}_k$, and estimates world positions $\hat{\mathcal{Y}}_k = \{\hat{\mathbf{p}}_{k,h} \in \mathbb{R}^3\}_{h=1}^{H}$ at $H$ requested timestamps. For multiple targets, these outputs follow query order. Figure 3 shows how RTGE encodes the observations, TRK retains the target state, and OSK-RA decodes its trajectory.

The target state is updated in dialogue order, while decoded positions follow the requested physical times. When a later question revisits an earlier interval, the retained state supplies entity context and the interval determines which positions to recover.

## 4.2 RTGE: RAY-TIME GEOMETRY ENCODING

Calibration places observations from different views and times in a common geometric frame. For patch $n$ with visual feature $\mathbf{f}_{v,\tau,n} \in \mathbb{R}^d$, the camera parameters determine the ray origin $\mathbf{o}_{v,\tau}$ and unit direction $\mathbf{d}_{v,\tau,n}$. RTGE encodes this ray by its Plücker descriptor $\mathbf{r}_{v,\tau,n} = [\mathbf{d}_{v,\tau,n}; \mathbf{o}_{v,\tau} \times \mathbf{d}_{v,\tau,n}]$. Its second term, the ray moment, specifies the supporting line together with the direction. A Fourier embedding $\boldsymbol{\gamma}(\tau) = [\sin(\omega_i\tau), \cos(\omega_i\tau)]_{i=1}^{m}$ encodes the observation time using $m$ frequencies $\omega_i$. Learned projections $\phi_r$ and $\phi_t$ map both embeddings to the visual feature dimension and add them to the token:

$$\tilde{\mathbf{f}}_{v,\tau,n} = \mathbf{f}_{v,\tau,n} + \phi_r(\mathbf{r}_{v,\tau,n}) + \phi_t(\boldsymbol{\gamma}(\tau)). \tag{1}$$

The resulting features $\mathcal{F} = \{\tilde{\mathbf{f}}_{v,\tau,n}\}_{v,\tau,n}$ condition the target state on where and when each patch was observed. Each feature retains its association with the calibrated ray, so selecting visual evidence also selects the corresponding line constraint for geometric decoding. Appendix C.3 gives the projection settings.

## 4.3 TRK STATE-MEMORY TOKEN

TRK retains a target state for each entity referenced in the dialogue. A reference table associates the entity's descriptions and aliases with a local key, allowing later mentions to retrieve the same state.

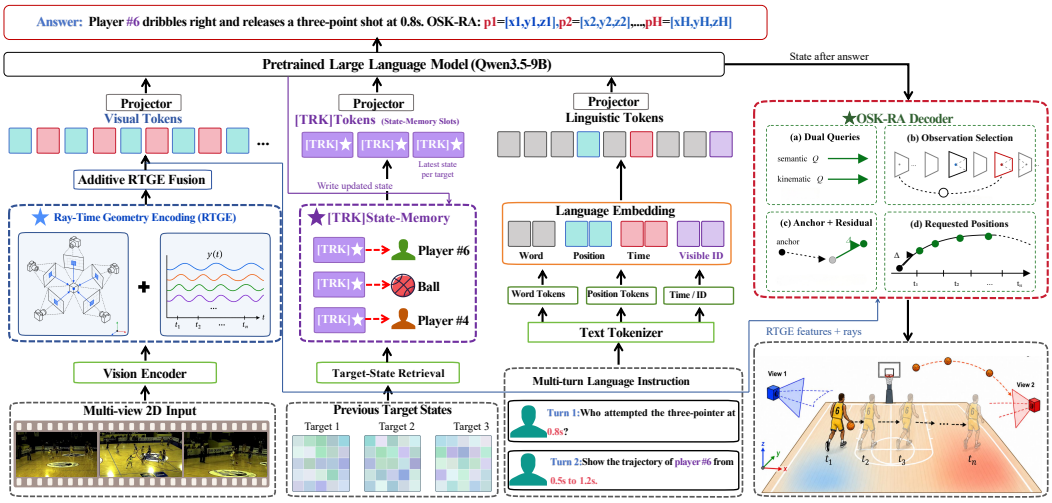



Figure 3: LMM-Track4D overview. RTGE adds calibrated ray and time embeddings to visual features. TRK retrieves the queried entity's latest state and writes its update after the language response, including any located interval. Position, time, and visible identifiers denote tokenized question content. OSK-RA receives the updated target state (black route) and RTGE features with calibrated rays (blue route) to select observations and predict positions through anchors and residuals. The purple route retains updated states for later turns. The dialogue and trajectory panels are schematic.

The backbone resolves each target from the question, generated history, and observations, then reads its latest state. A new entity starts from a learned initial state and is distinguished by its resolved description $\rho_k^{(e)}$, where $e$ denotes the local entity key.

Let $\mathcal{R}_{k-1}$ denote the retained reference table. Binding selects the local key through

$$e_k = \mathcal{B}_\theta(q_k, \mathcal{H}_{k-1}, \mathcal{F}, \mathcal{R}_{k-1}). \tag{2}$$

Here $\mathcal{B}_\theta$ denotes reference binding by the language backbone. The selected key identifies the state to retrieve; the current answer supplies the temporal context needed to update it before position decoding.

The backbone generates $a_k$, including any interval to be located, before processing a trailing TRK write token:

$$\mathbf{s}_k^{(e)} = U_\theta\left(\mathbf{s}_{k-1}^{(e)}, \rho_k^{(e)}, q_k, \mathcal{H}_{k-1}, \mathcal{F}, a_k\right). \tag{3}$$

The backbone computation $U_\theta$ reads $\mathbf{s}_{k-1}^{(e)}$, the latest state for entity $e$, even when the preceding turn concerned another entity. Writing after the response incorporates the current question and any located interval into $\mathbf{s}_k^{(e)}$. The updated state conditions OSK-RA and is retained for later references. For multiple targets, separate descriptions condition states in query order, aligning verbal and geometric outputs by entity. Generated history and target memory reset between clips. Source identities provide training correspondence and scoring; inference resolves local keys from visible inputs. Appendix C.2 specifies memory management and unresolved references.

### 4.4 OSK-RA: Evidence Selection and Trajectory Decoding

For each bound entity, OSK-RA uses the updated target state $\mathbf{s}_k$, suppressing the entity index $e$, to estimate $H$ positions on the source time grid within the requested interval. An explicit interval comes from the question; an interval that must be located comes from $a_k$. Learned semantic and kinematic projections, $\mathbf{q}_k^{\text{sem}} = W_{\text{sem}}\mathbf{s}_k$ and $\mathbf{q}_k^{\text{kin}} = W_{\text{kin}}\mathbf{s}_k$, score the visual evidence separately for each output step $h$:

$$z_{k,h,r} = g_{\text{score}}^h(\mathbf{q}_k^{\text{sem}}, \mathbf{q}_k^{\text{kin}}, \tilde{\mathbf{f}}_r), \qquad r = (v, \tau, n). \tag{4}$$

Here $r$ indexes a patch observation, and $g^h_{\text{score}}$ is its scoring function for step $h$. Each image contributes its highest-scoring valid patch. Confidence-based selection retains a set $\mathcal{S}_{k,h}$ of these observations, and their normalized scores weight the fused feature:

$$\alpha_{k,h,r} = \frac{\exp z_{k,h,r}}{\sum_{j\in\mathcal{S}_{k,h}} \exp z_{k,h,j}}, \qquad \mathbf{z}_{k,h} = \sum_{r\in\mathcal{S}_{k,h}} \alpha_{k,h,r}\tilde{\mathbf{f}}_r.$$

The same weights locate a coarse geometric anchor. With $P_r = I - \mathbf{d}_r\mathbf{d}_r^\top$, weighted line intersection gives

$$\hat{\mathbf{p}}^{\text{anchor}}_{k,h} \in \arg\min_{\mathbf{p}\in\mathbb{R}^3} \sum_{r\in\mathcal{S}_{k,h}} \alpha_{k,h,r} \left\| P_r(\mathbf{p} - \mathbf{o}_r) \right\|_2^2 . \tag{5}$$

The minimum-norm solution gives a coarse metric anchor from the selected line constraints. These constraints can come from different observation times, across which the entity moves. The fused feature retains their visual and temporal information, allowing a learned residual $g_{\text{res}}$ to correct the anchor for the requested step:

$$\hat{\mathbf{p}}_{k,h} = \hat{\mathbf{p}}^{\text{anchor}}_{k,h} + g_{\text{res}}([\mathbf{z}_{k,h}; \hat{\mathbf{p}}^{\text{anchor}}_{k,h}]). \tag{6}$$

All output steps share the target state and residual network, while their scoring channels select and weight evidence separately. Each position is decoded independently of preceding predicted positions. With multiple retained rays, position supervision reaches the target state through the normalized weights used for feature fusion and anchor estimation. This dependence trains the state to select evidence for the referenced entity. Appendix C.3 gives the selection rule, solver, and single-ray case.

### 4.5 Training

LMM-Track4D jointly learns to answer questions and estimate the associated geometry. Geometry-oriented supervision first initializes the target representation and decoder. Joint supervision then connects the verbal answer, the requested positions, and the retained entity state through

$$\mathcal{L} = w_{\text{text}}\mathcal{L}_{\text{NLL}} + w_{\text{traj}}\mathcal{L}_{\text{traj}} + w_{\text{sm}}\mathcal{L}_{\text{sm}} + w_{\text{id}}\mathcal{L}_{\text{id}}. \tag{7}$$

Here $\mathcal{L}_{\text{NLL}}$ supervises the verbal response, including any interval to be located, and $\mathcal{L}_{\text{traj}}$ penalizes position errors at valid target–time pairs. The smoothness term regularizes changes in consecutive predicted displacements; the identity term associates states of the same entity across turns. Structured XYZ fields receive geometric supervision, so the target state is trained through the coordinates recovered from its selected evidence.

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) subsequently aligns the answer policy using language and geometric rewards while fixing RTGE, TRK-specific parameters, and OSK-RA. Each generated response produces a target state and a decoded trajectory, whose scores contribute to the policy reward. The model adapts Qwen3.5-9B through low-rank adaptation (LoRA) (Hu et al., 2022); Appendix C defines the losses and training settings.

## 5 Experiments

### 5.1 Experimental Setup

**Data and inputs.** Track4D-Bench evaluates all eligible questions or reference steps in each evaluation set across its Short and Long regimes. Scene-level partitions keep overlapping source windows together, and question identifiers link the reported evaluation groups. Language and sparse-tracking comparisons use matched observations. Dialogue evaluation retains generated history and resets memory between clips. Section 3 defines the metrics; Appendix C.7 reports seed sensitivity.

