SCENEFORESEE: LEARNING SCENE EVOLUTION FOR SPATIOTEMPORAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large multimodal models (LMMs) need to understand how entities move and interact as a scene develops. This requires following the same entities over time and relating their observed motion to subsequent changes. We formulate Scene Evolution Understanding (SEU), a task for predicting and interpreting this change. Given observed video, a model estimates entity histories and predicts one scene continuation through future trajectories and Events, accompanied by Visual Evidence from observed motion and a scene description. We construct SceneForesee, a benchmark for evaluating entity continuity and scene evolution reasoning through paired observed and future scenes with shared entity identities. It contains 144.57K temporal pairs, 1.05M entity correspondences, and 272.49K Events paired with observed evidence. We propose a spatiotemporal reasoning framework, EvoSee, that learns future geometry and Event semantics through shared candidate states and joint assignment. It selects one continuation per entity from observation and uses the selected changes for entity refinement, retention, and scene representation. On SceneForesee, EvoSee outperforms Qwen3.5-9B by 10.45 points in Combined Point-F1 and 17.24 points in Event Global F1. On the 4D reasoning benchmark Dyn-Bench, it improves question answering accuracy and grounding J&F by 4.8 and 3.58 points over their respective base systems. These results support scene evolution as both a prediction target and a dynamic representation that brings predicted entity changes into spatiotemporal reasoning.

1 INTRODUCTION

Large multimodal models (LMMs) are moving from visual recognition toward reasoning about motion and interaction in videos (Bai et al., 2025a; Huang et al., 2026). An entity moving behind another retains its identity even as its position, visibility, and spatial relations change. This continuity makes the changes attributable to the same entity, while its observed motion informs what follows. **Scene evolution** concerns the development of persistent entities and their relations over time, placing continuity at the center of dynamic video reasoning.

Continuity and change appear across several lines of dynamic video research. Video language models interpret extended visual content (Bai et al., 2025b; Wang et al., 2025), while motion grounding and evidence localization relate answers to dynamic observations (Deng et al., 2025; Gao et al., 2026; Wei et al., 2026). Tracking and forecasting establish entity identities and predict their spatial continuation (Meinhardt et al., 2022; Dendorfer et al., 2022; Ashutosh et al., 2025). Event prediction and predictive representations address the changes that follow observation (Wang et al., 2026; Assran et al., 2025). At the intersection of these capabilities, we study how spatial and semantic predictions can refer to the same observed entities and jointly inform scene reasoning.

Relating spatial and semantic change requires knowing which observations belong to the same entity. Persistent identity supplies this reference when observation ends and future prediction begins. In Figure 1(a,b), following the console and the person's hands makes their changing positions and interaction attributable to the same participants. A trajectory describes where an entity moves, while an Event summarizes its state or relations. These spatial and semantic descriptions concern the same continuation and can be learned and evaluated with a common entity reference. Visual Evidence records the observed motion associated with that continuation, such as the changing grip and console

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.

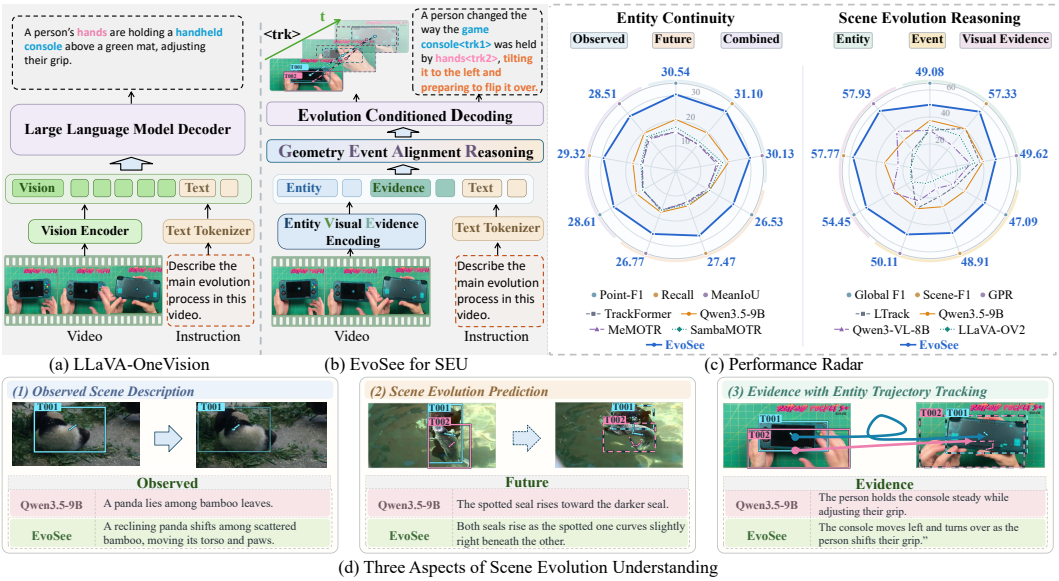


Figure 1. Scene Evolution Understanding. (a,b) Description and entity continuation from observation. (c) Eighteen continuity and reasoning metrics, plotted as percentage scores; labels give EvoSee values. (d) Observed movement, anticipated interaction, and evidence. Future frames show the evaluation reference. Tracking baselines denote task adaptations.

orientation in the example. Together, these descriptions relate predicted change to the entities and motion from which it is inferred, providing context for scene reasoning.

We formulate **Scene Evolution Understanding** (SEU) as a task for predicting and interpreting how persistent entities develop. From observed video, a model estimates entity histories and infers one scene continuation through future trajectories and Events. Its output also identifies associated Visual Evidence in observation and describes the evolving scene. We construct **SceneForesee**, a benchmark for evaluating scene evolution understanding through paired observed and future intervals with shared entity identities. Its **144.57K** temporal pairs contain **1.05M** entity correspondences and **272.49K** Events paired with observed evidence. This organization supports complementary evaluations of entity continuity and scene evolution reasoning on the same scenes.

To learn and use scene evolution, we propose a spatiotemporal reasoning framework, **EvoSee**. EvoSee treats the observed entity as the reference for both its future motion and the Events in which it participates. For each entity, it learns candidate continuations whose geometric and Event predictions share future states and joint assignment during training. It selects one candidate using the observed entity state and evidence. The selected change then refines the entity state and contributes to entity retention and scene representation.

EvoSee outperforms the evaluated baselines in entity continuity and Event category recognition on SceneForesee. Against Qwen3.5-9B, it also improves complete Event prediction, where the category and all participants must match. Ablations relate these gains to joint assignment, continuation selection, and the use of predicted change. The value of this change extends to question answering and grounding on Dyn-Bench (Huang et al., 2026), where the learned scene representation supports 4D reasoning.

Our contributions are summarized as follows.

- We formulate a new task, **Scene Evolution Understanding**, which aims to predict and interpret how persistent entities evolve over time. The task represents one scene continuation through future entity motion and Events, together with associated visual evidence from observation.
- We introduce **SceneForesee**, a benchmark for evaluating entity continuity and scene evolution reasoning. Shared identities associate observed motion, future trajectories, and Events, supporting complementary evaluations of continuity and change on the same scenes.

- We propose **EvoSee**, an entity-centric spatiotemporal reasoning framework that models future geometry and Event semantics as complementary aspects of scene evolution. Shared future states and joint assignment align geometric and Event predictions for each entity, and the selected changes update entity and scene representations.
- We demonstrate experimentally that modeling scene evolution improves future prediction and observed entity estimation. The learned representations also improve dynamic question answering and grounding on the **4D reasoning benchmark Dyn-Bench**.

2 Related Work

Following entities through time. Tracking establishes correspondence between observations of the same entity. Object query trackers maintain this correspondence across frames (Meinhardt et al., 2022; Gao & Wang, 2023), while other formulations predict identities (Gao et al., 2025) or associate tracked entities with descriptions and referring expressions (Li et al., 2024; Wu et al., 2023). These entity histories support forecasting of subsequent locations and interactions beyond observation (Dendorfer et al., 2022; Lu et al., 2026; Liang et al., 2019; Ashutosh et al., 2025; Nguyen et al., 2025). SEU builds on this identity correspondence to attribute future motion and Events to the same observed entities.

Predicting scene development. Beyond future locations, prediction concerns changes in entity states and relations. FutureBench and FeVOS study future events (Wang et al., 2026; Lan et al., 2026), while scene graph forecasting represents changes through entities and their relations (Pistilli et al., 2026). Anticipatory detection and predictive representations learn expected future states (Papais et al., 2025; Assran et al., 2025). SEU uses persistent identities to relate future trajectories, Events, and observed Visual Evidence within the same scene continuation.

Reasoning with motion and visual evidence. Video language models use dynamic visual content to produce descriptions and answers (Bai et al., 2025b; Wang et al., 2025; Zhang et al., 2025). Motion grounding and evidence localization identify observations relevant to these outputs (Deng et al., 2025; Galoaa et al., 2026; Wei et al., 2026). DynTrace represents observed motion through traces, and Dyn-Bench evaluates dynamic question answering and grounding (Gao et al., 2026; Huang et al., 2026). EvoSee uses predicted entity continuations to enrich the scene representation for such reasoning.

3 Task and Benchmark

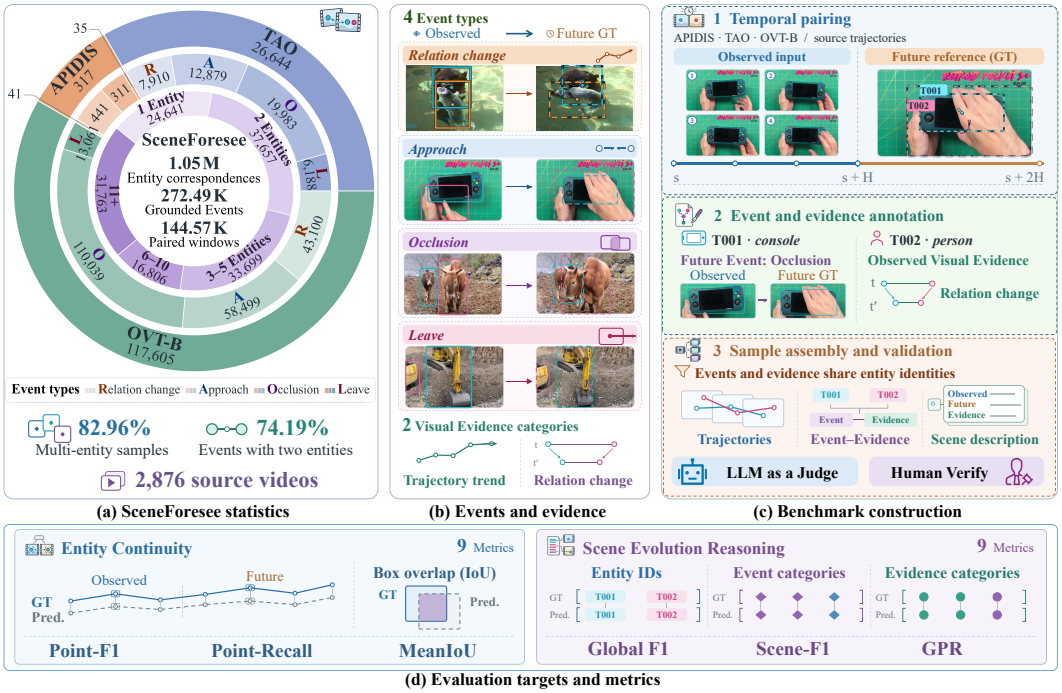
3.1 Scene Evolution Understanding

Scene Evolution Understanding requires a model to estimate entity histories from observed video and predict one scene continuation through future trajectories and Events with explicit participants. The observed entities provide the reference for these predictions. The output also identifies Visual Evidence in observation and describes the scene evolution. Figure 1(d) illustrates observed motion, anticipated interaction, and visual evidence through example scenes. Future frames supply training and evaluation targets; Appendix A specifies sampling and output format.

3.2 Benchmark Construction

SceneForesee pairs adjacent, equally long observed and future intervals from APIDIS, OVT-B, and TAO, covering sports interactions, open category entities, and varied natural motion (De Vleeschouwer et al., 2008; Liang & Han, 2024; Dave et al., 2020). Timestamped boxes preserve shared identities between observed and future intervals, with at least one continuing entity per sample. These identities anchor the annotations in Figure 2(c).

The four Event categories in Figure 2(b) summarize trajectory configurations, providing discrete targets for predicting evolution under shared identities. Approach uses decreasing separation. Occlusion uses box overlap throughout the image at the transition from observation to prediction. Leave uses proximity to the image edge at the future endpoint, while relation change covers remaining



(a) SceneForesee statistics (b) Events and evidence (c) Benchmark construction

(d) Evaluation targets and metrics

Figure 2. SceneForesee. (a) Corpus statistics. (b) Event and Visual Evidence categories. (c) Temporal pairing, annotation, and sample validation. (d) Entity Continuity and Scene Evolution Reasoning metrics. Ring labels give counts; sector angles are schematic.

configurations. Each Event records participants and a future interval, paired with observed Visual Evidence of individual trajectory trends or pairwise relation changes. Figure 2(c) pairs the console's anticipated occlusion with the entities' observed relation. A scene description expresses their motion and development.

Common annotation criteria apply to all sources, with LLM assessment and manual inspection of generated annotations. Data splits follow source videos and disjoint APIDIS episodes, keeping overlapping windows together. Figure 2(a) summarizes 2,876 videos, with multiple entities in 82.96% of samples and two participants in 74.19% of Events. Appendix A.2 details construction and splits.

3.3 Evaluation Metrics

Figure 2(d) presents complementary evaluations of entity continuity and scene evolution reasoning on the same temporal pairs.

Entity Continuity. Identity matching from observed trajectories remains fixed across all intervals. Point-F1@0.5 and Point-Recall@0.5 measure recovery of entity locations under this correspondence, while MeanIoU measures spatial overlap. Combined scores weight observed and future scores by their annotated point counts.

Scene Evolution Reasoning. Entity scores evaluate identities referenced in descriptions; Event and Visual Evidence scores evaluate category occurrences. Global F1 and the geometric mean of global precision and recall (GPR) pool matches across scenes; Scene-F1 weights scenes equally. Complete Event micro-F1 additionally requires the category and all participants to match. BERTScore F1 measures similarity between complete descriptions (Zhang et al., 2020). Appendix A defines matching and aggregation.

4 Method

EvoSee learns entity continuations from observed video and uses the predicted changes for scene reasoning. Figure 3 shows how observed entity states and evidence condition shared geometric and

Event predictions. A selector uses the observed representations to choose one continuation per entity. The selected change refines the entity state and contributes to retention and scene representation.

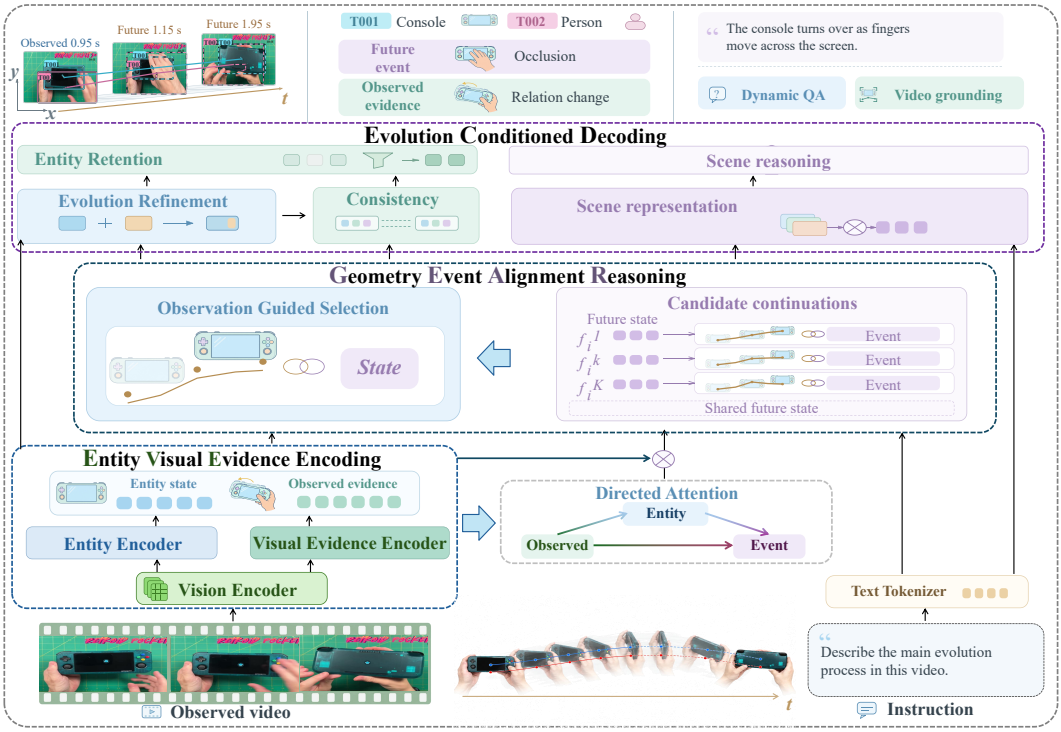


Figure 3. EvoSee. Entity state and evidence condition shared geometric and Event continuations. Selected changes inform refinement, retention, and scene representation. Entity in directed attention denotes geometric state.

4.1 Entity Visual Evidence Encoding

Entity Visual Evidence Encoding (EVE) uses a shared learned query q_i to encode entity state and observed motion. The query attends to observed visual memory X_{obs} for state a_i at the observation boundary and to frame features X_t for temporal states h_i^t (Meinhardt et al., 2022; Vaswani et al., 2017). Normalized relevance scores aggregate these temporal states into evidence e_i. Evidence intervals supervise temporal relevance, and evidence categories supervise the aggregated representation. Appendix B.1 gives the encoding equations.

4.2 Geometry Event Alignment Reasoning

Geometry Event Alignment Reasoning (GEAR) infers future geometry and Event semantics from entity state and evidence. Their sum $u_i^a = a_i + e_i$ initializes geometric and semantic states u_i^g, u_i^s. Stacking u_i^a, u_i^g, u_i^s in this order gives U_i. Each attention head h projects normalized U_i into queries Q_i^h, keys K_i^h, and values V_i^h of width d_h, computing output A_i^h under mask $\mathcal{M}$.

$$A_i^h = \text{softmax}\left(\frac{Q_i^h (K_i^h)^\top}{\sqrt{d_h}} + \mathcal{M}\right) V_i^h, \qquad \mathcal{M} = \begin{pmatrix} 0 & -\infty & -\infty \\ 0 & 0 & -\infty \\ 0 & 0 & 0 \end{pmatrix}. \tag{1}$$

The mask implements the dependencies in Figure 3, allowing each state to attend only to itself and preceding states. Concatenated head outputs are projected for a residual update of U_i.

Each updated state independently weights feedforward transformations, preserving these dependencies. The resulting geometric and semantic states condition a future sequence $f_i \in \mathbb{R}^{L \times d}$ with L time steps and feature width d. Each of K candidate continuations uses embedding $d_k \in \mathbb{R}^d$ and transition ϕ_k (Rupprecht et al., 2017). The vector $\mathbf{1}_L$ broadcasts the embedding across time. As depicted in Figure 3, decoders D_g and D_s share candidate state f_i^k to predict geometry g_i^k and Event logits s_i^k,

with task context embedding c informing semantics.

$$f_i^k = f_i + \phi_k(f_i + \mathbf{1}_L d_k^\top), \qquad g_i^k = D_g(f_i^k), \qquad s_i^k = D_s(f_i^k; c). \tag{2}$$

Joint assignment selects one candidate for training against annotated future geometry $g_i^\star$ and Event targets $s_i^\star$, using their combined prediction error.

$$\widehat{k}_i \in \arg\min_k \left\{ \ell_{\text{geo}}(g_i^k, g_i^\star) + \lambda_s \ell_{\text{sem}}(s_i^k, s_i^\star) \right\}. \tag{3}$$

Here ℓ_{geo} measures box error, ℓ_{sem} measures Event classification error, and $\lambda_s = 0.20$. Each Event supplies its category target to every participant, including both identities for pairwise Events. All future supervision uses the assigned index, coupling geometric and Event learning within the same candidate.

The assigned index also supervises selection from observation. A selector R maps entity state, evidence, and observed motion embedding o_i into candidate probabilities $\boldsymbol{\rho}_i$. It learns $\widehat{k}_i$ through negative log likelihood. At inference, the most probable candidate k_i^* determines future boxes, visibility, persistence, and Event scores, giving one continuation per retained entity. Appendix B.2 details selection and the endpoint separation regularizer used during training.

4.3 Evolution Conditioned Decoding

Evolution Conditioned Decoding (ECD) refines the observed entity state using predicted change. Projecting temporally averaged future boxes and the sigmoid of temporally averaged Event logits gives summaries $\mathbf{g}_i$ and $\mathbf{s}_i$. The summaries use probability weighted decoded predictions during training and the selected candidate at inference. Map Φ combines them with observed state and evidence into evolution representation $\mathbf{m}_i$. A sigmoid gate from Γ weights the projected evolution representation $P_f(\mathbf{m}_i)$.

$$\mathbf{m}_i = \Phi([a_i; \mathbf{g}_i; \mathbf{s}_i; e_i]), \qquad a_i^+ = a_i + \sigma(\Gamma([a_i; \mathbf{m}_i])) \odot P_f(\mathbf{m}_i). \tag{4}$$

Here σ is sigmoid and $\odot$ denotes elementwise multiplication. The residual keeps a_i as a reference, while the updated state supports observed trajectory estimation and entity confidence.

Entity retention combines confidence ℓ_i^{obj} from updated state a_i^+ with agreement among observed and predicted features. Six cosine distances measure relations among evidence, entity state, geometry, semantics, evolution representation, and task context. Their mean B_i enters retention score $\kappa_i = \ell_i^{\text{obj}} - \lambda_{\text{cons}} B_i$, with nonnegative weight λ_{cons}. A margin ranking loss favors lower distances for queries matched to annotations than for unmatched queries. The highest retention scores define set $\mathcal{I}^*$. Appendix B.3 specifies the relations and supervision.

Retained entities supply trajectories and Event scores from their selected continuations. Complete Events specify a category and all participants using predicted entity IDs, which evaluation maps through the observation correspondence. For scene reasoning, EvoSee averages entity states a_i, evidence e_i, and evolution representations $\mathbf{m}_i$ over all queries. Their concatenation with task context c is projected into scene representation h_{scene} (Appendix B.3). This representation supplies observed and predicted context for description, dynamic question answering, and grounding (Deng et al., 2025; Gao et al., 2026).

5 Experiments

SceneForesee comparisons share test samples, observation intervals, temporal extents, and metrics. Qwen3.5-9B and EvoSee predict trajectories; other language models use shared predictions from the TrackFormer-style task adaptation (Meinhardt et al., 2022). Tables 1–2 report eighteen metrics as percentages; ablation and difficulty tables average each metric triplet. Ablations share training and evaluation conditions. Dyn-Bench evaluates the learned scene representation for 4D reasoning. Appendix C.1 details both protocols.

5.1 Comparison with Baselines

Entity Continuity. Table 1 shows EvoSee leading all nine continuity metrics. Combined Point-F1 exceeds Qwen3.5-9B by 10.45 points, while Future Point-F1 exceeds SambaMOTR by 10.92 points.

Recall and overlap gains indicate more complete recovery of entity locations and more accurate spatial continuation.

Table 1. Entity Continuity. P-F1 and P-R denote Point-F1@0.5 and Point-Recall@0.5. Underlined best baselines define red gains; bold marks best results. Tracking adaptations are trained for SceneForesee.