**Comparisons.** Language baselines include general and 3D/4D LMMs, with supervised fine-tuning (SFT) controls trained on the same data. Component ablations use the standard objective and a controlled objective shared by the RTGE/TRK variants (Appendix C.5). Tracking baselines answer trajectory questions through calibrated geometric estimation and Qwen3.5-9B target binding. Appendix E specifies their sparse-input setting and a complementary dense-frame KITTI comparison. External evaluations assess question answering and dynamic object grounding.

Table 1: Comparison on Track4D-Bench. C denotes CIDEr; B-4 and M denote BLEU-4 and METEOR scaled by 100. Geometric scores are percentages over all eligible Short and Long requests. Dashes denote unreported scores. Bold and underline mark the best and second-best language scores.

| Method | C ↑ | B-4 ↑ | M ↑ | SAcc @0.5 ↑ | Traj-Acc @1.0 ↑ | TAcc @0.5 ↑ |
|---|---|---|---|---|---|---|
| *Proprietary LMMs* | | | | | | |
| GPT-4o (Hurst et al., 2024) | 0.246 | 13.2 | - | - | - | - |
| Gemini 2.5 Pro (Comanici et al., 2025) | 0.155 | 6.2 | - | - | - | - |
| *Open-source LMMs* | | | | | | |
| Qwen3-VL-8B-Instruct (Bai et al., 2025a) | 0.314 | 16.9 | - | - | - | - |
| LLaVA-Video (Zhang et al., 2024b) | 0.059 | 5.7 | 13.1 | - | - | - |
| Video-3D LLM (Zheng et al., 2025) | 0.000 | 0.0 | 3.0 | - | - | - |
| LLaVA-3D (Zhu et al., 2025) | 0.000 | 0.3 | 1.5 | - | - | - |
| MLLM-4D (Yin et al., 2026) | 0.133 | 9.4 | 16.4 | - | - | - |
| *Same-dataset SFT controls* | | | | | | |
| Video-3D LLM + SFT (Zheng et al., 2025) | 0.000 | 0.5 | 4.1 | - | - | - |
| LLaVA-Video + SFT (Zhang et al., 2024b) | 0.115 | 8.4 | 22.6 | - | - | - |
| Qwen2.5-VL-7B + SFT (Bai et al., 2025b) | 0.035 | 0.0 | 10.8 | - | - | - |
| Qwen3.5-9B + SFT (Qwen Team, 2026) | <u>0.451</u> | <u>22.0</u> | <u>34.8</u> | - | - | - |
| LMM-Track4D (Ours) | **0.698** | **32.2** | **45.3** | 75.8 | 62.6 | 61.9 |

### 5.2 Comparative Experiments

**Quantitative Results.** LMM-Track4D achieves the highest language scores in Table 1, improving CIDEr by 0.247, BLEU-4 by 10.2 points, and METEOR by 10.5 points over Qwen3.5-9B + SFT. This comparison measures the combined contribution of the target representation and training objectives. The model also produces metric trajectories with 62.6% Traj-Acc and 61.9% TAcc.

**Comparison with Tracking Models.** On trajectory questions with matched sparse RGB inputs, LMM-Track4D reaches 78.8% Traj-Acc and 77.8% TAcc, exceeding the strongest tracking baseline by 73.00 and 76.77 percentage points. Appendix E reports average displacement error (ADE) and the complementary dense-frame KITTI comparison.

### 5.3 Ablation Studies

**Effect of Geometric Encoding and Decoding.** Removing RTGE lowers TAcc from 61.9% to 48.2% in Table 2, with calibrated rays retained in the decoder. Replacing OSK-RA with direct coordinate regression lowers TAcc to 42.5%. Both comparisons support using geometry in the target state and position decoding under the standard objective.

Table 2: Component ablations. C: CIDEr; other scores in %.

| Variant | C ↑ | SAcc ↑ | Traj-Acc↑ | TAcc ↑ |
|---|---|---|---|---|
| Full model | **0.698** | **75.8** | **62.6** | **61.9** |
| w/o OSK-RA | <u>0.631</u> | 58.8 | 50.0 | 42.5 |
| w/o RTGE | 0.579 | <u>71.4</u> | <u>57.1</u> | <u>48.2</u> |
| w/o TRK | 0.509 | – | – | – |

**Effect of Cross-Turn Target Memory.** Removing TRK while retaining generated text history reduces CIDEr from 0.698 to 0.509 (Table 2). TRK retains entity context beyond the verbal history and uses that context to select geometric evidence for subsequent questions. Table 3 shares the decoder, data, and training across variants, with identity and smoothness losses disabled. Adding TRK to RTGE increases SAcc and Traj-Acc by 0.55 points each and TAcc by 0.72 points. TRK also improves all three metrics without RTGE. Together, RTGE and TRK improve TAcc by 11.21 points over the decoder alone. These gains support using persistent target context for both verbal answers and decoded positions.

Table 3: Controlled RTGE/TRK ablations. All scores in %.

| RTGE | TRK | SAcc ↑ | Traj-Acc↑ | TAcc ↑ |
|---|---|---|---|---|
| Off | Off | 69.78 | 55.49 | 48.99 |
| On | Off | 74.18 | 60.99 | 59.48 |
| Off | On | 70.33 | 56.04 | 49.52 |
| On | On | **74.73** | **61.54** | **60.20** |

### 5.4 Quantitative Analysis

Table 4: Track4D-Bench results grouped by semantic question type. BLEU-4 and METEOR are scaled by 100; CIDEr retains its original scale. Geometric scores are percentages. Scene-level questions have no object-centric geometric target (†).

| Type | C ↑ | B-4 ↑ | M ↑ | SAcc @0.5 ↑ | Traj-Acc @1.0 ↑ | TAcc @0.5 ↑ |
|---|---|---|---|---|---|---|
| Identity | **1.274** | 31.6 | **53.5** | 70.0 | 70.0 | 63.6 |
| Trajectory | 0.678 | **33.5** | 46.1 | 87.9 | 78.8 | 77.8 |
| Relation | 0.539 | 22.1 | 44.6 | **100.0** | **81.1** | **89.9** |
| Motion trend | 0.420 | 15.4 | 37.8 | 78.8 | 51.5 | 44.4 |
| Scene-level | 0.369 | 20.2 | 41.6 | –† | –† | –† |

**Results by Question Type.** Table 4 groups questions by their requested semantic operation. Identity questions ask which entity is being described and yield the highest CIDEr. Trajectory and relation questions instead emphasize motion over an interval and relationships between entities, reaching 78.8% and 81.1% Traj-Acc. Their language scores reflect different answer content, while their geometric scores use the same distance thresholds. Motion-trend questions concern changes in motion and reach 44.4% TAcc. Appendix C.7 reports the model's sensitivity to the training seed.

Table 5: Reference-state continuations under three velocity scales. Early, Middle, and Late denote turns 2, 3, and 4 onward. Each target grid contributes up to four uniformly sampled positions. All entries are percentages.

| | Traj-Acc @1.0 ↑ | | | TAcc @0.5 ↑ | | |
|---|---|---|---|---|---|---|
| Setting | Early | Middle | Late | Early | Middle | Late |
| Low velocity scale | 67.4 | 53.5 | 66.0 | 59.3 | 56.6 | 47.2 |
| Medium velocity scale | 74.4 | 65.1 | 71.7 | 65.7 | 68.1 | 59.3 |
| High velocity scale | 88.4 | 81.4 | 73.6 | 72.1 | 78.3 | 65.0 |

**Effect of Reference Velocity.** Table 5 evaluates trajectory continuations obtained by extrapolating velocity from two preceding reference positions. Larger tested velocity scales improve both metrics in every dialogue-turn group. In late turns, changing the scale from Low to High raises Traj-Acc from 66.0% to 73.6% and TAcc from 47.2% to 65.0%. These reference-state continuations characterize the effect of supplied motion on the target trajectories; the learned-memory comparison is given in Table 3. Appendix D defines the velocity scales and extends the diagnostic to future-only motion groups.

### 5.5 External Benchmark Evaluation

Table 6: Dyn-Bench comparisons by interaction category. Question answering reports accuracy; dynamic object grounding reports $\mathcal{J}\&\mathcal{F}$. All entries are percentages.

| Method | Inter-Object ↑ | Object-Scene ↑ | Camera-Object ↑ |
|---|---|---|---|
| *Question answering (Accuracy)* | | | |
| Qwen3-VL-8B (Bai et al., 2025a) | 59.00 | 76.20 | 56.00 |
| DynTrace (Gao et al., 2026) | 66.33 | 80.53 | 58.63 |
| Qwen3.5-9B (Qwen Team, 2026) | 59.20 | 79.35 | 58.87 |
| LMM-Track4D | **67.60** | **80.80** | **60.93** |
| *Dynamic object grounding ($\mathcal{J}\&\mathcal{F}$)* | | | |
| UniPixel-7B (Liu et al., 2026) | 66.00 | 71.10 | 58.60 |
| Sa2VA-Qwen2.5-VL-7B (Yuan et al., 2026) | 73.40 | 75.90 | 69.10 |
| Sa2VA-InternVL2.5-8B (Yuan et al., 2026) | 76.80 | 80.20 | 70.10 |
| LMM-Track4D | **78.60** | **80.80** | **74.93** |

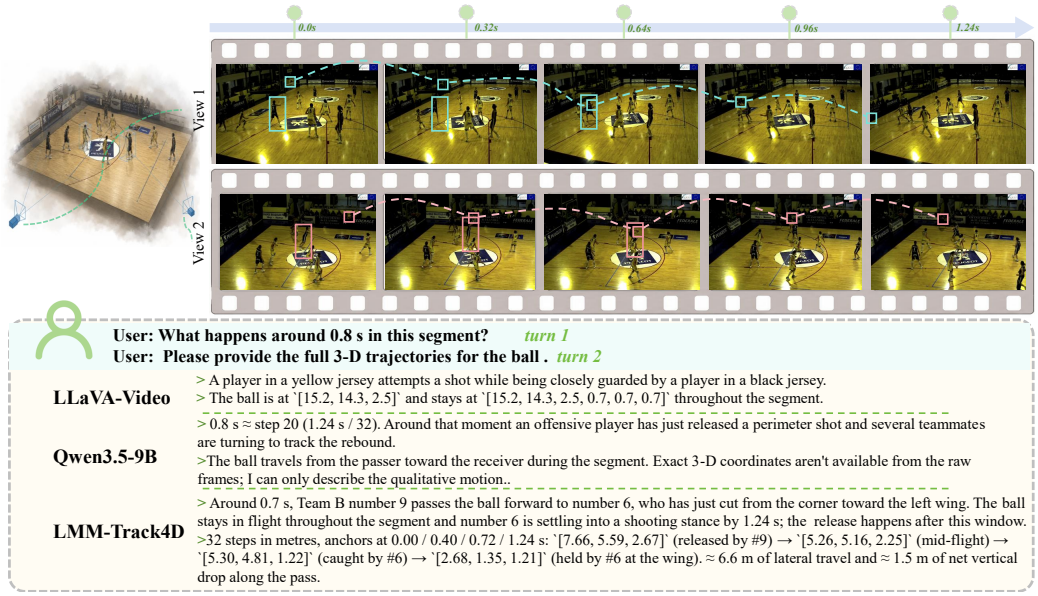



Figure 4: Basketball dialogue with LLaVA-Video, Qwen3.5-9B, and LMM-Track4D responses. An event question is followed by a request for the ball trajectory over the clip interval. The filmstrip illustrates how the scene changes over the queried interval.

**Comparison on Dyn-Bench.** Dyn-Bench (Huang et al., 2026b) evaluates interactions between objects, between objects and the scene, and between cameras and objects. In Table 6, LMM-Track4D improves question-answering accuracy over the strongest listed model in each category by 1.27, 0.27, and 2.06 percentage points, respectively. For dynamic object grounding, its predicted target reference conditions a frozen video-segmentation interface, yielding gains of 1.80, 0.60, and 4.83 points. These grounding scores evaluate the complete language-conditioned system. Appendix F describes the official evaluation protocol and model use.

**Evaluation on VLM4D.** The jointly supervised model reaches 58.15% accuracy on VLM4D without benchmark-specific training. It answers these video questions through its frozen RGB representation, extending evaluation beyond the calibrated Track4D-Bench setting. Appendix F specifies the checkpoint and reports the real and synthetic video groups, alongside a separate Dyn-Bench prompt analysis.

### 5.6 Qualitative Comparison

**Qualitative Results.** Figure 4 compares LMM-Track4D with LLaVA-Video and Qwen3.5-9B in a two-turn basketball dialogue. LLaVA-Video returns fixed positions over the queried interval, and Qwen3.5-9B describes the motion verbally. LMM-Track4D describes the pass and supplies timestamped metric positions of the ball. This comparison illustrates the output capability targeted by trajectory-grounded dialogue: an event question can be followed by a request for explicit positions of the referenced entity over time.

## 6 Conclusion

Trajectory-grounded dialogue makes an entity's changing metric positions an explicit object of conversation. Track4D-Bench associates dialogue references with timestamped trajectories, allowing verbal answers and geometric outputs to be evaluated within the same temporal context. LMM-Track4D learns a geometry-conditioned target state that retains entity context across turns and selects observations for decoding the requested positions. Its language and trajectory gains, together with component analyses and external dynamic-scene evaluations, support this connection between persistent entity references and metric reasoning in LMMs.

## AI USE STATEMENT

Generative AI tools assisted with synthetic data generation and language polishing. The authors take responsibility for the final data and manuscript.

## REFERENCES


Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. URL https://arxiv.org/abs/2206.14651.

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. *arXiv preprint arXiv:2502.13923*, 2025b.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE International Conference on Image Processing*, 2016. URL https://arxiv.org/abs/1602.00763.

Federico Castanedo, Jesus Garcia, Miguel A Patricio, and José M Molina. A multi-agent architecture based on the bdi model for data fusion in visual sensor networks. *Journal of Intelligent & Robotic Systems*, 62(3):299–328, 2011.

Tatjana Chavdarova, Pierre Baqué, Stéphane Bouquet, Andrii Maksai, Cijo Jose, Timur Bagautdinov, Louis Lettry, Pascal Fua, Luc Van Gool, and François Fleuret. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5030–5039, 2018.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465. IEEE, 2024. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.01370.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jiajun Deng, Tianyu He, Li Jiang, Tianyu Wang, Feras Dayoub, and Ian Reid. 3D-LLaVA: Towards generalist 3D LMMs with omni superpoint transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3772–3782, 2025. doi: 10.1109/CVPR52734.2025.00357.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2694–2703. IEEE, 2023. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.00254.

Mingkang Dong, Muxin Pu, Jie Li, Bohan Guo, Songruo Chen, Bin Ren, Xu Zheng, Chen Zhao, Tianwen Qian, Mohamed Elhoseiny, and Yuqian Fu. ObjectStream: Latent objects as memory anchors for streaming video understanding. *arXiv preprint arXiv:2607.28312*, 2026. URL https://arxiv.org/abs/2607.28312.

Bishoy Galoaa, Shayda Moezzi, Xiangyu Bai, and Sarah Ostadabbas. Motion-o: Trajectory-grounded video reasoning. *arXiv preprint arXiv:2603.18856*, 2026. URL https://arxiv.org/abs/2603.18856.

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

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Amol Harsh, Zongyan Han, Jean Lahoud, Ye Liu, Rao Muhammad Anwer, Hisham Cholakkal, Salman Khan, and Fahad Khan. Ground3D-LMM: Fine-grained 3d point grounding and spatial reasoning with lmm. *arXiv preprint arXiv:2607.05493*, 2026. URL https://arxiv.org/abs/2607.05493.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.

Jiangyong Huang, Xiaojian Ma, Xiongkun Linghu, Junchao He, Qing Li, Song-Chun Zhu, Yixin Chen, Baoxiong Jia, and Siyuan Huang. Leo-vl: Efficient scene representation for scalable 3d vision-language learning. *arXiv preprint arXiv:2506.09935*, 2025a. URL https://arxiv.org/abs/2506.09935.

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

Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv preprint arXiv:2507.23478*, 2025b.

Yuzhi Huang, Kairun Wen, Rongxin Gao, Dongxuan Liu, Yibin Lou, Jie Wu, Jing Xu, Jian Zhang, Zheng Yang, Yunlong Lin, et al. Thinking in dynamics: How multimodal large language models perceive, track, and reason dynamics in physical 4d world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Huang_Thinking_in_Dynamics_How_Multimodal_Large_Language_Models_Perceive_Track_CVPR_2026_paper.pdf.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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. *arXiv preprint arXiv:2506.03135*, 2025.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. Videomem: Enhancing ultra-long video understanding via adaptive memory management. *arXiv preprint arXiv:2512.04540*, 2025a.

Hongbo Jin, Jiayu Ding, Siyi Xie, Guibo Luo, and Ge Li. Vista: Mitigating semantic inertia in video-llms via training-free dynamic chain-of-thought routing. *arXiv preprint arXiv:2505.11830*, 2026a.

Hongbo Jin, Siyi Xie, Jiayu Ding, Kuanwei Lin, and Ge Li. Tir-flow: Active video search and reasoning with frozen vlms. *arXiv preprint arXiv:2601.06176*, 2026b.

Hongbo Jin, Rongpeng Zhu, Jiayu Ding, Wenhao Zhang, and Ge Li. Himac: Hierarchical macro-micro learning for long-horizon llm agents. *arXiv preprint arXiv:2603.00977*, 2026c.

Hongbo Jin, Rongpeng Zhu, Zhongjing Du, Xu Jiang, Jingqi Tian, Qiaoman Zhang, and Jiayu Ding. Dgpo: Distribution guided policy optimization for fine grained credit assignment. *arXiv preprint arXiv:2605.03327*, 2026d.

Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Hołynski. Stereo4D: Learning how things move in 3D from internet stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10497–10509, 2025b. doi: 10.1109/CVPR52734.2025.00982.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. DynamicStereo: Consistent dynamic depth from stereo videos. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13229–13239. IEEE, 2023. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01271.

Junlong Ke, Zichen Wen, Boxue Yang, Yantai Yang, Xuyang Liu, Chenfei Liao, Zhaorun Chen, Shaobo Wang, and Linfeng Zhang. Flash-unified: A training-free and task-aware acceleration framework for native unified models. *arXiv preprint arXiv:2603.15271*, 2026.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8592–8603, 2025a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat : Chat-centric video understanding. *Science China Information Sciences*, 68 (10):200102, 2025b. doi: 10.1007/s11432-024-4321-9.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.342.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Chen. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. *Advances in Neural Information Processing Systems*, 38:126078–126108, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Torben Teepe, Philipp Wolters, Johannes Gilg, Fabian Herzog, and Gerhard Rigoll. Lifting multi-view detection and tracking to the bird's eye view. *arXiv preprint arXiv:2403.12573*, 2024. URL `https://arxiv.org/abs/2403.12573`.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

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

Xiyang Wang, Shouzheng Qi, Jieyou Zhao, Hangning Zhou, Siyu Zhang, Guoan Wang, Kai Tu, Songlin Guo, Jianbo Zhao, Jian Li, and Mu Yang. Mctrack: A unified 3d multi-object tracking framework for autonomous driving. *arXiv preprint arXiv:2409.16149*, 2024. URL https://arxiv.org/abs/2409.16149.

Yiyu Wang, Xuyang Liu, Xiyan Gui, Xinying Lin, Boxue Yang, Chenfei Liao, Tailai Chen, and Linfeng Zhang. Accelerating streaming video large language models via hierarchical token compression. *arXiv preprint arXiv:2512.00891*, 2025c.

Yuxin Wang, Lei Ke, Boqiang Zhang, Tianyuan Qu, Hanxun Yu, Zhenpeng Huang, Meng Yu, Dan Xu, and Dong Yu. N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. *arXiv preprint arXiv:2512.16561*, 2025d. URL https://arxiv.org/abs/2512.16561.

Hongkai Wei, Yang Yang, Shijie Sun, Mingtao Feng, Xiangyu Song, Qi Lei, Hongli Hu, Rong Wang, Huansheng Song, Naveed Akhtar, and Ajmal Saeed Mian. Mono3DVLT: Monocular-video-based 3d visual language tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13886–13896, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wei_Mono3DVLT_Monocular-Video-Based_3D_Visual_Language_Tracking_CVPR_2025_paper.html.

Zichen Wen, Yiyu Wang, Chenfei Liao, Boxue Yang, Junxian Li, Weifeng Liu, Haocong He, Bolong Feng, Xuyang Liu, Yuanhuiyi Lyu, et al. Ai for service: Proactive assistance with ai glasses. *arXiv preprint arXiv:2510.14359*, 2025.

Zichen Wen, Boxue Yang, Shuang Chen, Yaojie Zhang, Yuhang Han, Junlong Ke, Cong Wang, Yicheng Fu, Jiawang Zhao, Jiangchao Yao, et al. Innovator-vl: A multimodal large language model for scientific discovery. *arXiv preprint arXiv:2601.19325*, 2026a.

Zichen Wen, Boxue Yang, Junlong Ke, Jiajie Huang, Chenfei Liao, Junxi Wang, Xuyang Liu, and Linfeng Zhang. Evostreaming: Your offline video model is a natively streaming assistant. *arXiv preprint arXiv:2605.10343*, 2026b.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *Advances in neural information processing systems*, 38: 13569–13597, 2026.