Method	Observed			Future			Combined		
	P-F1	P-R	mIoU	P-F1	P-R	mIoU	P-F1	P-R	mIoU
TrackFormer-style	15.59	13.94	17.38	14.46	13.11	17.08	15.05	13.53	17.24
MeMOTR-style	15.62	13.71	16.55	15.00	13.44	16.43	15.35	13.60	16.50
LTrack-style	14.59	13.19	16.13	12.78	11.60	15.58	13.72	12.42	15.87
SambaMOTR	17.46	14.98	18.94	15.61	14.16	16.61	15.28	14.16	17.84
Qwen3.5-9B	20.55	19.72	21.58	15.49	14.78	17.55	18.16	17.35	19.65
EvoSee (Ours)	**30.54** (+9.99)	**31.10** (+11.38)	**30.13** (+8.55)	**26.53** (+10.92)	**27.47** (+12.69)	**26.77** (+9.22)	**28.61** (+10.45)	**29.32** (+11.97)	**28.51** (+8.86)

Scene Evolution Reasoning. Table 2 shows Global F1 gains of 8.62 points over Qwen2.5-VL-7B for Entity, 17.24 over Qwen3.5-9B for Event, and 22.22 over Qwen3-VL-8B for Visual Evidence. Scene-F1 gains also hold when scenes receive equal weight. Figure 1(c) places these improvements alongside entity continuity across all eighteen metrics.

Table 2. Scene Evolution Reasoning. Event and Evidence match category occurrences. Underlined best baselines define red gains; bold marks best results. OV1.5 and OV2 are 8B LLaVA-OneVision models.

Method	Entity			Event			Visual Evidence		
	GF1	SF1	GPR	GF1	SF1	GPR	GF1	SF1	GPR
Tracking and entity continuity methods									
TrackFormer-style	29.59	40.42	37.18	19.56	16.97	28.33	12.17	10.59	17.63
MeMOTR-style	32.45	42.63	40.31	21.03	18.13	28.72	14.18	13.29	19.37
LTrack-style	30.18	41.06	37.60	22.30	20.32	29.87	15.03	14.04	20.12
SambaMOTR	33.91	42.25	43.02	21.87	20.06	28.10	16.46	13.46	21.93
Multimodal and video language models									
Qwen3.5-9B	37.45	40.73	40.14	29.85	28.20	29.87	29.82	34.23	29.84
Qwen3-VL-8B	30.04	25.03	30.39	19.90	13.78	23.56	32.23	23.90	38.17
InternVL3.5-8B	32.92	30.53	33.92	15.86	12.56	17.20	30.13	27.32	32.66
Qwen2.5-VL-7B	40.46	51.21	43.24	11.67	12.05	11.93	27.48	34.29	28.10
LLaVA-Video-7B	35.84	38.11	36.16	10.16	8.49	10.67	20.43	22.10	21.47
LLaVA-OV1.5	33.69	34.81	34.63	12.79	9.58	13.42	24.11	21.04	25.30
LLaVA-OV2	33.54	34.14	33.75	9.21	7.86	9.81	15.90	14.79	16.94
Proposed method									
EvoSee (Ours)	**49.08** (+8.62)	**57.33** (+6.12)	**49.62** (+6.38)	**47.09** (+17.24)	**48.91** (+20.71)	**50.11** (+20.24)	**54.45** (+22.22)	**57.77** (+23.48)	**57.93** (+19.76)

Table 3 evaluates complete Events and scene descriptions. EvoSee raises micro-F1 from Qwen3.5-9B's 19.94 to 40.76 when category and all participants must match together. Although the two methods have similar BERTScore F1, EvoSee is more accurate when Event categories and participants are evaluated jointly. Appendix C.6 gives present, future, and evidence description scores.

Table 3. Description similarity and complete Events.

Method	BERTScore F1	Micro-F1
Qwen3.5-9B	95.98	19.94
EvoSee (Ours)	**96.26**	**40.76**

5.2 Ablation Study and Discussion

Table 4 tests continuation learning and the use of predicted change. Appendix C gives complete metrics and additional variants.

Role of joint assignment. Rows 1 and 8 of Table 4 compare geometric and joint assignment. With Event supervision retained, assignment from geometric error alone reduces Future and Event means by 5.03 and 7.43 points. Using both errors in Eq. 3 improves geometric and Event predictions.

How does predicted change inform refinement? Observation-only Refinement (row 3) restores the update removed in row 2 and improves future prediction. EvoSee (row 8) uses predicted change,

Table 4. Ablations on SceneForesee. Columns report dimension means.

ID	Variant	Trajectory			Reasoning		
		Observed	Future	Combined	Entity	Event	Evidence
1	Geometry-only Assignment	27.82	21.89	24.03	44.83	41.27	43.15
2	w/o Evolution Refinement	29.22	18.27	22.11	48.59	42.71	50.12
3	Observation-only Refinement	29.10	22.60	25.93	49.36	44.52	52.41
4	w/o Continuation Selection	25.37	18.89	22.00	45.61	38.70	43.65
5	Single Hypothesis	27.86	21.88	24.16	46.98	40.67	46.47
6	Full Attention	28.49	24.52	25.97	49.90	44.13	50.07
7	w/o Visual Evidence	27.02	19.73	22.57	46.68	39.89	36.90
8	**EvoSee (Ours)**	**30.59**	**26.92**	**28.81**	**52.01**	**48.70**	**56.72**

adding 4.32, 4.18, and 4.31 points over row 3 in Future, Event, and Evidence, and 1.49 in Observed. The Observed gain supports refining entity estimates with predicted change.

Choice of continuation. Against EvoSee (row 8), Single Hypothesis (row 5) reduces Future and Event means, while removing Continuation Selection (row 4) gives the lowest Combined mean. Both candidate diversity and selection from observation contribute to entity continuation.

Direction of attention. Full Attention (row 6) reduces Future and Event means by 2.41 and 4.58 points relative to EvoSee (row 8). Directed attention in Eq. 1 conditions geometry on observation and Event semantics on observed and geometric states.

Contribution of observed evidence. Removing Visual Evidence (row 7 vs. row 8) reduces the Evidence mean by 19.82 points, the largest drop in this dimension, and lowers Future and Event means. Observed motion cues thus support both evidence recognition and entity continuation.

5.3 Further Analysis

Effect of entity count. Table 5 shows EvoSee gains in all six dimensions for every entity count group. With at least seven entities, its Entity and Event means exceed Qwen3.5-9B by 11.27 and 14.11 points, alongside improved future continuity. Evidence recognition also improves by 35.88 points. The gains persist as more identities and interactions must be distinguished. Its Event scores vary less than its trajectory scores across groups.

Table 5. Results by observed entity count.

Method	Count	Trajectory			Reasoning		
		Obs.	Fut.	All	Ent.	Evt.	Evd.
Qwen3.5-9B	1–3	28.35	21.90	25.26	58.06	28.30	50.72
	4–6	22.99	17.30	20.31	53.46	28.28	24.95
	≥ 7	9.36	7.49	8.47	19.68	31.39	14.81
EvoSee (Ours)	1–3	**44.06**	**38.96**	**41.57**	**71.50**	**52.96**	**65.12**
	4–6	**29.08**	**26.90**	**28.03**	**69.59**	**48.30**	**54.00**
	≥ 7	**13.68**	**11.18**	**12.49**	**30.94**	**45.50**	**50.70**

Effect of temporal extent. Table 6 compares equal observation and future durations, varying both the available motion history and the extent of prediction. EvoSee leads all six dimensions in every group. Its Future gain over Qwen3.5-9B is 10.18 points in the shortest group and 13.87 in the longest. Evidence gains exceed 25 points throughout, accompanying improved future localization across temporal scales. Combined scores also decline less than Qwen3.5-9B's.

Table 6. Results across temporal extents H.

Method	H (s)	Trajectory			Reasoning		
		Obs.	Fut.	All	Ent.	Evt.	Evd.
Qwen3.5-9B	1–3	21.97	17.29	19.76	41.35	27.20	31.89
	4–6	15.78	11.75	13.82	32.50	37.21	29.41
	≥ 7	13.64	7.95	10.78	31.15	38.76	28.15
EvoSee (Ours)	1–3	**30.46**	**27.47**	**29.04**	**53.45**	**49.02**	**57.26**
	4–6	**31.29**	**26.37**	**28.85**	**46.77**	**48.12**	**54.45**
	≥ 7	**30.81**	**21.82**	**26.32**	**46.17**	**46.25**	**54.46**

Variation across video sources. Table 7 shows distinct source profiles. OVT-B leads in trajectory means, TAO in Entity, and APIDIS in Event and Evidence. The source ranking thus depends on whether evaluation concerns spatial continuity, entity references, or semantic change. Appendix C.5 gives scores and sample counts for each source.

Table 7. EvoSee dimension means by source.

Source	Trajectory			Reasoning		
	Obs.	Fut.	All	Ent.	Evt.	Evd.
APIDIS	24.40	5.23	14.53	50.51	**65.63**	**81.25**
OVT-B	**32.27**	**29.45**	**30.88**	50.41	50.26	57.18
TAO	23.37	16.25	20.05	**59.92**	41.48	54.36

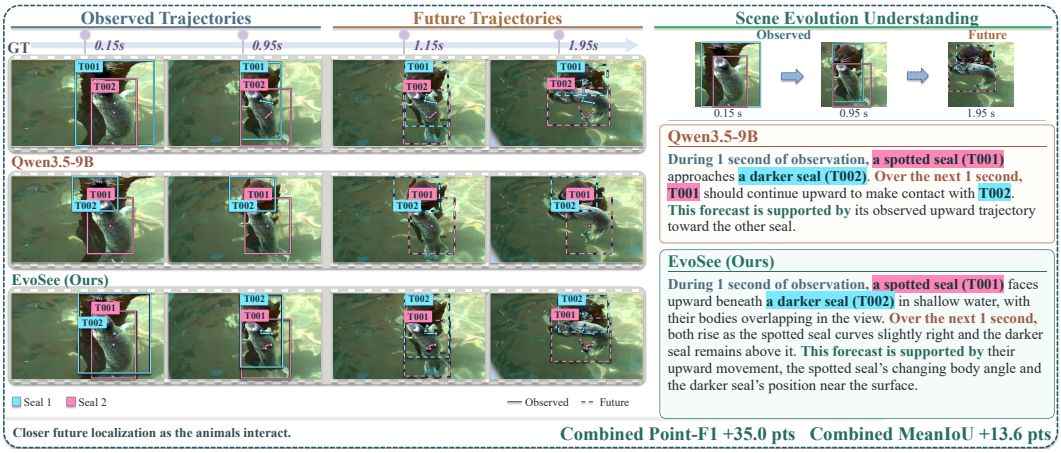


Figure 4. Seal interaction at matched timestamps. Identity colors distinguish the seals; solid and dashed paths show observed and future centers. EvoSee retains both through changing overlap.

Qualitative Scene Evolution. Figure 4 shows EvoSee following both seals through changing overlap and capturing the spotted seal's curved continuation. Its description relates this movement to the other seal and their observed upward motion, connecting the predicted interaction with observed evidence. Appendix D presents six further scenes.

5.4 Scene Evolution for 4D Reasoning

EvoSee's scene representation augments Qwen3.5-9B for VQA and Sa2VA-InternVL2.5-8B for grounding on Dyn-Bench (Huang et al., 2026; Yuan et al., 2026). Observation-only Refinement retains each base system and substitutes observed summaries, testing the value of predicted change for these tasks.

Table 8. Dyn-Bench grounding J&F. Red gains use the Sa2VA-InternVL2.5-8B base.

Method	Overall	Inter.	Object	Cam.
UniPixel-7B	65.23	66.00	71.10	58.60
Sa2VA-Qwen2.5-VL-7B	72.80	73.40	75.90	69.10
Sa2VA-InternVL2.5-8B	75.70	76.80	80.20	70.10
Observation-only Refinement	77.02	77.12	80.21	73.73
EvoSee (Ours)	**79.28** (+3.58)	**78.03** (+1.23)	**80.28** (+0.08)	**79.54** (+9.44)

Table 9. Dyn-Bench VQA accuracy. Red gains use the Qwen3.5-9B base. Overall aggregates all questions.

Method	Overall	Inter.	Object	Cam.
Qwen3-VL-8B	61.4	59.00	76.20	56.00
DynTrace	65.8	66.33	80.53	58.63
Qwen3.5-9B	63.5	59.20	79.35	58.87
Observation-only Refinement	66.7	67.40	80.62	60.83
EvoSee (Ours)	**68.3** (+4.8)	**69.80** (+10.60)	**80.85** (+1.50)	**61.50** (+2.63)

Dynamic grounding. Table 8 shows overall J&F gains of 3.58 points over Sa2VA-InternVL2.5-8B and 2.26 over Observation-only Refinement. Camera motion shows the largest gain over this control, 5.81 points, supporting predicted change as context for grounding under changing viewpoints.