Xiyang Wu, Zongxia Li, Jihui Jin, Guangyao Shi, Gouthaman KV, Vishnu Raj, Nilotpal Sinha, Jingxi Chen, Fan Du, and Dinesh Manocha. MASS: Motion-aware spatial-temporal grounding for physics reasoning and comprehension in vision-language models. *arXiv preprint arXiv:2511.18373*, 2025. URL https://arxiv.org/abs/2511.18373.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15073, pp. 98–115. Springer Nature Switzerland, 2025. ISBN 978-3-031-72632-3 978-3-031-72633-0. doi: 10.1007/978-3-031-72633-0_6.

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.

Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_Inst3D-LMM_Instance-Aware_3D_Scene_Understanding_with_Multi-modal_Instruction_Tuning_CVPR_2025_paper.html.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2VA: Marrying SAM2 with MLLM for dense grounded understanding of images and videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3720379. URL https://arxiv.org/abs/2501.04001.

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.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. *arXiv preprint arXiv:2110.06864*, 2021. URL https://arxiv.org/abs/2110.06864.

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 2024a. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning position-aware video representation for 3D scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, 2025. doi: 10.1109/CVPR52734.2025.00841.

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.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19798–19808. IEEE, 2023. ISBN 979-8-3503-0718-4. doi: 10.1109/ICCV51070.2023.01818.

Hanyu Zhou and Gim Hee Lee. Uni4d-llm: A unified spatiotemporal-aware vlm for 4d understanding and generation. *arXiv preprint arXiv:2509.23828*, 2025a.

Hanyu Zhou and Gim Hee Lee. Llava-4d: Embedding spatiotemporal prompt into lmms for 4d scene understanding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/312c92e0cb862422eaa49452cdf55caf-Abstract-Conference.html.

Haoran Zhou and Gim Hee Lee. Motion4D: Learning 3D-consistent motion and semantics for 4D scene understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-3869. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/a84953147312ea2e8b020e53a267321b-Abstract-Conference.html.

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*, pp. 8600–8612, 2025.

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. doi: 10.1109/ICCV51701.2025.00409.

# Supplementary Material

## A Additional Benchmark Statistics

Table 7 reports how the three source datasets contribute clips, dialogue turns, frames, and entity tracks to Track4D-Bench. An entity track counts once within a clip, independently of its number of timestamps, views, or dialogue references. The same source entity can contribute a track to more than one clip. Frame counts sum camera–timestamp image records across clips.

Table 7: Statistics of Track4D-Bench. Entity tracks count distinct entities within each clip and are summed over clips. Frame counts likewise sum camera–timestamp image records within clips.

| Source dataset | Clips | QA turns | Frames | Entity tracks |
|---|---|---|---|---|
| APIDIS | 77 | 336 | 3,904 | 1,154 |
| KITTI | 383 | 1,585 | 18,688 | 4,205 |
| WildTrack | 66 | 274 | 912 | 2,165 |
| Total | 526 | 2,195 | 23,504 | 7,524 |

## B Additional Benchmark Construction Details

Track4D-Bench uses source identities and timestamps to associate each dialogue reference with its geometric targets. The construction below describes how these associations are preserved through state preparation, clip formation, and dialogue generation.

### B.1 State Preparation

Each source dataset is represented by scene-indexed entities with timestamped positions and associated observations. An entity record links the source, sequence, and track identity to its XYZ coordinates, time in seconds, reference-validity mask, and visible views. Times are measured relative to the clip start while preserving the physical intervals between observations. Camera intrinsics and extrinsics associate the images with a fixed coordinate system for the clip.

APIDIS and WildTrack person tracks use ground locations. WildTrack positions are embedded as $(X, Y, 0)$, with the third coordinate denoting the ground plane. APIDIS ball tracks use ball centers reconstructed from synchronized multiview ball-center annotations and camera calibration, retaining the ball's height above the court. KITTI tracks use the bottom centers of the source 3D boxes. Camera-coordinate positions are transformed into the fixed clip frame with the corresponding camera poses:

$$\mathbf{p}_{\text{clip}}(t) = R_{\text{cam}\to\text{clip}}(t)\mathbf{p}_{\text{cam}}(t) + \mathbf{t}_{\text{cam}\to\text{clip}}(t). \tag{8}$$

The rotation and translation account for camera motion, and all positions are expressed in meters. Each entity retains its physical reference-point definition throughout the clip.

The source records also contain box attributes used to obtain these reference points. APIDIS person records use $(c_x, c_y, c_z, d_x, d_y, d_z)$ in centimeters, where $\mathbf{c}$ is the box center and $\mathbf{d}$ gives its extents. WildTrack person records use the same fields in meters with an additional yaw angle. The person-box centers are converted to ground locations using the vertical extent and calibrated ground plane; WildTrack stores the resulting location as $(X, Y, 0)$. KITTI records use $(h, w, l, x, y, z, \psi)$, with dimensions followed by the box-bottom position and yaw $\psi$. Their position fields undergo Eq. 8, whereas APIDIS ball centers retain their height above the court. The qualitative figures display the original source-record formatting; geometric evaluation uses the converted metric positions.

The reference-validity mask records whether each position is available, independently of visibility in a particular image. A position with a valid reference therefore remains eligible during occlusion. The valid positions preserve the source's reference-point definition, providing planar supervision for ground locations and height information for ball centers.

Each dialogue record associates these geometric targets with a clip, turn order, question, verbal answer, queried entities, requested times, semantic type, and eligible output fields. Source identities establish

the correspondence during construction and scoring. The model instead resolves the requested entities from visible questions and observations, together with its generated history. Source-specific cues, including ego motion, event annotations, and synchronized views when available, remain associated with the clip.

### B.2 Clip Formation

Each clip defines the visual and temporal context for a multi-turn dialogue. Short clips pair full-clip trajectory requests with qualitative questions about the same event. Long clips provide broader context for requests about local positions or trajectory segments. For example, a Short APIDIS dialogue requests the ball trajectory during a 1.24-second pass. A Long dialogue supplies 5.08 seconds of context, asks who possesses the ball at 0.4 seconds, and requests player and ball trajectories over 0.4–1.2 seconds. These intervals determine the requested outputs; Appendix E describes the sparse observations used in the Short tracking comparisons.

Source sequences or scenes are assigned to train, validation, and test partitions before overlapping windows are formed. Windows from the same sequence or scene remain within one partition, together with all synchronized cameras observing that scene. Dialogue turns retain their original order within each clip.

### B.3 Dialogue Construction

Dialogue generation uses the prepared state records for geometric targets and the associated images for scene content. The generated questions concern identity, trajectory, relation, motion trend, and scene context. Later questions can change the queried property or time interval while retaining an earlier entity reference. The entity's source identity links that reference to the requested trajectory. Geometric answers inherit identities, timestamps, and coordinates from the state records, while language generation expresses the corresponding questions and verbal answers.

### B.4 Quality Control

Checks against the state records validate entity references, requested intervals, coordinate entries, and required answer fields. A large language model (LLM)/vision-language model (VLM) judge then filters dialogues for consistency between verbal descriptions, visual evidence, and geometric targets. The authors spot-check visual grounding, cross-turn references, and event descriptions and handle inconsistencies found in these checks. The retained release contains 428 Short clips and 98 Long clips. Manual checking was performed by the authors without external crowdsourcing.

## C Additional Evaluation, Method, and Experimental Details

This appendix details the evaluation protocol, target-state management, geometric decoding, and training objectives.

### C.1 Evaluation Protocol Details

The evaluation sets $\mathcal{T}_{\mathrm{lang}}$, $\mathcal{T}_{\mathrm{pt}}$, and $\mathcal{T}_{\mathrm{traj}}$ contain turns eligible for language, point, and trajectory scoring, respectively. Eligibility depends on the requested output and available reference, as in Section 3. A natural-language reference permits language scoring; a point or trajectory request with a valid reference permits the corresponding geometric scoring. These conditions are independent of the five semantic question types. For language scoring, structured coordinate lists are removed from both predictions and references, retaining verbal identity, action, time, and relation content. CIDEr, BLEU-4, and METEOR are computed directly on each table's fixed language set, including for aggregate results. BLEU-4 and METEOR are multiplied by 100 for reporting; training rewards use the original metric values.

For $i \in \mathcal{T}_{\mathrm{pt}}$, let $\mathcal{P}_i$ contain the requested targets with valid point references, and let $\hat{\mathbf{p}}_{i,u}, \mathbf{p}_{i,u} \in \mathbb{R}^3$ be the predicted and reference positions of target $u$. Define

$$e_i = \frac{1}{|\mathcal{P}_i|} \sum_{u \in \mathcal{P}_i} \|\hat{\mathbf{p}}_{i,u} - \mathbf{p}_{i,u}\|_2, \qquad \mathrm{SAcc@0.5} = \frac{1}{|\mathcal{T}_{\mathrm{pt}}|} \sum_{i \in \mathcal{T}_{\mathrm{pt}}} \mathbf{1}[e_i < 0.5].$$

For $j \in \mathcal{T}_{\mathrm{traj}}$, let $\mathcal{V}_j$ contain the valid reference target–time pairs after removing reference positions marked as missing. The index $t$ below enumerates these pairs in query-target and timestamp order; it can cover multiple targets in one question. Denote the predicted and reference positions for a pair by $\hat{\mathbf{p}}_{j,t}$ and $\mathbf{p}_{j,t}$. The pointwise and per-question errors are

$$e_{j,t} = \|\hat{\mathbf{p}}_{j,t} - \mathbf{p}_{j,t}\|_2, \qquad \bar{e}_j = \frac{1}{|\mathcal{V}_j|} \sum_{t \in \mathcal{V}_j} e_{j,t}.$$

The corresponding trajectory and step accuracies are

$$\mathrm{Traj\text{-}Acc@1.0} = \frac{1}{|\mathcal{T}_{\mathrm{traj}}|} \sum_{j \in \mathcal{T}_{\mathrm{traj}}} \mathbf{1}[\bar{e}_j < 1.0],$$

$$\mathrm{TAcc@0.5} = \frac{\sum_{j \in \mathcal{T}_{\mathrm{traj}}} \sum_{t \in \mathcal{V}_j} \mathbf{1}[e_{j,t} < 0.5]}{\sum_{j \in \mathcal{T}_{\mathrm{traj}}} |\mathcal{V}_j|}.$$

Here SAcc@0.5 measures success per point-grounding turn, and TAcc@0.5 measures accuracy per reference target–time pair. Multiple targets follow the order specified by the query, and their positions are aligned to the declared output timestamps. Missing, duplicate, or non-finite predictions that cannot be uniquely aligned have infinite error at the affected required positions. An incorrectly located interval leaves its missing required reference timestamps in the denominator. Reference missingness is masked before scoring; prediction failure does not change that mask. Error uses the $(x, y, z)$ coordinates in meters, with strict thresholds as written above. Thresholded geometric metrics are reported as corpus-level micro averages,

$$\mathrm{MicroAvg} = \frac{\sum_k s_k}{\sum_k n_k},$$

where $k$ indexes clips, $n_k$ counts reference evaluation units in clip $k$, and $s_k$ counts successes. The unit is a point-grounding turn for SAcc@0.5, a trajectory turn for Traj-Acc@1.0, and a valid reference target–time pair for TAcc@0.5. Missing or non-finite outputs contribute zero to $s_k$.

### C.2 Reference Binding and Target Memory

The reference table $\mathcal{R}_{k-1}$ associates each clip-local entity key with descriptions and aliases from visible questions or generated answers, together with its retained target state. The language backbone uses this table, the conversation, and visual features to resolve references in the current question. Extending Eq. 2 to multiple targets gives an ordered list of local keys:

$$\mathbf{e}_k = \mathcal{B}_\theta(q_k, \mathcal{H}_{k-1}, \mathcal{F}, \mathcal{R}_{k-1}). \tag{9}$$

Reference binding uses constrained generation by the same backbone that answers the question. In query order, the generated output selects an existing key, introduces a new target description, or leaves the reference unresolved. Descriptions interpreted as the same entity reuse a key, including pronouns whose antecedents have been resolved. Ambiguous references remain unresolved. A new entity receives a fresh key and the learned initial target state.

For a resolved key $e$, let $\rho_k^{(e)}$ be its description in the current query, resolved using the reference table when necessary. This description distinguishes newly introduced targets that share the learned initial state. The model reads the latest state and includes the description at the corresponding TRK prompt position. It generates the verbal response $a_k$, including any interval that the question asks it to locate, before processing the assistant-side TRK write positions. The hidden state at the write position implements the update defined by Eq. 3. The notation $\mathbf{s}_{k-1}^{(e)}$ denotes the latest state for key $e$ even when the preceding turn concerns another entity. Target descriptions and their read/write positions

follow query order. Each updated state conditions its target's OSK-RA output, whose coordinates are emitted as geometric answer fields, and is retained under the same key.

Target states are stored in entity-specific memory slots. Each slot retains up to eight states, including the current update; earlier states are cached without becoming additional input tokens. When the queried entity changes, the other entities retain their entries. With 32 occupied slots, a new entity evicts the least recently queried entity outside the current query. A later reference to an evicted entity can retain its local key but restarts from the learned initial state. Both the reference table and target memory reset between clips. A scene-level question with no individual target leaves target memory unchanged.

An unresolved reference produces no geometric estimate for that target. Its valid reference positions remain in the evaluation denominator and count as failures. Source track identities define positive pairs for training the identity objective, while inference selects local keys from the question, preceding generated conversation, and observations. Generated answers are retained as produced, including failed answers; reference answers do not replace them between turns.

### C.3 RTGE and Geometric Anchor Details

RTGE uses two-layer projections for ray and time embeddings. The time projection maps 32-D Fourier features, corresponding to $m = 16$ sine–cosine frequency pairs, through a multi-layer perceptron (MLP) with hidden width 128. The resulting visual features retain their calibrated rays for OSK-RA, which expresses predicted positions in meters. The number of positions follows the requested interval on the source time grid. This grid uses time metadata independently of reference positions and their validity mask. The decoder's fixed output-channel capacity covers the longest local request in Track4D-Bench. Its scoring channels share the observation set; a request of length $H$ activates the first $H$ channels and masks the rest.

A view observation is an image from camera $v$ at timestamp $\tau$, so images from the same camera at different times are distinct observations. For each output step, the highest-scoring valid patch ray represents its image, using $z_{k,h,r}$; a fixed patch order breaks exact ties. Images without a valid ray contribute no candidate. The representatives form $\mathcal{C}_{k,h}$, and softmax over their scores defines observation confidence:

$$\pi_{k,h,r} = \frac{\exp z_{k,h,r}}{\sum_{j \in \mathcal{C}_{k,h}} \exp z_{k,h,j}}.$$

The retained set $\mathcal{S}_{k,h}$ consists of the top seven candidates filtered by a confidence threshold of 0.2. Since their probabilities sum to one, at most five candidates satisfy this threshold. If none does, the highest-scoring valid candidate is retained. The selected confidences are then normalized as $\alpha_{k,h,r} = \pi_{k,h,r} / \sum_{j \in \mathcal{S}_{k,h}} \pi_{k,h,j}$, giving weights that sum to one for both feature fusion and anchor estimation.

The weighted line-intersection objective in Eq. 5 can be expressed as a stacked least-squares system. With $P_r = I - \mathbf{d}_r \mathbf{d}_r^\top$, define

$$B_{k,h} = \mathrm{stack}_{r \in \mathcal{S}_{k,h}}(\sqrt{\alpha_{k,h,r}} P_r), \qquad \mathbf{c}_{k,h} = \mathrm{stack}_{r \in \mathcal{S}_{k,h}}(\sqrt{\alpha_{k,h,r}} P_r \mathbf{o}_r). \tag{10}$$

With $B_{k,h} \in \mathbb{R}^{3|\mathcal{S}_{k,h}| \times 3}$ and $\mathbf{c}_{k,h} \in \mathbb{R}^{3|\mathcal{S}_{k,h}|}$, the minimum-norm solution is

$$\hat{\mathbf{p}}_{k,h}^{\mathrm{anchor}} = B_{k,h}^{\dagger} \mathbf{c}_{k,h}, \tag{11}$$

where † denotes the Moore–Penrose pseudoinverse.

For the retained rays, the system in Eq. 10 satisfies

$$\|B_{k,h}\mathbf{p} - \mathbf{c}_{k,h}\|_2^2 = \sum_{r \in \mathcal{S}_{k,h}} \alpha_{k,h,r} \|P_r(\mathbf{p} - \mathbf{o}_r)\|_2^2. \tag{12}$$

In the singular value decomposition $B_{k,h} = U \Sigma V^\top$, the pseudoinverse is $B_{k,h}^{\dagger} = V \Sigma^{\dagger} U^\top$. The diagonal entries of $\Sigma^{\dagger}$ invert the nonzero singular values and leave zeros at zero. The full set of least-squares solutions is

$$\mathbf{p} = B_{k,h}^{\dagger} \mathbf{c}_{k,h} + (I - B_{k,h}^{\dagger} B_{k,h}) \mathbf{u}, \qquad \mathbf{u} \in \mathbb{R}^3. \tag{13}$$

Choosing $\mathbf{u} = \mathbf{0}$ gives the minimum-norm anchor. Full column rank gives a unique solution. For one unit ray, the anchor reduces to $P_r\mathbf{o}_r$, the point on its supporting line nearest the world origin; depth along the line is left to the learned residual.

Within a fixed selected set, the target state affects the decoded position through the normalized observation weights. A single retained ray has weight one and therefore provides no gradient to the target state through these weights. If every output step retains the same single ray, the fused feature and anchor are identical, so the shared residual yields a constant trajectory. Variation across output positions depends on their scoring channels selecting or weighting evidence differently. The residual then corrects the resulting anchors using the fused visual and temporal features.

The pseudoinverse is computed by singular value decomposition, with numerically zero singular values excluded from inversion. Working directly with the stacked system avoids squaring its condition number through normal equations.

Valid candidates have finite camera centers and scores, with nonzero finite directions normalized to unit length. When no valid ray exists or the solver fails to return a finite result, the affected output is invalid and counts as a failure. Otherwise, the selected rays and their features provide the calibrated and temporal evidence used to estimate positions on the requested grid, including times between observations.

### C.4 Training Objective and Schedule Details

Training comprises geometry-oriented supervision (Stage 1), joint language and trajectory supervision (Stage 2), and GRPO policy alignment (Stage 3) (Shao et al., 2024). Stage 1 uses trajectory and smoothness supervision, with the identity objective applied to eligible cross-turn pairs and the text-likelihood term inactive. Stages 1 and 2 update the LoRA adapters (Hu et al., 2022) and the RTGE, TRK, and OSK-RA parameters involved in their objectives. Stage 3 updates LoRA while fixing those parameters. The remaining backbone weights stay frozen throughout.

The Stage 2 objective is given in Eq. 7. Each supervised term is defined on its eligible entries; a term with no eligible entries contributes zero. Here $\mathcal{L}_{\mathrm{NLL}}$ is the negative log-likelihood on the verbal assistant response, including any interval it is trained to locate. Structured XYZ fields are excluded from this token loss and supervised through the geometric decoder. Let $\Omega$ contain valid supervised target–time pairs, indexed by turn $k$ and output position $h$, with target indices suppressed. Let $\mathbf{p}^*_{k,h} \in \mathbb{R}^3$ denote the reference world-coordinate position. The trajectory loss is

$$\mathcal{L}_{\mathrm{traj}} = \frac{1}{|\Omega|} \sum_{(k,h)\in\Omega} \mathrm{SmoothL1}_{\beta=1.0}(\hat{\mathbf{p}}_{k,h}, \mathbf{p}^*_{k,h}),$$

where SmoothL1 averages the three coordinate losses and $\Omega$ weights valid target–time pairs uniformly. Let $\hat{\mathbf{v}}_{k,h} = \hat{\mathbf{p}}_{k,h+1} - \hat{\mathbf{p}}_{k,h}$ denote displacement between consecutive positions, and let $\Omega'$ contain indices with valid positions at $h-1$, $h$, and $h+1$. The smoothness regularizer penalizes changes in these displacements:

$$\mathcal{L}_{\mathrm{sm}} = \frac{1}{|\Omega'|} \sum_{(k,h)\in\Omega'} \|\hat{\mathbf{v}}_{k,h} - \hat{\mathbf{v}}_{k,h-1}\|_2^2 .$$

For identity supervision, let $\mathcal{P}$ contain same-entity TRK pairs across turns, with identity scoped by source and scene. Let $\mathcal{N}(u)$ contain the positive counterpart and sampled different-entity states for anchor $u$, excluding the anchor itself and padding. The function $\mathrm{sim}$ is cosine similarity between L2-normalized states, and $\tau_{\mathrm{id}} = 0.07$ is the contrastive temperature. Each anchor samples up to 32 negatives from a first-in-first-out (FIFO) queue of capacity 64. We use

$$\mathcal{L}_{\mathrm{id}} = -\frac{1}{|\mathcal{P}|} \sum_{(u,v)\in\mathcal{P}} \log \frac{\exp(\mathrm{sim}(\mathbf{s}_u, \mathbf{s}_v)/\tau_{\mathrm{id}})}{\sum_{w\in\mathcal{N}(u)} \exp(\mathrm{sim}(\mathbf{s}_u, \mathbf{s}_w)/\tau_{\mathrm{id}})}.$$

The reported run completed 1,438 geometry-oriented updates, 700 joint supervised updates, and 250 GRPO updates. GRPO uses a frozen reference policy and eight rollouts per prompt. Each rollout generates its own response and trailing TRK state, from which the fixed OSK-RA decoder produces its geometric output. Text and geometric scores form a scalar reward used in the group-relative advantage. This reward is treated as constant during policy differentiation, and the sampled text's

log-probabilities provide the gradient to LoRA. The geometry decoder is learned by supervised position losses in Stages 1 and 2; during Stage 3, its outputs contribute to the reward for the answer policy. TRK states continue to change with the response even though the TRK-specific parameters are fixed.