Dynamic question answering. Table 9 shows overall gains of 4.8 points over Qwen3.5-9B and 1.6 over Observation-only Refinement, with a 2.40 point interaction gain over this control. Across VQA and grounding, predicted change most benefits interactions and changing viewpoints, while object motion scores remain close.

6 Conclusion

SceneForesee evaluates how persistent entities move and how their relations develop through continuity and change in the same scenes. EvoSee learns entity continuations through shared future states and joint assignment, then uses predicted change to update entity and scene representations. Results on SceneForesee and Dyn-Bench support scene evolution as both a prediction target and a dynamic representation for spatiotemporal reasoning. SceneForesee currently represents motion in the image plane and evaluates four Event categories. Extending this formulation to 3D trajectories and a broader Event vocabulary is a direction for future work.

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

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

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

Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. FIction: 4D future interaction prediction from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17613–17625. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ashutosh_FIction_4D_Future_Interaction_Prediction_from_Video_CVPR_2025_paper.html.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

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. URL https://arxiv.org/abs/2511.21631.

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

Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *European Conference on Computer Vision*, pp. 436–454. Springer, 2020.

Christophe De Vleeschouwer, Fan Chen, Damien Delannay, Christophe Parisot, Christophe Chaudy, Eric Martrou, and Andrea Cavallaro. Distributed video acquisition and annotation for sport-event summarization. In *NEM Summit*, 2008. URL https://ispgroup.gitlab.io/code/apidis/.

Patrick Dendorfer, Vladimir Yugay, Aljosa Osep, and Laura Leal-Taixé. Quo Vadis: Is trajectory forecasting the key towards long-term multi-object tracking? In *Advances in Neural Information Processing Systems*, volume 35, pp. 15657–15671, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/647dc4a76b3efdd676f50f32949299a8-Abstract-Conference.html.

Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, and Chen Chen. Motion-grounded video reasoning:

Understanding and perceiving motion at pixel level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8625–8636. IEEE, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Deng_Motion-Grounded_Video_Reasoning_Understanding_and_Perceiving_Motion_at_Pixel_Level_CVPR_2025_paper.html.

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, et al. 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.

Ruopeng Gao and Limin Wang. MeMOTR: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9901–9910, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Gao_MeMOTR_Long-Term_Memory-Augmented_Transformer_for_Multi-Object_Tracking_ICCV_2023_paper.html.

Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as ID prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27883–27893, 2025.

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 (CVPR)*, pp. 33446–33456, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Huang_Thinking_in_Dynamics_How_Multimodal_Large_Language_Models_Perceive_Track_CVPR_2026_paper.html.

Pere-Lluís Huguet Cabot and Roberto Navigli. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2370–2381, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.204. URL https://aclanthology.org/2021.findings-emnlp.204/.

Kehan Lan, Kaining Ying, and Henghui Ding. FeVOS: Foresight expression video object segmentation, 2026. URL https://arxiv.org/abs/2606.25585.

Yunhao Li, Qin Li, Hao Wang, Xue Ma, Jiali Yao, Shaohua Dong, Heng Fan, and Libo Zhang. Beyond MOT: Semantic multi-object tracking. In *European Conference on Computer Vision*, pp. 276–293. Springer, 2024.

Haiji Liang and Ruize Han. OVT-B: A new large-scale benchmark for open-vocabulary multi-object tracking. *Advances in Neural Information Processing Systems*, 37:14849–14863, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/1adeeac24ce6168e20bcee85645720e9-Abstract-Datasets_and_Benchmarks_Track.html.

Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander G Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5725–5734, 2019.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. UniPixel: Unified object referring and segmentation for pixel-level visual reasoning. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2509.18094.

Xuanchen Lu, Ang Cao, Chao Feng, and Andrew Owens. Generative point tracking and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28167–28178, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lu_Generative_Point_Tracking_and_Forecasting_CVPR_2026_paper.html.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. TrackFormer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8844–8854, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Meinhardt_TrackFormer_Multi-Object_Tracking_With_Transformers_CVPR_2022_paper.html.

Trong-Thuan Nguyen, Pha Nguyen, Jackson Cothren, Alper Yilmaz, and Khoa Luu. HyperGLM: HyperGraph for video scene graph generation and anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29150–29160, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Nguyen_HyperGLM_HyperGraph_for_Video_Scene_Graph_Generation_and_Anticipation_CVPR_2025_paper.html.

Sandro Papais, Letian Wang, Brian Cheong, and Steven L Waslander. ForeSight: Multi-view streaming joint object detection and trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25474–25484, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Papais_ForeSight_Multi-View_Streaming_Joint_Object_Detection_and_Trajectory_Forecasting_ICCV_2025_paper.html.

Francesca Pistilli, Simone Alberto Peirone, and Giuseppe Averta. Learning to evolve scenes: Reasoning about human activities with scene graphs, 2026. URL https://arxiv.org/abs/2607.02425.

Qwen Team. Qwen3.5: Towards native multimodal agents. https://qwen.ai/blog?id=qwen3.5, 2026.

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3591–3600, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Rupprecht_Learning_in_an_ICCV_2017_paper.html.

Mattia Segu, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Luc Van Gool, and Bernt Schiele. Samba: Synchronized set-of-sequences modeling for multiple object tracking. In *International Conference on Learning Representations*, pp. 30057–30070, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4a32a646254d2e37fc74a38d65796552-Abstract-Conference.html.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8nUgzuvskm.

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, 2025. URL https://arxiv.org/abs/2508.18265.

Zhaoyang Wei, Guoliang Wang, Guohua Gao, Yanchao Hao, Mingda Li, Wenchao Ding, Xi Chen, Shizhu He, and Xuehui Yu. Seeing is believing: Grounding long-video understanding in spatio-temporal visual evidence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10584–10592, 2026. doi: 10.1609/aaai.v40i13.38031. URL https://doi.org/10.1609/aaai.v40i13.38031.

Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, and Jianbing Shen. Referring multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14633–14642, 2023.

En Yu, Songtao Liu, Zhuoling Li, Jinrong Yang, Zeming Li, Shoudong Han, and Wenbing Tao. Generalizing multiple object tracking to unseen domains by introducing natural language representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3304–3312, 2023. doi: 10.1609/aaai.v37i3.25437. URL https://ojs.aaai.org/index.php/AAAI/article/view/25437.

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/2501.04001v4.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09675.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=EElFGvt39K.

APPENDIX

A BENCHMARK CONSTRUCTION AND EVALUATION

A.1 TEMPORAL SAMPLING AND ANNOTATIONS

The observation boundary separates the video available for inference from the future continuation used for training and evaluation. For video V and a window beginning at s, duration H defines the boundary t_b and the paired observed and future intervals.

$$t_b = s + H, \qquad V^{\mathrm{obs}} = V[s, s+H], \qquad V^{\mathrm{fut}} = V(s+H, s+2H], \qquad H \in \{1, \ldots, 10\}\ \mathrm{s}. \tag{A.1}$$

Videos are normalized to 5 frames per second, and candidate windows advance by $H/4$. Each observed interval supplies four uniformly sampled keyframes. The model receives the observed video and keyframes, with H specified by the task context. Annotations from this interval supervise entity states and Visual Evidence, while future annotations supervise trajectories and Events. Each sample calls for one scene evolution prediction comprising observed and future trajectories, Events, Visual Evidence, and a scene description.

Entity states. Each entity annotation includes an identity within the sample, timestamped boxes in the image plane, visibility, and view. Entities that continue across the observation boundary retain this identity in their future persistence and state annotations. Every sample contains at least one continuing entity. Counting shared identities over all temporal pairs gives 1.05M entity correspondences. Counting observed and future trajectory records separately gives 2.14M records. Both totals accumulate over samples.

Events and Visual Evidence. The four Event categories, relation change, approach, occlusion, and leave, summarize entity configurations using observed motion and states at the transition to prediction or the future endpoint. Each Event records its participating identities and the requested future interval. Its paired Visual Evidence records an individual trajectory trend or pairwise relation change, together with the participating identities, observed keyframes, and an observed interval. The paired evidence annotations record associated motion in observation and supervise evidence localization and category prediction from observed frames. All 272.49K Events have paired evidence, and each of the 144.57K samples has a scene description.

A.2 ANNOTATION CONSTRUCTION AND DATA SPLITS

SceneForesee constructs annotations around identities shared across each temporal pair. These identities associate timestamped boxes and visibility with continuing entities, whose trajectories determine the Event annotations. The same construction criteria apply to APIDIS, OVT-B, and TAO. Together, these sources cover sports interactions, open category entities, and varied natural motion (De Vleeschouwer et al., 2008; Liang & Han, 2024; Dave et al., 2020).

Event assignment describes relative motion, spatial overlap, and position in the field of view under these shared identities. The categories provide discrete targets for predicting the configurations represented by continuous trajectories. Approach requires the distance between entity centers to decrease by more than 10% during observation and remain at most 1.10 times the observation endpoint distance at the future endpoint.

Occlusion uses box overlap at the temporal transition from observation to prediction, with pairwise IoU above 0.05 at the last observed or first future state. This criterion applies throughout the image, including its interior. Leave uses position relative to the image edge at the future endpoint, with the box center within the outer 10% of either image axis. Relation change records the remaining configurations when none of these rules applies and is assigned to the first continuing identity in source order. The geometric criteria define Event supervision for the same entities and temporal pairs used in trajectory prediction.

Pairwise Events are enumerated in source identity order, with approach preceding occlusion for each pair, followed by individual leave Events. Retaining the first three records and their paired evidence defines the subset of configurations covered by Event evaluation. A sample is retained when

Table A1. Scale and Event distribution of SceneForesee. Entity correspondences count identities shared across paired windows within each sample. Trajectory records count observed and future windows separately. Each Event has one associated Visual Evidence annotation. Val denotes validation.

Statistic	Count
Train / Val / Test	116,681 / 14,259 / 13,626
Total samples	144,566
Source videos	2,876
Entity correspondences	1,046,165
Observed trajectory records	1,098,141
Future trajectory records	1,046,165
Entity trajectory records	2,144,306
Event and evidence pairs	272,487
Scene descriptions	144,566
Event categories	4
Visual Evidence categories	2

Source	Samples	Relation change	Approach	Occlusion	Leave	Events
APIDIS	317	41	441	311	35	828
TAO	26,644	7,910	12,879	19,983	6,188	46,960
OVT-B	117,605	43,100	58,499	110,039	13,061	224,699
All sources	144,566	51,051	71,819	130,333	19,284	272,487

it contains an entity annotated across the observation boundary, an Event assigned to the requested continuation, and its evidence in observation.

Visual Evidence associates each Event with observed motion. Its category follows the Event's participant structure, assigning relation change to pairwise Events and trajectory trend to individual Events. The evidence description uses the first observed motion segment containing a participating identity. If no segment shares an Event participant, it uses the first available observed segment as scene context. If no segment is available, the observed interval supplies the description and time span. The resulting evidence record carries the Event participants and observed keyframes, with its temporal extent confined to observation.

Construction included LLM assessment and manual inspection of the generated annotations. The train, validation, and test sets contain 116,681, 14,259, and 13,626 samples, respectively. Partitions use source videos, with disjoint chronological episodes defining the APIDIS partitions, so overlapping windows remain together. Each source contributes the windows whose continuing entities and Events meet the common construction criteria. Table A1 reports the resulting scale, counting units, and Event distribution.

Table A1 shows the distribution of scene evolution in the retained annotations. Occlusion and approach are the most frequent Events, capturing spatial overlap and relative motion, while relation change and leave cover residual configurations and proximity to the image edge. Separate counts for observed trajectories and continuing identities distinguish entities present in observation from those followed into the future.

The retained scenes also vary in entity count and Event composition. Observed windows contain a mean of 7.60 entities and a median of 3, ranging from 1 to 89. Multiple Events occur in 47.72% of samples and multiple Event categories in 27.34%. Two entities participate in 74.19% of Events. These distributions support analysis of continuity and change across scene densities.