Cross-turn target states are detached with a truncated backpropagation through time (TBPTT) window of one turn; the contrastive queue supplies negatives for the supervised identity objective. Table 8 lists the optimization and decoding settings.

Table 8: Shared model design and optimization settings. Component ablations modify the named component; external evaluations freeze the specified checkpoint.

| Setting | Value |
|---|---|
| Backbone and adaptation | Qwen3.5-9B; LoRA rank 32, scaling 64, dropout 0.05; other backbone weights frozen |
| Context and target memory | Maximum sequence length 12,288; 32 target slots, eight states per slot; TBPTT window 1 with detached propagation |
| RTGE and OSK-RA | 32-D Fourier time features, MLP width 128; top-7 view observations before confidence threshold 0.2; query-dependent $H$ XYZ steps |
| Joint supervised optimization | 700 updates; learning rate $5 \times 10^{-6}$ with linear decay; 50 warmup updates; weight decay 0; per-worker batch/accumulation 1/1; bfloat16; gradient norm 1.0; seed 43 |
| Supervised loss weights | Text 1.0; trajectory 100.0; smoothness $10^{-5}$; identity 0.05; identity temperature 0.07 |
| GRPO optimization | LoRA only; 250 updates; learning rate $5 \times 10^{-7}$; 30 warmup updates; weight decay 0; gradient norm 0.5; seed 47; group size 8; KL coefficient 0.04 |
| GRPO sampling | Temperature 0.3; top-$p$ 0.9; top-$k$ 40; maximum 768 generated tokens |
| Reward weights | CIDEr/BLEU-4/METEOR: 0.4/0.3/0.3; format/Traj-Acc/TAcc/SAcc: 0.2/0.4/0.4/0.2 |
| Evaluation decoding | Greedy; repetition penalty 1.1; no-repeat 6-gram |

### C.5 Model and Comparison Settings

The LMM-Track4D component variants share the backbone and model settings, with the named computation changed in each ablation. The RTGE-free variant removes additive ray/time embeddings and retains calibrated rays in OSK-RA. The OSK-RA-free variant directly regresses coordinates from the target state with the same output order and requested-length mask. The TRK-free variant removes recurrent reads and writes while retaining generated text history. The standard-objective component variants in Table 2 start from the same pretrained backbone and follow all three training stages with applicable supervision. The backbone-only control receives 700 language SFT updates and is the Qwen3.5-9B + SFT row in Table 1.

The controlled variants in Table 3 share the decoder, training data, evaluation set, and 700 Stage-2 updates. Every variant disables identity and smoothness auxiliary losses, isolating RTGE and TRK under a common objective that differs from the standard objective. Table 8 gives the shared architecture and training settings.

Language baselines and sparse-tracking comparisons use the matched observations specified in Section 5.1, with each model retaining its native image preprocessing. Dialogue evaluation retains generated answers in turn order and applies the same verbal-content extraction. Geometric eligibility is determined from references and held fixed across methods.

### C.6 Inference and Requested Time Grids

Multi-turn inference retains target memory across turns and resets it at clip boundaries. Trajectory timestamps follow the source annotation and requested interval. For example, the displayed APIDIS trajectory contains 32 samples from 0 to 1.24 s, spaced 0.04 s apart; Long questions request a local segment within the longer clip. Each predicted position is associated with its requested timestamp before geometric scoring.

### C.7 SEED SENSITIVITY

Table 9: Sensitivity to the Stage-2 SFT seed with GRPO settings, including seed 47, fixed. Geometric scores are percentages. The last row gives the mean and sample standard deviation.

| Stage-2 seed | SAcc @0.5 ↑ | Traj-Acc @1.0 ↑ | TAcc @0.5 ↑ |
|---|---|---|---|
| 17 | 75.27 | 62.09 | 61.41 |
| 43 | 75.82 | 62.64 | 61.89 |
| 91 | 75.82 | 62.09 | 61.60 |
| Mean ± SD | $75.64 \pm 0.32$ | $62.27 \pm 0.32$ | $61.63 \pm 0.24$ |

Table 9 varies the Stage-2 seed while keeping GRPO settings fixed. The sample standard deviation is at most 0.32 percentage points for each geometric metric. The main full model uses Stage-2 seed 43.

### C.8 ASSET USE, RELEASE, AND BROADER IMPACTS

Track4D-Bench is derived from three public source datasets: APIDIS under the original academic research release, KITTI Tracking under CC BY-NC-SA 3.0, and WildTrack under its academic/non-commercial research release. We use these assets only for non-commercial research and retain citation and attribution to the original releases. The pretrained models and open-source baselines used in our experiments follow their original release terms, including Qwen3.5-9B and baselines such as LLaVA-Video, LLaVA-3D, Video-3D LLM, and MLLM-4D under their respective original model terms. Upon acceptance, we will release the benchmark-conversion scripts, training/evaluation code, model checkpoints, environment specification, and concise benchmark/model cards describing intended use, fields, evaluation protocol, and known limitations.

The benchmark concerns sports, driving, and pedestrian scenes, making it relevant to research on sports analytics, robot perception, and assistive scene understanding. Tracking across views also presents a risk of use in cross-camera surveillance. Our research uses publicly released datasets and plans a research-use release without face, gait, or biometric identification components. Manual verification consists of author spot checks; no external participants are recruited or compensated.

## D REFERENCE-STATE MOTION DIAGNOSTICS

### D.1 DIALOGUE-TURN CONTINUATIONS

For Table 5, two preceding reference positions $\mathbf{p}_{-1}, \mathbf{p}_0$ at distinct times $t_{-1} < t_0$ define

$$\mathbf{v}_0 = \frac{\mathbf{p}_0 - \mathbf{p}_{-1}}{t_0 - t_{-1}}, \qquad \hat{\mathbf{p}}(t_h) = \mathbf{p}_0 + \lambda_{m,g}\mathbf{v}_0(t_h - t_0), \quad t_h > t_0.$$

Here $\lambda_{m,g}$ is the velocity coefficient for setting $m$ and dialogue-turn group $g$. All settings use the same questions with two valid preceding states. Early, Middle, and Late denote turn 2, turn 3, and turns 4 onward, respectively; turn 1 is excluded. The Low, Medium, and High settings use coefficients $(0.30, 0.25, 0.30)$, $(0.60, 0.65, 0.76)$, and $(0.74, 0.85, 1.00)$ across these groups. Up to four positions are sampled uniformly in time from each target grid; shorter sequences use all valid positions. Physical timestamps determine velocity, while the turn groups locate the question within the conversation.

### D.2 Future-Only Motion Groups

Table 10: Future-only reference-state diagnostic. Repeat-current holds the last reference position fixed; Oracle-CV extrapolates reference velocity. All-source values macro-average the three sources, whereas main model accuracies are micro averages. Both methods use reference states; this table characterizes target motion.

| Regime and subset | Trajectories | Repeat-current TAcc@0.5 (%) | Oracle-CV TAcc@0.5 (%) |
|---|---|---|---|
| Short, all sources, full future | 123 | 77.45 | 90.11 |
| Short, all sources, high motion | 62 | 58.06 | 81.39 |
| Short, all sources, high acceleration | 32 | 57.17 | 82.13 |
| Short, KITTI, high motion | 45 | 2.96 | 67.41 |
| Long, all sources, full future | 251 | 77.91 | 91.70 |
| Long, all sources, high motion | 126 | 58.37 | 84.57 |

Repeat-current holds the latest reference position fixed; Oracle-CV extrapolates the velocity from consecutive reference states. Both use the same questions with two valid preceding states at distinct times, and evaluation includes timestamps after the latest initial state. Questions without the two states are excluded from this shared diagnostic set. All-source TAcc is computed within each source and then averaged equally over APIDIS, KITTI, and WildTrack.

Motion score is the reference trajectory's path length divided by its duration. Acceleration score averages the magnitudes of consecutive velocity differences divided by the time between their velocity samples. Within each regime, the high-motion group contains scores at or above the median, and the high-acceleration group contains scores at or above the 75th percentile, including boundary ties. Source-specific subsets are selected from these regime-level groups. Table 10 reports the resulting subset sizes. Holding position is less accurate in these groups. In the KITTI high-motion subset, Repeat-current attains 2.96% and Oracle-CV reaches 67.41%, showing how velocity information changes this reference-based estimate of the target trajectory.

## E Additional Tracking Comparisons

The sparse-RGB comparison evaluates all eligible trajectory questions across the three sources, corresponding to the Trajectory group in Table 4. All methods share this question set and reference eligibility rules. The KITTI comparison evaluates 58 trajectory questions with MCTrack's dense-frame inputs and LMM-Track4D's sparse observations. Missing predictions count as failures in Traj-Acc and TAcc; ADE averages errors for valid predictions.

Table 11: Tracking comparison on all eligible trajectory questions across the three sources. Methods use matched sparse RGB observations and Qwen3.5-9B target binding. ADE averages valid predictions; missing outputs count as failures in thresholded accuracies.

| Method | Traj-Acc@1.0 (%) ↑ | TAcc@0.5 (%) ↑ | ADE (m) ↓ |
|---|---|---|---|
| SORT (Bewley et al., 2016) | 4.35 | 1.03 | 12.02 |
| ByteTrack (Zhang et al., 2021) | 5.80 | 1.03 | 11.08 |
| BoT-SORT (Aharon et al., 2022) | 4.35 | 1.03 | 12.40 |
| TrackTacular-JDE (Teepe et al., 2024) | 4.35 | 0.44 | 13.26 |
| LMM-Track4D | **78.8** | **77.8** | **0.62** |

Table 12: Tracking comparison on 58 KITTI trajectory questions. ADE averages valid predictions. MCTrack (Wang et al., 2024) receives dense frames and released VirConv detections; LMM-Track4D uses benchmark-visible sparse RGB.

| Method | Traj-Acc@1.0 (%) ↑ | TAcc@0.5 (%) ↑ | ADE (m) ↓ |
|---|---|---|---|
| MCTrack + Qwen3.5-9B binding | 43.10 | 28.35 | 6.66 |
| LMM-Track4D | **60.34** | **59.42** | **0.71** |

The KITTI comparison complements the sparse-RGB setting in Table 11. MCTrack uses dense KITTI frames and released VirConv detections, followed by Qwen3.5-9B target binding; LMM-Track4D uses the benchmark's sparse observations.

The sparse-RGB baselines receive YOLO11x detections and estimate positions in the clip's calibrated coordinate frame. For APIDIS and WildTrack people, detected footpoints are projected onto the ground plane, and predicted appearance and ground locations establish correspondences across synchronized views. For APIDIS balls, associated detections provide calibrated rays from which ball centers are recovered, retaining height above the court.

For KITTI, the trackers associate detections across observed keyframes and share a frozen monocular 3D estimator with the same RGB inputs, candidate boxes, and calibration. Estimated box-bottom centers supply metric positions. Camera poses transform them into the fixed clip frame through Eq. 8, before the shared temporal completion and dialogue binding.

Temporal completion operates on valid position estimates of the same track. Two or more observed timestamps permit linear interpolation between consecutive positions; requested times outside their range use the nearest endpoint. One observed position is held throughout the requested interval, and a track with no valid position produces an invalid output. APIDIS and WildTrack supply one synchronized timestamp per Short clip, so their multiple cameras constrain position at that timestamp. KITTI supplies four temporal keyframes.

Target binding follows geometric adaptation and temporal completion. Qwen3.5-9B receives the question, preceding generated conversation, and predicted track candidates with their identifiers, appearance information, and positions. It selects the candidate corresponding to the dialogue reference. Source reference identities and positions remain in the evaluation records; unresolved or missing required outputs count as failures.

# F External Evaluation Details

Dyn-Bench (Huang et al., 2026b) evaluates spatiotemporal question answering and dynamic object grounding across Inter-Object, Object-Scene, and Camera-Object interactions. We follow its official zero-shot evaluation protocol and task groupings. Question answering uses exact matching of the selected multiple-choice answer. Grounding uses $\mathcal{J}\&\mathcal{F}$, the mean of mask intersection-over-union ($\mathcal{J}$) and boundary F-measure ($\mathcal{F}$) for the referred object. Within each category, the official evaluator averages question accuracies and per-example grounding scores, respectively. Table 6 reports the three categories separately for each task. The DynTrace question-answering row uses Qwen3-VL-8B; each category value averages its three published subtype accuracies (Gao et al., 2026).

The model answers external video questions through its frozen RGB representation, without calibrated RTGE ray embeddings or metric trajectory decoding. Dyn-Bench uses the final Stage-3 checkpoint, and VLM4D uses the jointly supervised Stage-2 checkpoint. Neither receives benchmark-specific updates.

For dynamic object grounding, LMM-Track4D resolves the queried entity from video and language. Its predicted reference conditions a frozen two-dimensional grounding and video-segmentation interface to produce per-frame masks. Neither component receives Dyn-Bench-specific updates. Official $\mathcal{J}\&\mathcal{F}$ scores evaluate this complete language-conditioned grounding system.

Table 13: Dyn-Bench prompt sensitivity on a paired 700-question subset. Each row uses the same frozen checkpoint, videos, and decoding settings.

| Prompt | Questions | Correct | Accuracy (%) ↑ | Parsed |
|---|---|---|---|---|
| Original | 700 | 425 | 60.71 | 693 |
| Variant 1 | 700 | 429 | 61.29 | 699 |
| Variant 2 | 700 | 427 | 61.00 | 700 |

Table 13 holds the 700-question set fixed across prompt formulations. Accuracy varies by 0.57 percentage points across the three prompts.

Table 14: Frozen Stage-2 zero-shot evaluation on VLM4D (Zhou et al., 2025). The overall video-cluster bootstrap 95% interval is [55.88, 60.36]%.

| Subset | Questions | Correct | Accuracy (%) ↑ | Parse rate (%) |
|---|---|---|---|---|
| Real video | 1,371 | 816 | 59.52 | 98.47 |
| Synthetic video | 445 | 240 | 53.93 | 99.33 |
| Overall | 1,816 | 1,056 | 58.15 | 98.68 |

The frozen Stage-2 checkpoint uses eight uniformly sampled frames per VLM4D question with deterministic decoding. It attains 59.52% on real videos and 53.93% on synthetic videos. These results provide an external evaluation of the jointly supervised representation, alongside the Stage-3 answer-policy evaluation on Dyn-Bench.

## G Additional Related Work

Visual-language alignment (Li et al., 2023) extends to video interaction in VideoChat and InternVL (Li et al., 2025b; Wang et al., 2025a) and general multimodal models (Hurst et al., 2024; Comanici et al., 2025; Bai et al., 2025a). LLaVA-ST, Compositional 4D Dynamic Scenes Understanding, and Uni4D-LLM further organize temporal or geometric representations (Li et al., 2025a; Wang et al., 2025b; Zhou & Lee, 2025a).

SpatialVLM, OmniSpatial, Flatland, and Spatial-MLLM study spatial supervision and evaluation (Chen et al., 2024; Jia et al., 2025; Zhang et al., 2026; Wu et al., 2026). Scene representations include 3D-LLaVA's point-cloud superpoints (Deng et al., 2025), instance geometry in Inst3D-LMM (Yu et al., 2025), condensed grids in LEO-VL (Huang et al., 2025a), and native object grounding in N3D-VLM (Wang et al., 2025d). Learning Videos in 3D, Chat-Scene, and 3D-R1 address cross-view learning, scene dialogue, and reasoning (Zheng et al., 2026; Huang et al., 2024; 2025b).

MeViS and VISA evaluate temporally grounded language understanding (Ding et al., 2023; Yan et al., 2025). PointOdyssey, Dynamic Stereo, Stereo4D, Motion4D, and Understanding Dynamic Scenes in 4D provide tracking, motion, or reconstruction signals (Zheng et al., 2023; Karaev et al., 2023; Jin et al., 2025b; Zhou & Lee, 2025b; Huang et al., 2026a). Track4D-Bench links dialogue references and requested intervals to explicit metric trajectories, so the conversation determines the entity and physical times being evaluated.

## H Limitations

Track4D-Bench covers calibrated scenes, finite clips, and five dialogue categories, with KITTI supplying the largest share of the three-source collection. Geometric evaluation measures metric trajectories within these scenes. The component analyses measure language and trajectory accuracy; they do not separately score reference resolution or retrospective interval selection. Dyn-Bench extends evaluation to dynamic-scene question answering and mask grounding, and VLM4D evaluates external video questions. Metric trajectory recovery across additional scene sources and longer conversations remains a direction for further study.

## I Additional Qualitative Examples

Figures 5, 6, and 7 illustrate the dialogue format in sports, driving, and pedestrian scenes. The images show changes over the queried interval, and the accompanying answers display entity references and trajectory fields at the source dialogue's temporal resolution. Within these figures, long and short label answer verbosity for the same question; the benchmark's Short and Long regimes describe temporal scope. Quantitative performance is reported in the experimental tables.

## APIDIS



Q1 Q2 Q3 Q4 Q5

**Q1:Who is the Team A player near the basket at the start of the segment?**

GT (long):
> At the start of the segment, the Team A player closest to the basket (near x = -0.76 m) is number 9, wearing yellow. He remains relatively stationary in the corner near the baseline throughout the early part of the play. I can track his full movement if you'd like.

GT (short):
> Team A number 9 (yellow), standing in the corner near the baseline at x ≈ -0.76 m and barely moving early in the clip.

Ours (long):
> The closest Team A player to the basket at the opening frames is number 9 in yellow, positioned in the baseline corner around x ≈ -0.8 m. He stays nearly fixed there during the opening half-second; I can pull his full per-step trajectory on request.

Ours (short):
> Team A #9 (yellow) — corner of the baseline at x ≈ -0.8 m, almost stationary at the start.

**Q2:Yes, please track Team A number 9 over the full segment.**

GT (long):
> Certainly. Here is the full trajectory for Team A number 9 over the segment: '[[-76, 957, 88, 72, 72, 175], [-76, 957, 88, . . .], . . ., [-78, 940, 88, 72, 72, 175]]'. He stands near the corner baseline and slides only ≈ 0.2 m forward over the 1.24 s window.

GT (short):
> 32 steps of #9: starts '[-76, 957, 88]', ends '[-78, 940, 88]' cm — barely moves (~ 0.2 m forward).

Ours (long):
> Tracking Team A #9 for the full clip (32 steps, cm): '[[-76, 957, 88, 72, 72, 175], . . ., [-78, 940, 88, 72, 72, 175]]'. He is essentially anchored in the baseline corner — about 0.2 m of forward shift across the 1.24 s window, no lateral motion.

Ours (short):
> #9 trajectory (32 steps, cm): '[-76, 957, 88]' → '[-78, 940, 88]'; net motion ≈ 0.2 m forward.

**Q3:How does Team A number 4's position relate to the referees during the first half-second?**

GT (long):
> During the first half-second, Team A number 4 starts near the top of the key and moves slightly rightward. He remains well away from both referees — the one near the baseline stays far behind him, and the sideline referee is positioned on the opposite side of the court, near the far wing. There's no close interaction between him and either official in that window.

GT (short):
> Team A #4 starts near the top of the key, drifts right, and stays well clear of both referees in the first 0.5 s.

Ours (long):
> Within the first ≈ 0.5 s, Team A number 4 sits near the top of the key and shifts a touch to the right. Neither referee comes close: the baseline official is several metres behind him and the sideline official is on the opposite wing of the court, so there is no proximity event between them in this window.

Ours (short):
>#4 near the top of the key, drifting right; both referees stay several metres away — no close interaction in the first 0.5 s.

**Q4:What happens around 0.8 s in this segment?**

GT (long):
> Around 0.8 s, Team B number 6 releases a three-point shot, assisted by Team B number 9. The ball is already descending by then, and the shot ultimately misses, leading to a rebound situation. I can identify the shooter or track the ball trajectory if you'd like.

GT (short):
> Around 0.8 s, Team B #6 releases a three-pointer (assisted by #9); the ball is descending and the shot misses.

Ours (long):
> Near the 0.8 s mark, Team B #6 lifts for a three-point attempt off an assist from #9; by that frame the ball is already on its way down. The shot does not fall, so a rebound contest follows. Trajectory of the shooter or the ball is available if needed.

Ours (short):
> ~ 0.8 s: #6 takes a three (assist #9), ball descending, shot misses → rebound to follow.

**Q5:Describe the overall scene in this segment.**

GT (long):
> This segment shows Team A (yellow) and Team B (black) in a half-court setup near one end of the court. Early on, several players are spread across the wings and baseline areas, with noticeable movement from both teams. The ball starts low near the paint and gradually moves outward and upward over time, suggesting a developing offensive action. Two referees are visible — one near the sideline close to the baseline, and another higher up near the top of the key.