A.3 Entity Continuity

Let P and G denote all predicted and annotated identities in one sample. Trajectory overlap in the observed interval determines matched subsets $P_{\mathrm{m}} \subseteq P$ and $G_{\mathrm{m}} \subseteq G$, with a bijection $\pi : P_{\mathrm{m}} \to G_{\mathrm{m}}$. This partial matching remains fixed for future evaluation. Unmatched predictions remain in the prediction counts, and unmatched annotations remain in the reference counts. For interval W, let

$\mathcal{A}_W$ contain point pairs aligned by π, frame, and view, and define

$$m_W = \sum_{(g,p)\in\mathcal{A}_W} \mathbf{1}[\mathrm{IoU}(g,p) \geq 0.5]. \tag{A.2}$$

Let N_W^{gt} and N_W^{pred} count annotated and predicted points in interval W. Point-Recall and Point-F1 normalize the matched count by these totals.

$$\mathrm{PointRecall}_W = \frac{m_W}{N_W^{\mathrm{gt}}}, \qquad \mathrm{PointF1}_W = \frac{2m_W}{N_W^{\mathrm{gt}} + N_W^{\mathrm{pred}}}. \tag{A.3}$$

MeanIoU averages spatial overlap over the same annotated points.

$$\mathrm{MeanIoU}_W = \frac{1}{N_W^{\mathrm{gt}}} \sum_{(g,p)\in\mathcal{A}_W} \mathrm{IoU}(g,p). \tag{A.4}$$

Unmatched annotated points contribute zero to MeanIoU. The observed interval measures entity state estimation, and the future interval measures trajectory prediction under the same π. Combined scores are formed within each sample. For any of the three interval scores M, the paired result is

$$M_{\mathrm{comb}} = \frac{N_{\mathrm{obs}}^{\mathrm{gt}} M_{\mathrm{obs}} + N_{\mathrm{fut}}^{\mathrm{gt}} M_{\mathrm{fut}}}{N_{\mathrm{obs}}^{\mathrm{gt}} + N_{\mathrm{fut}}^{\mathrm{gt}}}. \tag{A.5}$$

The weights are the annotated point counts of that sample's observed and future intervals. The reported score is the arithmetic mean over test samples $\mathcal{S}$.

$$\overline{M}_W = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} M_{s,W}, \qquad W \in \{\mathrm{obs}, \mathrm{fut}, \mathrm{comb}\}. \tag{A.6}$$

Thus annotated point counts weight intervals within a sample, while samples receive equal weight in the final result. A score with a zero denominator is assigned zero.

A.4 Scene Evolution Reasoning

Scene Evolution Reasoning evaluates entity references and the categories of Events and Visual Evidence. Entity scoring extracts the identities referenced in predicted and reference descriptions, maps predictions through π, and compares the resulting sets. Event and Visual Evidence scoring compares category multisets within each scene, retaining repeated occurrences. For dimension d in scene s, let $\mathcal{P}_s^d$ and $\mathcal{G}_s^d$ denote predicted and annotated collections, and t_s^d their intersection count, with multiplicity for Event and Visual Evidence. Global precision and recall pool these counts across scenes.

$$P_g^d = \frac{\sum_s t_s^d}{\sum_s |\mathcal{P}_s^d|}, \qquad R_g^d = \frac{\sum_s t_s^d}{\sum_s |\mathcal{G}_s^d|}. \tag{A.7}$$

Global F1 summarizes matching counts across the test set.

$$\mathrm{GlobalF1}^d = \frac{2\sum_s t_s^d}{\sum_s |\mathcal{P}_s^d| + \sum_s |\mathcal{G}_s^d|}. \tag{A.8}$$

Scene-F1 assigns equal weight to each scene.

$$\mathrm{SceneF1}^d = \frac{1}{|\mathcal{S}|} \sum_{s\in\mathcal{S}} \frac{2t_s^d}{|\mathcal{P}_s^d| + |\mathcal{G}_s^d|}. \tag{A.9}$$

GPR is the geometric mean of global precision and recall.

$$\mathrm{GPR}^d = \sqrt{P_g^d R_g^d}. \tag{A.10}$$

A zero denominator gives zero. Entity, Event, and Visual Evidence collections are matched independently, measuring identity references and category multiplicities. Reported scores multiply the ratios above by 100.

Visual Evidence scores compare individual trajectory trend and pairwise relation change categories from the observed context defined in Appendix A.2. They measure category recognition, without testing temporal localization or the support a segment provides for a particular Event. Complete Event matching separately tests the predicted category and its participants.

Exact matching of complete Events. Complete Event micro-F1 follows the structured matching principle used in relation extraction (Huguet Cabot & Navigli, 2021). A correct Event must match its category and complete participant record. The score is

$$F_{1,\mathrm{micro}} = 100 \times \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{A.11}$$

TP counts matched Events, FP unmatched predictions, and FN unmatched references, pooled across test scenes.

Complete Event records contain a category and all participating identities. These identities are part of the model output before evaluation and refer to its predicted entities. Participant roles or ordering follow the annotation schema. Evaluation maps the predicted identities through the observation correspondence π used for trajectory evaluation. An Event with any unmapped or incorrect participant cannot match a reference and remains in the prediction count. Within each scene, maximum one-to-one exact matching retains multiplicities. Each predicted and annotated record can be used at most once. Unmatched predictions and references contribute FP and FN, respectively. Counts are pooled over all test scenes before computing micro-F1; a zero denominator gives zero. The matching key comprises the Event category and all participant identities, with no additional temporal localization condition. Thus the score tests a category and its participants jointly, while Event category scores compare category occurrences alone.

A.5 Dimension Means and Group Aggregation

The main ablation and difficulty tables summarize six dimensions. Observed, Future, and Combined each average Point-F1, Point-Recall, and MeanIoU for the corresponding interval. Entity, Event, and Visual Evidence each average Global F1, Scene-F1, and GPR. These arithmetic means give equal weight to the three metrics within a dimension. Table 4 averages the corresponding scores reported in Section C and rounds the resulting means to two decimal places. Differences between dimension means are calculated before this final rounding. Event micro-F1 with exact matching is reported separately and does not enter these means.

For a sample group, trajectory scores and Scene-F1 are averaged across its samples. Global F1 and GPR are calculated from the group's pooled true positive, prediction, and reference counts. Combining groups therefore requires pooling these counts before computing the global semantic metrics. Overall results use the complete test population, and source results apply the same definitions to each source's retained samples. The dimension mean is taken after the three constituent scores have been computed for the group.

A.6 Scene Description Similarity

BERTScore F1 with RoBERTa-large evaluates similarity between the predicted and reference scene descriptions (Zhang et al., 2020). The complete description covers observed motion, predicted development, and evidence. We also score the present, future, and evidence parts separately to examine which content contributes to the overall result. Appendix C.6 reports these component scores.

B Method Details

EvoSee encodes observed entity states and Visual Evidence, then uses these representations to learn candidate continuations. Each candidate shares future states between geometry and Event semantics, and joint assignment determines which candidate receives future supervision. A selector learns this assignment from observation. The predicted change then enters entity refinement, consistency ranking, and scene representation.

B.1 Entity Visual Evidence Encoding

Visual features from the observed video and its keyframes form memory X_{obs}. A learned query q_i attends to this memory to obtain entity state a_i at the observation boundary. The same query attends

to frame features X_t at each valid observed time $t \in \mathcal{O}_i$, producing temporal entity state h_i^t. These two uses of cross attention are

$$a_i = \mathrm{CrossAttn}(q_i, X_{\mathrm{obs}}, X_{\mathrm{obs}}), \qquad h_i^t = \mathrm{CrossAttn}(q_i, X_t, X_t). \tag{B.1}$$

A learned map ψ_e assigns each temporal state a relevance logit r_i^t. Normalizing its sigmoid response over observed times gives weights α_i^t for aggregating the states into evidence e_i.

$$\alpha_i^t = \frac{\sigma(r_i^t)}{\sum_{\ell \in \mathcal{O}_i} \sigma(r_i^\ell)}, \qquad e_i = \sum_{t \in \mathcal{O}_i} \alpha_i^t h_i^t. \tag{B.2}$$

Here $r_i^t = \psi_e(h_i^t)$, and σ is the sigmoid function. Evidence intervals supervise temporal relevance, while evidence categories supervise the aggregated features. These two objectives train the representation to identify the observed motion associated with each entity.

B.2 Geometry Event Alignment Reasoning

GEAR uses the entity and evidence representations to initialize its geometric and semantic states. Their sum $u_i^a = a_i + e_i$ is projected into $u_i^g, u_i^s \in \mathbb{R}^d$. Stacking these states as observation, geometry, and semantics gives $U_i \in \mathbb{R}^{3\times d}$. For attention head h of width d_h, layer normalization followed by learned projections gives queries, keys, and values.

$$Q_i^h = \mathrm{LN}(U_i) W_Q^h, \qquad K_i^h = \mathrm{LN}(U_i) W_K^h, \qquad V_i^h = \mathrm{LN}(U_i) W_V^h. \tag{B.3}$$

The mask in Eq. 1 permits a receiving state to attend to itself and the states preceding it in this order. Thus $\mathcal{M}_{rs} = 0$ for $s \le r$ and $\mathcal{M}_{rs} = -\infty$ for $s > r$. The resulting head outputs A_i^h return to the state space through concatenation, learned projection W_O, and a residual update.

$$\widetilde{U}_i = U_i + \mathrm{Concat}_h \left[\mathrm{softmax}\left(\frac{Q_i^h (K_i^h)^\top}{\sqrt{d_h}} + \mathcal{M} \right) V_i^h \right] W_O. \tag{B.4}$$

The observed state receives information only from observation. The geometric state additionally uses this observed context, and the semantic state uses both observation and geometry.

The updated states use separate transformation weights to preserve these dependencies. With normalized states $Z_i = \mathrm{LN}(\widetilde{U}_i)$, learned matrix W_R and bias b_R assign weights over E feedforward transformations; $\mathbf{1}_3$ broadcasts the bias across the three states.

$$\Pi_i = \mathrm{softmax}\big(Z_i W_R + \mathbf{1}_3 b_R^\top \big). \tag{B.5}$$

Softmax acts over the E transformations independently for each state. Writing $\Pi_{i,:,m}$ for the weights of transformation FFN_m, the update is

$$U_i' = \widetilde{U}_i + \sum_{m=1}^{E} \mathrm{diag}(\Pi_{i,:,m})\ \mathrm{FFN}_m(Z_i). \tag{B.6}$$

Each transformation acts separately on the three states, preserving the dependencies defined by the attention mask. The refined geometric and semantic states jointly condition the future sequence from which candidate continuations are learned.

For L future steps, this sequence is $f_i \in \mathbb{R}^{L\times d}$. Each of K candidate continuations transforms it with embedding $d_k \in \mathbb{R}^d$ and transition ϕ_k, as in Eq. 2. The vector $\mathbf{1}_L$ broadcasts the embedding across future time steps. Both decoders use the resulting state f_i^k. The Event decoder also receives embedding c, which represents task context in the entity feature space. One candidate index therefore determines future boxes, visibility, persistence, view, and Event semantics.

Joint assignment compares each candidate with future annotations using box error and Event classification error. Let $\mathcal{F}_i$ contain valid annotated future steps, $\mathbf{b}_{i,t}^k \in \mathbb{R}^4$ be a predicted box, and $\mathbf{b}_{i,t}^\star$ its reference. The geometric error is

$$\ell_{\mathrm{geo}}^{i,k} = \frac{1}{4|\mathcal{F}_i|} \sum_{t \in \mathcal{F}_i} \left\| \mathbf{b}_{i,t}^k - \mathbf{b}_{i,t}^\star \right\|_1. \tag{B.7}$$

The Event classification error uses logits averaged over the L future steps, $\bar{s}^k_{i,j} = L^{-1}\sum_{t=1}^{L} s^k_{i,t,j}$. For each reference Event of category j, every recorded participant i receives $y_{i,j} = 1$, so both identities in a pairwise Event carry its category target. These labels over J output categories define the semantic error.

$$\ell^{i,k}_{\text{sem}} = -\frac{1}{J}\sum_{j=1}^{J}\left[y_{i,j}\log\sigma(\bar{s}^k_{i,j}) + (1-y_{i,j})\log\left(1-\sigma(\bar{s}^k_{i,j})\right)\right]. \tag{B.8}$$

The assigned candidate minimizes $\ell^{i,k}_{\text{geo}} + 0.20\ell^{i,k}_{\text{sem}}$, and its index $\widehat{k}_i$ is shared across all future supervision terms. This index also supplies the target for selecting a continuation from observation.