GT (short):
> Half-court set: yellow (A) vs black (B) spread across wings and baseline; ball moves from the paint outward and upward, two referees visible.

Ours (long):
> The clip captures a half-court possession on one end of the floor. Team A (yellow) and Team B (black) are deployed across the wings and the baseline; player motion is active on both sides. The ball begins low near the paint and progressively shifts outward and upward, consistent with the early stage of an offensive set. Two officials are framed in shot — one along the baseline-side sideline, one higher up near the top of the key.

Ours (short):
> Half-court possession; A (yellow) and B (black) on wings/baseline; ball starts low in the paint and rises outward; two referees visible.

Figure 5: APIDIS dialogue example showing sports entities, timestamped questions, and trajectory-form answers with different levels of detail. Original centimeter-scale records include person box centers and dimensions. Geometric evaluation uses metric ground-reference positions for people and centers for balls, following the conversion in Appendix B.

## KITTI



Q1 Q2 Q3 Q4 Q5

**Q1:Which object matches this coordinate snippet: '[1.76, 1.602, 3.84, -25.952, -3.213, 355.503, -1.742]'?**

GT (long):
> That corresponds to a silver car moving in the rightmost lane. It's steadily progressing forward throughout the clip. I can provide its full trajectory if needed.

GT (short):
> A silver car in the rightmost lane, moving steadily forward.

Ours (long):
> Those coordinates belong to a silver car driving in the right-most lane; it advances with a roughly constant forward speed across the whole clip. Its full per-step trajectory is available on request.

Ours (short):
> Silver car in the rightmost lane, advancing at roughly constant speed.

**Q2:Track the silver car moving in the rightmost lane over the full clip, and note when it is visible (seconds).**

GT (long):
> The silver car in the rightmost lane is visible throughout the entire 3.1-second clip. Here is its full trajectory: '[[1.76, 1.602, 3.84, -16.079, -2.798, 290.631, -1.712], . . ., [1.76, 1.602, 3.84, -2.85, 293.10, -1.71]]' (32 steps).

GT (short):
> Silver car visible 0.0 s – 3.1 s; 32-step trajectory from '(-16.08, -2.80, 290.63)' to roughly '(-2.85, 293.10, · · ·)' m.

Ours (long):
> The silver car stays in shot for the full 3.1 s window. Its 32-step pose trajectory in metres (w/h/l identical at 1.76/1.60/3.84): '[[1.76, 1.60, 3.84, -16.08, -2.80, 290.63, -1.71], . . ., [1.76, 1.60, 3.84, -2.85, 293.10, -1.71]]'. Heading remains roughly constant (yaw ≈ -1.71 rad).

Ours (short):
> 32 steps over 0–3.1 s; (x, y, z) goes from '(-16.08, -2.80, 290.63)' to '(-2.85, 293.10, · · ·)' m, yaw ≈ -1.71 rad.

**Q3:How does the lateral position of the red car change relative to the white car over the first 1.5 seconds?**

GT (long):
> Over the first 1.5 seconds, the red car remains consistently to the left of the white car. The gap between them stays fairly stable, with the red car maintaining its position along the left side of the road while the white car travels in the centre lane. I can also describe how their longitudinal spacing evolves during that time if you'd like.

GT (short):
> Red car stays to the left of the white car throughout the first 1.5 s; lateral gap is roughly constant.

Ours (long):
> Across the first 1.5 s the red car holds a consistent position to the left of the white car, with their lateral separation remaining stable; the red car keeps to the left side of the road while the white car runs the centre lane. Longitudinal spacing trend is available on request.

Ours (short):
> Red car stays left of white car over 0–1.5 s; lateral gap roughly constant (red on left edge, white in centre lane).

**Q4:How does the motion of the silver car in the rightmost lane change over the clip?**

GT (long):
> The silver car in the rightmost lane maintains a steady forward motion throughout the clip, gradually increasing its speed slightly while staying in its lane without significant lateral movement. I can also compare its motion with other nearby vehicles if you'd like.

GT (short):
> Silver car keeps a steady forward motion, mildly accelerating, no lateral deviation.

Ours (long):
> The silver car in the right-most lane drives forward steadily across the whole clip. There is a small, gradual increase in speed, and its lateral position barely changes — it stays inside its lane the entire time. A relative-motion comparison with neighbouring vehicles is available on request.

Ours (short):
> Silver car: steady forward motion, mild speed gain, no lane change.

**Q5:What moving vehicles are present in the scene, and roughly when do they appear?**

GT (long):
> There are three main moving vehicles: a white car traveling in the middle of the road (visible throughout the full 3.2 seconds), a red car on the left side of the road (also visible the entire time), and a dark van moving forward in the middle lane (visible from start to end as well). Additionally, a dark car appears briefly on the left road only during the last 0.6 seconds (from 2.6 s to 3.1 s).

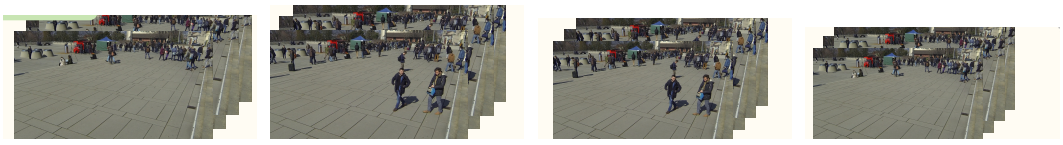
GT (short):
> White car, red car, and a dark van all visible 0.0–3.1 s; an extra dark car appears briefly 2.6–3.1 s on the left.

Ours (long):
> Three vehicles are visible throughout the 3.2 s clip: a white car in the middle of the road, a red car on the left, and a dark van in the centre lane. A fourth, a dark car, enters from the left and is only present in the last 0.6 s (≈ 2.6 s – 3.1 s).

Ours (short):
> 3 vehicles for the full clip (white centre, red left, dark van centre) + 1 dark car briefly on the left during 2.6–3.1 s.

Figure 6: KITTI dialogue example showing object references and local trajectories in a driving scene. The abbreviated source-format records contain box dimensions, bottom-center position, and yaw. The original response labels dimensions as w/h/l; the source records use $(h, w, l)$, as defined with the coordinate conversion in Appendix B.

## WildTrack



Q1 Q2 Q3 Q4

**Q1:Who is the pedestrian standing still near y = 9.4 early in the clip?**

GT (long):
> That appears to be a young woman with light brown hair tied back, wearing a dark puffer jacket, a light-coloured scarf, blue jeans, and light gray sneakers, carrying a small shoulder bag and holding an object in her hands. She remains nearly stationary near that spot until about 2.0 s before making slight movements. I can track her full trajectory if you'd like.

GT (short):
> Young woman, dark puffer + light scarf + jeans + gray sneakers + small shoulder bag; stays at y ≈ 9.4 until ≈ 2.0 s.

Ours (long):
> Near y ≈ 9.4 in the early frames is a young woman with brown hair tied back, in a dark puffer jacket and a light scarf, blue jeans and light gray sneakers, carrying a small shoulder bag and holding something in her hands. She stays nearly fixed until roughly 2 s before showing minor movement. Per-frame trajectory available on request.

Ours (short):
> Young woman in dark puffer + light scarf, jeans, gray sneakers, shoulder bag — stationary at y ≈ 9.4 until ~ 2 s.

**Q2:Yes, please track that person over the full clip.**

GT (long):
> Certainly. Here is the trajectory for the person in the dark beanie with a shoulder bag over the full clip: '[[-1.98, 0.50, 0.9, . . .], [-1.80, 0.42, 0.9, . . .], . . ., [-1.07, 0.23, 0.9, 0.49, 0.49, 1.81, 0.28]]'. He is continuously observable from 0.0 s to 3.5 s, steadily advancing and slightly adjusting his orientation.

GT (short):
> 8-step trajectory of the man in the dark beanie + shoulder bag, 0.0–3.5 s, from '(-1.98, 0.50)' to '(-1.07, 0.23)' m; mild yaw adjustment.

Ours (long):
> The man in the dark beanie with a shoulder bag is continuously visible for the full 3.5 s. Trajectory in metres: '[[-1.98, 0.50, 0.9, . . .], . . ., [-1.07, 0.23, 0.9, 0.49, 0.49, 1.81, 0.28]]' — he advances steadily and adjusts heading slightly across the segment.

Ours (short):
> 8 steps, 0.0–3.5 s, from '(-1.98, 0.50)' → '(-1.07, 0.23)' m; small heading change.

**Q3:Is anyone walking close to him during the clip?**

GT (long):
> He remains mostly separate from others, with no one consistently near him. A few pedestrians pass at a short distance later in the clip, but none appear to be walking together with him or closely following. His path stays relatively isolated through the centre corridor. I can identify the nearest passerby around 2.5 s if useful.

GT (short):
> No, he stays relatively isolated; a few pedestrians pass nearby later but no one is walking with him.

Ours (long):
> No one accompanies him through the clip. His route along the central corridor is largely isolated; later in the segment a couple of pedestrians cross within a short distance, but none is following or walking alongside. Identification of the nearest passerby around the 2.5 s mark is available on request.

Ours (short):
> Largely isolated — a couple of pedestrians pass close later, but none walks with him.

**Q4:What does the overall pedestrian activity look like in this scene?**

GT (long):
> The scene shows moderate pedestrian activity concentrated along a central corridor of the platform, with several individuals moving steadily westward or lingering near key zones. Early in the clip, movement is active across multiple lanes, while by around 2.0 s, some pedestrians begin to slow or change direction near the centre. A few individuals remain stationary for extended periods, particularly near the southern and northern edges. Most reliable tracks are grounded within the main platform area, with densest motion occurring between y ≈ 9 and y ≈ 17.

GT (short):
> Moderate pedestrian flow along a central corridor; activity densest between y ≈ 9 and y ≈ 17; some slow or change direction near 2.0 s.

Ours (long):
> The scene captures moderate foot traffic along a central platform corridor. Early on, motion is spread across multiple lanes, with most pedestrians heading westward or pausing in key zones; around the 2 s mark, several start to slow or change direction near the centre. A handful of people stay still for extended periods near the north and south edges. The densest motion sits in the y ≈ 9–17 band, where most reliable tracks are grounded.

Ours (short):
> Moderate flow along the central corridor (y ≈ 9–17); some slow or turn near the centre at ≈ 2 s; a few stationary individuals at the north / south edges.

Figure 7: WildTrack dialogue example illustrating pedestrian references and trajectory-form answers. The displayed source-format records contain box centers, dimensions, and yaw. Geometric evaluation uses pedestrian ground-reference positions after the conversion in Appendix B.