The selector R combines entity state a_i, evidence e_i, and embedding o_i of the entity's observed trajectory history. Its K output logits define candidate probabilities $\boldsymbol{\rho}_i$ and the most probable candidate k^*_i.

$$\boldsymbol{\rho}_i = \text{softmax}(R(a_i + e_i + o_i)), \qquad k^*_i \in \arg\max_k \rho^k_i. \tag{B.9}$$

The selector learns the assigned candidate by minimizing

$$\mathcal{L}_{\text{prior}} = -\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\log\rho^{\widehat{k}_i}_i, \tag{B.10}$$

where $\mathcal{I}$ contains matched entities with valid future annotations. Future annotations determine the training target, while the selection distribution is inferred from observation.

Trajectory diversity is encouraged through the separation of candidate endpoints. For final predicted center $\mathbf{p}^k_i$, the regularizer is

$$\mathcal{L}_{\text{diversity}} = \frac{1}{|\mathcal{I}|K(K-1)}\sum_{i\in\mathcal{I}}\sum_{k\neq\ell}\left[m - \|\mathbf{p}^k_i - \mathbf{p}^\ell_i\|_1\right]_+, \qquad m = 0.05. \tag{B.11}$$

Here $[x]_+ = \max(0, x)$, and coordinates are normalized to the image extent. This expression applies for $K > 1$ and nonempty $\mathcal{I}$. With $K = 1$, the diversity term is zero. The regularizer acts on endpoint separation while candidate assignment uses the complete valid trajectory and its Event prediction.

The candidate probabilities also define a weighted future state.

$$\widetilde{f}_i = \sum_{k=1}^{K}\rho^k_i f^k_i. \tag{B.12}$$

During training, ECD uses weighted predictions formed from the boxes and Event logits decoded for each candidate, as defined below. Inference uses $k^*_i \in \arg\max_k \rho^k_i$ and $\widetilde{f}_i = f^{k^*_i}_i$. The same index determines future boxes, visibility, persistence, and Event predictions, preserving one geometric and semantic continuation for each entity. The selected Event scores express the categories associated with that entity. Complete Event records additionally specify all participants and are evaluated by joint category and identity matching in Appendix A. With $K = 1$, its sole future state directly defines the continuation.

B.3 Evolution Conditioned Decoding

ECD summarizes geometric and Event predictions as features for entity refinement. At inference, it uses the selected continuation, whose box at future time t is $\mathbf{b}^{k^*_i}_{i,t} \in \mathbb{R}^4$ and whose Event logits are $s^{k^*_i}_{i,t}$. Temporal averaging and projection give the geometric and semantic summaries.

$$\mathbf{g}_i = P_g\left(\frac{1}{L}\sum_{t=1}^{L}\mathbf{b}^{k^*_i}_{i,t}\right), \qquad \mathbf{s}_i = P_s\left(\sigma\left(\frac{1}{L}\sum_{t=1}^{L}s^{k^*_i}_{i,t}\right)\right). \tag{B.13}$$

Here P_g and P_s project into $\mathbb{R}^d$. During training, the corresponding boxes and logits are weighted by selection probabilities.

$$\mathbf{b}_{i,t} = \sum_k \rho^k_i \mathbf{b}^k_{i,t}, \qquad s_{i,t} = \sum_k \rho^k_i s^k_{i,t}. \tag{B.14}$$

These weighted predictions undergo the same temporal averaging and projection as the selected predictions at inference. Equation 4 combines the resulting summaries with entity state and observed evidence into evolution representation $\mathbf{m}_i$.

The map Γ in Eq. 4 uses entity state and evolution representation to compute a vector gate.

$$\boldsymbol{\gamma}_i = \sigma(W_\gamma[a_i; \mathbf{m}_i] + b_\gamma)\,, \qquad \boldsymbol{\gamma}_i \in (0,1)^d. \tag{B.15}$$

Normalization and projection give $P_f(\mathbf{m}_i) = W_f\,\mathrm{LN}(\mathbf{m}_i) + b_f$. The update is

$$a_i^+ = a_i + \boldsymbol{\gamma}_i \odot P_f(\mathbf{m}_i). \tag{B.16}$$

The gate weights the contribution of predicted change to each feature dimension of the entity state. The updated state supports observed trajectory estimation and produces entity confidence ℓ_i^{obj}.

Entity retention uses feature agreement alongside this confidence. Six ordered relation pairs specify the features compared by Consistency.

$$\mathcal{R}_i = \big((e_i, a_i), (a_i, \mathbf{g}_i), (\mathbf{g}_i, \mathbf{s}_i), (\mathbf{s}_i, e_i), (\mathbf{g}_i, \mathbf{m}_i), (\mathbf{s}_i, c)\big). \tag{B.17}$$

The pairs connect evidence to entity state, entity state to geometry, geometry to semantics, and semantics to evidence. They also compare geometry with the evolution representation and semantics with task context. Prediction losses supervise the predicted content, while relation supervision distinguishes queries matched to annotated entities from unmatched queries.

For relation r, let $(u_{i,r}, v_{i,r})$ denote the corresponding pair in $\mathcal{R}_i$. Its cosine distance is

$$b_{i,r} = 1 - \frac{u_{i,r}^\top v_{i,r}}{\max(\|u_{i,r}\|_2, \epsilon)\max(\|v_{i,r}\|_2, \epsilon)}, \qquad \epsilon = 10^{-6}. \tag{B.18}$$

The stabilizer ϵ keeps the distance defined for zero norm features. Averaging the six distances gives B_i.

$$B_i = \frac{1}{6}\sum_{r=1}^{6} b_{i,r}. \tag{B.19}$$

A smaller B_i indicates closer feature agreement. Its geometric and semantic inputs come from weighted predictions during training and the selected candidate at inference. The same mean distance contributes to relation supervision and entity retention.

Relation supervision. Annotation matching divides entity queries into matched set $\mathcal{I}_{\mathrm{m}}$ and unmatched set $\mathcal{I}_{\mathrm{u}}$. Let $\bar{B}_{\mathrm{m}}$ and $\bar{B}_{\mathrm{u}}$ be their respective mean distances. The ranking loss is

$$\mathcal{L}_{\mathrm{rank}} = \big[m + \bar{B}_{\mathrm{m}} - \bar{B}_{\mathrm{u}}\big]_+, \qquad m = 0.25. \tag{B.20}$$

Training applies this loss to the aggregate distance and to each individual relation. It favors lower distances for entities associated with annotated trajectories. The term is evaluated when both sets are nonempty and is zero otherwise. Geometric, semantic, and evidence supervision jointly train the representations on which the relation distances depend.

Supervision and continuation learning. Observed annotations supervise entity estimation, while the evidence intervals and categories supervise temporal relevance and the evidence representation. Future box and Event losses determine the shared candidate assignment and supervise its geometric and semantic outputs. The assigned index trains the selector through $\mathcal{L}_{\mathrm{prior}}$, and $\mathcal{L}_{\mathrm{diversity}}$ encourages separation among candidate endpoints. Relation ranking encourages lower feature distances for matched queries than for unmatched queries. At inference, observed video and task context determine candidate selection, entity retention, and scene representation.

Entity retention combines entity confidence with the mean relation distance.

$$\kappa_i = \ell_i^{\mathrm{obj}} - \lambda_{\mathrm{cons}} B_i. \tag{B.21}$$

The entities with the highest scores form the retained set $\mathcal{I}^*$. Smaller relation distances raise an entity's rank at fixed confidence, linking the final entity set to both observed support and predicted continuation.

The evolution representations also contribute to scene reasoning. Together with entity states and evidence, they are averaged over all N_q queries.

$$\bar{a} = \frac{1}{N_q}\sum_i a_i, \qquad \bar{e} = \frac{1}{N_q}\sum_i e_i, \qquad \bar{m} = \frac{1}{N_q}\sum_i \mathbf{m}_i. \tag{B.22}$$

Concatenating these averages with task context c and projecting them gives the scene representation.

$$h_{\text{scene}} = P_{\text{scene}}([\bar{a}; \bar{e}; \bar{m}; c])\,. \tag{B.23}$$

The scene representation combines observed states with predicted geometry and Event semantics across all queries. Alongside this representation, the retained set $\mathcal{I}^*$ supplies structured trajectories and Events with one continuation per entity, while a_i^+ supplies observed trajectory estimates and entity confidence. SceneForesee evaluates these spatial and semantic predictions directly. Dyn-Bench evaluates the scene representation in question answering and grounding, testing the value of learned scene evolution for 4D reasoning.

C Experimental Protocols and Additional Results

The additional results examine entity continuity and recognition of change across model variants, entity counts, temporal extents, and video sources. Each of the six evaluation dimensions contains three metrics. P-F1 and P-R denote Point-F1@0.5 and Point-Recall@0.5. GF1, SF1, and GPR denote Global F1, Scene-F1, and the geometric mean of global precision and recall. Tables report scores as percentages and place group size N before the metrics. Bold indicates the leading value within each comparison.

The main ablation and difficulty tables summarize these dimensions using Obs., Fut., All, Ent., Evt., and Evd. for Observed, Future, Combined, Entity, Event, and Visual Evidence. Each dimension mean averages its three constituent metrics. Complete Event micro-F1 and description BERTScore F1 are reported separately.

C.1 Comparison Protocols

SceneForesee comparisons. SceneForesee comparisons use the same test samples, observed video intervals, temporal extents, and metrics, with one scene prediction per method and sample. The TrackFormer-style, MeMOTR-style, LTrack-style, and SambaMOTR adaptations (Meinhardt et al., 2022; Gao & Wang, 2023; Yu et al., 2023; Segu et al., 2025) implement the cited tracking principles and are trained for this task rather than using the released checkpoints. Their outputs include trajectories and categories for both evaluations. Qwen3.5-9B (Qwen Team, 2026) and EvoSee predict trajectories and scene descriptions from observation. The remaining multimodal and video language models use shared trajectories from the TrackFormer-style adaptation as geometric context, including LLaVA-Video (Zhang et al., 2025), LLaVA-OneVision-1.5 (An et al., 2025), and LLaVA-OneVision-2 (An et al., 2026). For each method, matching its observed trajectories to annotations establishes the correspondence used for future continuity and Entity scoring.

Ablation variants. Nine ablations use the full model's training and evaluation conditions. Single Hypothesis sets $K = 1$ to test the role of candidate continuations. Removing Consistency, Continuation Selection, Evolution Refinement, Entity Retention, or Visual Evidence tests feature agreement, selection from observation, refinement using predicted change, retention of continuing entities, or observed evidence, respectively. Every variant produces one scene prediction.

Observation-only Refinement retains the projection and gated residual update in ECD. It replaces the predicted geometric and Event inputs to the evolution representation with observed entity, motion, and evidence summaries of the same input dimensions. The resulting representation enters both the entity update and h_{scene}, replacing predicted change in both paths shown in Figure 3. The candidate continuations and their prediction supervision remain in the model. This comparison tests the information used for refinement while retaining the update structure; w/o Evolution Refinement removes the update itself.

Geometry-only Assignment chooses $\widehat{k}_i$ using geometric error while retaining Event classification supervision with its original loss weight. Both geometric and Event predictions use this assigned index

for supervision, and the same index trains the selector. Full Attention instead removes the directional constraint by setting every entry of $\mathcal{M}$ to zero, allowing observation, geometry, and semantics to attend to one another. Both variants retain EvoSee's candidate continuations and learning conditions.

Dyn-Bench transfer. On Dyn-Bench, scene representation h_{scene} carries predicted scene evolution into 4D reasoning. It augments Qwen3.5-9B for question answering and the Sa2VA-InternVL2.5-8B mask decoder for grounding. Each gain is measured against the corresponding base system. Additional comparisons include Qwen3-VL-8B and DynTrace for question answering, and UniPixel and Sa2VA for grounding (Huang et al., 2026; Gao et al., 2026; Liu et al., 2025; Yuan et al., 2026). DynTrace uses Qwen3-VL-8B, and the published Qwen3-VL and grounding reference results follow Dyn-Bench.

Inter., Object, and Cam. denote interaction, object motion, and camera motion. VQA Overall is accuracy across all questions, while each dimension averages the accuracies of its three constituent tasks. Grounding Overall averages the three dimension J&F scores. Gains use the corresponding overall scores of the matched base systems. Observation-only Refinement retains these respective base systems and uses the observed summaries defined above. Comparing this control with EvoSee tests the value of predicted change in the scene representation for VQA and grounding. These tasks evaluate 4D reasoning downstream of SceneForesee, whose trajectories describe the image plane over time.

C.2 Complete Ablation Results

Table A2 examines how continuation learning and refinement affect entity continuity. EvoSee leads every trajectory metric. Geometry-only Assignment retains Event supervision but reduces all future metrics, supporting the contribution of semantic error to joint assignment. Full Attention also lowers every future metric relative to directed attention.

Removing Evolution Refinement produces the largest loss in each future metric and in Combined Point-F1; Future Point-F1 falls by 8.44 points. Restoring the update with observed information recovers 4.62 points, and using predicted change adds a further 3.82 points. This additional gain extends to all three future metrics. EvoSee also improves the Observed mean to 30.59, exceeding Observation-only Refinement at 29.10 and the variant without refinement at 29.22 by 1.49 and 1.37 points. These gains support using the predicted continuation to refine entity estimates within the observed interval. Continuation Selection affects a different aspect of the prediction, with its removal producing the largest declines in Combined Point-Recall and MeanIoU. Single Hypothesis also lowers future accuracy. These comparisons distinguish the contributions of predicted change, selection from observation, and candidate continuations.

Table A2. Complete ablation results for Entity Continuity.

	Observed			Future			Combined		
Variant	**P-F1**	**P-R**	**mIoU**	**P-F1**	**P-R**	**mIoU**	**P-F1**	**P-R**	**mIoU**
Qwen3.5-9B	20.55	19.72	21.58	15.49	14.78	17.55	18.16	17.35	19.65
Geometry-only Assignment	26.93	28.48	28.05	22.39	21.24	22.05	24.04	23.09	24.96
w/o Evolution Refinement	29.77	28.67	29.23	18.09	17.03	19.68	21.44	21.63	23.25
Observation-only Refinement	29.31	28.96	29.02	22.71	22.36	22.74	25.74	25.88	26.17
w/o Continuation Selection	25.49	25.46	25.17	18.74	17.24	20.69	22.63	21.22	22.16
Single Hypothesis	28.65	26.56	28.38	21.39	21.68	22.58	24.55	22.85	25.08
Full Attention	28.00	29.51	27.95	24.32	24.03	25.20	25.07	26.36	26.49
w/o Visual Evidence	26.10	27.97	26.98	18.88	19.78	20.54	22.85	21.46	23.39
w/o Consistency	27.43	24.73	27.36	22.01	22.14	21.80	24.01	22.26	24.23
w/o Entity Retention	28.58	28.13	27.77	22.23	22.34	23.66	22.96	24.82	24.23
EvoSee (Ours)	**30.54**	**31.10**	**30.13**	**26.53**	**27.47**	**26.77**	**28.61**	**29.32**	**28.51**

Table A3 shows that the same ablations also affect Event prediction. Geometry-only Assignment lowers all Event scores, including Global F1 by 6.16 points. The concurrent trajectory and Event declines support using both errors in joint assignment. Full Attention also reduces all three Event metrics.

Removing Entity Retention produces the largest decline in all Entity metrics. Removing Consistency most affects Event Global F1 and GPR, whereas removing Continuation Selection most affects Event Scene-F1. Removing Visual Evidence produces the largest losses in evidence category recognition, including 19.42 points in Global F1, and also lowers Event scores. Observation-only Refinement improves Event and Visual Evidence scores over removing the update. Using predicted change further improves their Global F1 scores by 4.23 and 4.14 points, respectively. EvoSee leads all nine semantic metrics.

Table A3. Complete ablation results for Scene Evolution Reasoning.

	Entity			Event			Visual Evidence		
Variant	**GF1**	**SF1**	**GPR**	**GF1**	**SF1**	**GPR**	**GF1**	**SF1**	**GPR**
Qwen3.5-9B	37.45	40.73	40.14	29.85	28.20	29.87	29.82	34.23	29.84
Geometry-only Assignment	41.38	48.42	44.69	40.93	40.26	42.63	41.84	43.84	43.78
w/o Evolution Refinement	45.50	53.61	46.67	40.63	42.10	45.40	48.43	50.31	51.62
Observation-only Refinement	46.82	54.11	47.15	42.86	45.03	45.67	50.31	53.18	53.74
w/o Continuation Selection	43.46	49.00	44.36	38.54	36.96	40.59	41.31	45.43	44.20
Single Hypothesis	43.88	49.88	47.18	40.15	38.79	43.06	45.17	47.41	46.83
Full Attention	46.33	54.97	48.39	42.92	45.59	43.87	49.16	49.81	51.24
w/o Visual Evidence	43.31	50.33	46.39	38.74	38.41	42.53	35.03	38.65	37.01
w/o Consistency	44.45	50.60	45.46	36.01	37.87	37.83	46.32	50.22	48.32
w/o Entity Retention	39.93	44.15	41.74	40.33	43.32	41.49	46.23	50.43	49.35
EvoSee (Ours)	**49.08**	**57.33**	**49.62**	**47.09**	**48.91**	**50.11**	**54.45**	**57.77**	**57.93**

C.3 Entity Count

Table A4 shows that EvoSee improves every trajectory metric in each entity count group. Its trajectory scores decrease across the three groups, while Future MeanIoU retains a 5.90 point advantage over Qwen3.5-9B in the highest group. Gains in both point recovery and overlap show that this continuity advantage holds across scene densities.

Table A4. Entity Continuity by observed entity count.

			Observed			Future			Combined		
Entities	**Method**	N	**P-F1**	**P-R**	**mIoU**	**P-F1**	**P-R**	**mIoU**	**P-F1**	**P-R**	**mIoU**
1–3	Qwen3.5-9B	6,400	26.41	27.65	31.00	19.81	20.58	25.31	23.28	24.25	28.26
1–3	**EvoSee (Ours)**	6,400	**43.94**	**46.65**	**41.58**	**38.32**	**41.18**	**37.38**	**41.21**	**43.95**	**39.55**
4–6	Qwen3.5-9B	2,335	25.08	21.56	22.32	18.35	15.75	17.81	21.92	18.81	20.19
4–6	**EvoSee (Ours)**	2,335	**29.69**	**28.89**	**28.65**	**27.30**	**27.23**	**26.16**	**28.55**	**28.09**	**27.45**
≥ 7	Qwen3.5-9B	4,891	10.71	8.47	8.91	8.48	6.73	7.27	9.65	7.64	8.12
≥ 7	**EvoSee (Ours)**	4,891	**13.41**	**11.79**	**15.84**	**10.74**	**9.63**	**13.17**	**12.14**	**10.77**	**14.57**

Table A5 shows gains in all semantic metrics for every entity count group. EvoSee's Entity scores decrease more than its Event and Visual Evidence scores as entity count rises. Its Visual Evidence Global F1 margin over Qwen3.5-9B grows to 33.09 points in the highest group, accompanied by Scene-F1 gains. Evidence category recognition thus retains an advantage even where entity references become less accurate.

Table A5. Scene Evolution Reasoning by observed entity count.

			Entity			Event			Visual Evidence		
Entities	**Method**	N	**GF1**	**SF1**	**GPR**	**GF1**	**SF1**	**GPR**	**GF1**	**SF1**	**GPR**
1–3	Qwen3.5-9B	6,400	59.77	54.54	59.86	28.76	26.50	29.65	48.74	53.16	50.25
1–3	**EvoSee (Ours)**	6,400	**70.82**	**72.68**	**71.01**	**52.44**	**53.72**	**52.73**	**64.28**	**66.44**	**64.64**
4–6	Qwen3.5-9B	2,335	53.63	50.29	56.46	27.90	28.68	28.26	24.61	25.32	24.93
4–6	**EvoSee (Ours)**	2,335	**69.33**	**69.68**	**69.75**	**46.66**	**47.57**	**50.66**	**51.81**	**53.93**	**56.26**
≥ 7	Qwen3.5-9B	4,891	18.57	18.11	22.35	31.57	30.19	32.42	15.16	13.71	15.57
≥ 7	**EvoSee (Ours)**	4,891	**30.10**	**31.37**	**31.36**	**43.33**	**43.25**	**49.93**	**48.25**	**48.25**	**55.59**

Individual entity counts. Individual counts resolve variation within the broad density groups. Counts above 30 are pooled into > 30, with trajectory scores and Scene-F1 averaged over samples and Global F1 and GPR recomputed from pooled matching counts.

Table A6 shows that EvoSee leads every trajectory metric for one to four entities and for reported counts from sixteen onward, including > 30. Qwen3.5-9B has higher Point-F1 in several intermediate groups, including five to eleven entities. The advantage therefore varies within the broad entity count groups.

Table A6. Entity Continuity at each observed entity count.

			Observed			Future			Combined		
Entities	**Method**	N	**P-F1**	**P-R**	**mIoU**	**P-F1**	**P-R**	**mIoU**	**P-F1**	**P-R**	**mIoU**
1	Qwen3.5-9B	947	20.68	26.63	33.28	12.79	16.35	24.77	16.88	21.68	29.18
1	**EvoSee (Ours)**	947	**64.73**	**70.92**	**56.06**	**53.57**	**57.74**	**47.54**	**59.29**	**64.47**	**51.89**
2	Qwen3.5-9B	3,458	30.82	31.39	34.22	23.44	23.68	27.90	27.28	27.63	31.16
2	**EvoSee (Ours)**	3,458	**46.30**	**46.96**	**39.88**	**39.41**	**40.37**	**35.48**	**42.93**	**43.69**	**37.74**
3	Qwen3.5-9B	1,995	21.50	21.65	24.33	16.85	17.21	21.08	19.38	19.60	22.81
3	**EvoSee (Ours)**	1,995	**29.97**	**34.60**	**37.66**	**29.19**	**34.72**	**35.86**	**29.65**	**34.67**	**36.82**
4	Qwen3.5-9B	1,351	20.16	17.96	20.67	15.28	13.68	17.38	17.90	15.95	19.11
4	**EvoSee (Ours)**	1,351	**36.34**	**35.07**	**33.25**	**34.64**	**34.13**	**31.41**	**35.54**	**34.63**	**32.36**
5	Qwen3.5-9B	492	**32.68**	**27.00**	**25.08**	**22.69**	18.33	18.32	**28.03**	**22.95**	21.93
5	**EvoSee (Ours)**	492	22.56	22.91	24.08	18.80	**19.45**	**20.17**	20.81	21.29	**22.25**
6	Qwen3.5-9B	492	**31.00**	**26.02**	**24.08**	**22.43**	**18.83**	**18.49**	**26.87**	**22.53**	**21.40**
6	**EvoSee (Ours)**	492	18.54	17.92	20.58	15.67	16.04	17.74	17.12	16.94	19.18
7	Qwen3.5-9B	578	**30.34**	**24.19**	**22.88**	**24.83**	19.86	18.84	**27.72**	**22.12**	**20.95**
7	**EvoSee (Ours)**	578	24.29	21.46	20.99	22.28	**19.95**	**18.92**	23.33	20.75	20.03
8	Qwen3.5-9B	395	**27.93**	**23.66**	22.47	**22.23**	**18.96**	18.73	**25.28**	**21.45**	20.71
8	**EvoSee (Ours)**	395	16.92	16.80	**22.48**	14.45	14.65	**19.61**	15.82	15.85	**21.21**
9	Qwen3.5-9B	378	**28.06**	**23.69**	**21.96**	**21.55**	**18.09**	17.65	**24.95**	**21.00**	19.89
9	**EvoSee (Ours)**	378	16.98	16.97	21.92	13.56	13.69	**19.09**	15.36	15.42	**20.59**
10	Qwen3.5-9B	196	**15.01**	12.43	13.65	**10.36**	8.51	9.60	**12.86**	10.57	11.69
10	**EvoSee (Ours)**	196	12.78	**12.80**	**17.34**	8.80	**8.93**	**12.90**	10.89	**10.95**	**15.20**
11	Qwen3.5-9B	100	**18.46**	**13.19**	13.53	**16.41**	**11.63**	11.34	**17.47**	**12.43**	12.46
11	**EvoSee (Ours)**	100	12.72	12.04	**16.74**	9.52	9.01	**12.95**	11.18	10.58	**14.92**
12	Qwen3.5-9B	81	6.48	4.11	5.34	5.01	3.06	3.71	5.86	3.65	4.50
12	**EvoSee (Ours)**	81	**10.57**	**9.92**	**14.75**	**8.61**	**8.48**	**12.80**	**9.69**	**9.29**	**13.72**
13	Qwen3.5-9B	88	**10.46**	**6.93**	6.95	**9.04**	5.99	5.70	**9.77**	6.47	6.33
13	**EvoSee (Ours)**	88	6.71	6.32	**10.37**	6.63	**6.69**	**9.03**	6.66	**6.48**	**9.68**
14	Qwen3.5-9B	121	5.25	3.39	4.37	4.29	2.74	3.23	4.79	3.08	3.82
14	**EvoSee (Ours)**	121	**7.52**	**5.91**	**9.90**	**4.88**	**3.89**	**7.25**	**6.24**	**4.93**	**8.61**
15	Qwen3.5-9B	121	**5.81**	3.57	4.17	**4.91**	3.03	3.21	**5.38**	3.31	3.71
15	**EvoSee (Ours)**	121	5.71	**5.10**	**10.10**	3.77	**3.67**	**6.50**	4.80	**4.43**	**8.41**
16	Qwen3.5-9B	102	4.19	2.58	3.31	3.35	2.06	2.29	3.81	2.34	2.84
16	**EvoSee (Ours)**	102	**6.62**	**5.35**	**10.41**	**3.90**	**3.09**	**6.11**	**5.33**	**4.28**	**8.41**
17	Qwen3.5-9B	62	1.74	1.08	1.90	1.24	0.73	0.69	1.52	0.93	1.35
17	**EvoSee (Ours)**	62	**5.49**	**4.84**	**9.25**	**2.02**	**1.80**	**4.95**	**3.87**	**3.41**	**7.23**
18	Qwen3.5-9B	65	0.22	0.14	1.03	0.03	0.02	0.32	0.12	0.08	0.68
18	**EvoSee (Ours)**	65	**5.74**	**4.82**	**9.05**	**2.69**	**2.22**	**5.04**	**4.28**	**3.57**	**7.12**
19	Qwen3.5-9B	58	1.72	1.16	2.82	1.20	0.78	1.94	1.46	0.97	2.38
19	**EvoSee (Ours)**	58	**7.07**	**6.42**	**13.12**	**5.29**	**4.85**	**8.65**	**6.21**	**5.66**	**10.97**
20	Qwen3.5-9B	72	3.96	2.91	5.59	3.15	2.53	5.13	3.62	2.74	5.38
20	**EvoSee (Ours)**	72	**14.68**	**13.13**	**20.73**	**12.38**	**11.38**	**18.15**	**13.57**	**12.27**	**19.51**
21	Qwen3.5-9B	67	3.42	2.46	5.73	3.16	2.29	5.12	3.29	2.38	5.44
21	**EvoSee (Ours)**	67	**15.22**	**13.08**	**17.95**	**11.27**	**9.92**	**15.71**	**13.33**	**11.58**	**16.93**
22	Qwen3.5-9B	116	5.84	4.77	7.94	5.25	4.37	7.66	5.56	4.58	7.81
22	**EvoSee (Ours)**	116	**19.33**	**16.49**	**23.28**	**15.49**	**13.51**	**22.13**	**17.47**	**15.05**	**22.73**
23	Qwen3.5-9B	153	3.99	2.75	6.68	3.16	2.22	6.26	3.60	2.49	6.47
23	**EvoSee (Ours)**	153	**22.42**	**19.27**	**24.49**	**17.22**	**15.04**	**22.09**	**19.88**	**17.21**	**23.33**

Continued on next page

Table A6 (continued)

Entities	Method	N	Observed			Future			Combined		
			P-F1	P-R	mIoU	P-F1	P-R	mIoU	P-F1	P-R	mIoU
24	Qwen3.5-9B	130	2.55	1.68	5.15	2.15	1.43	4.68	2.35	1.56	4.92
24	**EvoSee (Ours)**	130	**25.85**	**22.07**	**24.45**	**22.02**	**19.10**	**22.41**	**23.98**	**20.62**	**23.45**
25	Qwen3.5-9B	166	3.26	2.01	5.55	2.77	1.74	5.21	3.03	1.88	5.38
25	**EvoSee (Ours)**	166	**26.55**	**22.83**	**25.50**	**23.49**	**20.22**	**22.92**	**25.06**	**21.55**	**24.23**
26	Qwen3.5-9B	51	3.65	2.43	6.62	2.53	1.89	6.04	3.11	2.16	6.33
26	**EvoSee (Ours)**	51	**22.85**	**19.43**	**21.98**	**16.55**	**14.54**	**18.39**	**19.89**	**17.13**	**20.28**
27	Qwen3.5-9B	76	0.84	0.54	1.06	0.78	0.51	0.85	0.81	0.53	0.96
27	**EvoSee (Ours)**	76	**6.25**	**4.77**	**7.34**	**5.03**	**3.95**	**6.36**	**5.67**	**4.38**	**6.87**
28	Qwen3.5-9B	68	1.04	0.58	1.14	0.76	0.43	0.70	0.90	0.50	0.91
28	**EvoSee (Ours)**	68	**11.74**	**9.02**	**12.69**	**9.32**	**7.32**	**10.92**	**10.57**	**8.20**	**11.81**
29	Qwen3.5-9B	130	0.83	0.48	0.99	0.70	0.39	0.81	0.77	0.44	0.91
29	**EvoSee (Ours)**	130	**12.79**	**9.71**	**16.64**	**11.14**	**8.64**	**15.07**	**12.03**	**9.22**	**15.89**
30	Qwen3.5-9B	130	1.28	0.71	1.13	0.74	0.41	0.85	1.03	0.57	0.99
30	**EvoSee (Ours)**	130	**11.25**	**8.66**	**14.63**	**7.49**	**5.87**	**11.44**	**9.46**	**7.33**	**13.11**
> 30	Qwen3.5-9B	1,387	1.22	0.67	0.93	0.74	0.41	0.66	0.99	0.54	0.80
> 30	**EvoSee (Ours)**	1,387	**6.31**	**4.74**	**8.97**	**4.11**	**3.11**	**6.35**	**5.27**	**3.97**	**7.71**

Table A7 shows that EvoSee improves Visual Evidence Global F1, Scene-F1, and GPR at every reported count. Its Entity and Event results have local reversals, including Entity GPR and Event F1. For > 30 entities, it leads all nine metrics, with a 39.15 point Visual Evidence Global F1 gain. The advantage in evidence category recognition therefore covers more individual counts than the advantages in Entity and Event recognition.

Table A7. Scene Evolution Reasoning at each observed entity count.

Entities	Method	N	Entity			Event			Visual Evidence		
			GF1	SF1	GPR	GF1	SF1	GPR	GF1	SF1	GPR
1	Qwen3.5-9B	947	58.62	62.84	61.08	36.05	31.49	37.57	71.50	79.13	74.51
1	**EvoSee (Ours)**	947	**68.62**	**79.86**	**71.05**	**81.73**	**81.73**	**81.73**	**84.69**	**84.69**	**84.69**
2	Qwen3.5-9B	3,458	58.86	51.46	59.12	29.56	26.63	31.04	51.57	54.75	54.15
2	**EvoSee (Ours)**	3,458	**69.55**	**70.71**	**69.61**	**49.83**	**50.38**	**49.87**	**64.44**	**65.28**	**64.49**
3	Qwen3.5-9B	1,995	61.09	55.94	61.37	25.00	23.89	25.25	36.58	38.07	36.94
3	**EvoSee (Ours)**	1,995	**72.67**	**72.67**	**72.86**	**45.38**	**46.23**	**46.66**	**56.62**	**59.79**	**58.21**
4	Qwen3.5-9B	1,351	50.44	46.91	54.21	32.17	34.21	32.54	22.92	23.81	23.18
4	**EvoSee (Ours)**	1,351	**76.63**	**76.56**	**76.88**	**52.35**	**53.89**	**55.72**	**56.43**	**59.15**	**60.07**
5	Qwen3.5-9B	492	57.95	56.37	59.76	23.48	21.08	23.90	29.61	30.49	30.14
5	**EvoSee (Ours)**	492	**61.50**	**60.98**	**61.97**	**39.75**	**38.28**	**44.22**	**47.52**	**48.04**	**52.86**
6	Qwen3.5-9B	492	56.06	53.46	58.19	22.95	21.10	23.23	23.49	24.31	23.77
6	**EvoSee (Ours)**	492	**59.63**	**59.49**	**60.44**	**40.36**	**39.50**	**45.19**	**45.39**	**45.49**	**50.82**
7	Qwen3.5-9B	578	41.02	36.65	44.75	18.22	16.41	18.93	17.23	16.46	17.91
7	**EvoSee (Ours)**	578	**48.88**	**48.97**	**49.99**	**42.57**	**42.21**	**48.75**	**46.21**	**46.31**	**52.92**
8	Qwen3.5-9B	395	35.38	31.53	40.28	23.64	21.76	24.25	21.17	21.00	21.71
8	**EvoSee (Ours)**	395	**49.23**	**49.37**	**50.08**	**43.08**	**42.70**	**49.24**	**49.22**	**49.28**	**56.26**
9	Qwen3.5-9B	378	33.69	31.27	36.86	23.80	21.60	24.16	21.52	21.16	21.84
9	**EvoSee (Ours)**	378	**48.52**	**48.39**	**49.48**	**46.14**	**46.08**	**53.19**	**48.54**	**48.46**	**55.95**
10	Qwen3.5-9B	196	28.99	27.12	33.76	35.06	32.87	35.81	19.50	17.87	19.92
10	**EvoSee (Ours)**	196	**37.96**	**38.13**	**39.11**	**41.75**	**41.92**	**48.05**	**46.91**	**47.02**	**53.98**
11	Qwen3.5-9B	100	38.32	36.53	**43.43**	29.30	28.03	29.74	21.48	19.90	21.81
11	**EvoSee (Ours)**	100	**38.91**	**40.00**	40.17	**43.00**	**43.00**	**49.65**	**46.00**	**46.00**	**53.12**
12	Qwen3.5-9B	81	29.01	28.31	**35.92**	30.69	27.78	31.92	14.29	13.66	14.86
12	**EvoSee (Ours)**	81	**32.18**	**32.43**	33.05	**42.86**	**42.59**	**49.39**	**45.34**	**45.06**	**52.25**
13	Qwen3.5-9B	88	**27.88**	25.92	**31.23**	29.46	26.78	29.95	20.09	17.95	20.42
13	**EvoSee (Ours)**	88	26.43	**26.56**	27.04	**45.45**	**45.45**	**52.49**	**48.86**	**48.86**	**56.42**
14	Qwen3.5-9B	121	17.19	14.53	23.05	**41.82**	40.66	42.88	11.13	9.72	11.41
14	**EvoSee (Ours)**	121	**29.41**	**29.48**	**30.56**	40.99	**41.05**	**47.30**	**48.45**	**48.48**	**55.90**

Continued on next page

Table A7 (continued)

Entities	Method	N	Entity GF1	Entity SF1	Entity GPR	Event GF1	Event SF1	Event GPR	Visual Evidence GF1	Visual Evidence SF1	Visual Evidence GPR
15	Qwen3.5-9B	121	21.71	19.88	29.85	36.77	34.05	37.85	6.19	5.90	6.37
15	**EvoSee (Ours)**	121	**30.39**	**30.60**	**31.54**	**44.07**	**43.94**	**50.79**	**46.15**	**46.01**	**53.18**
16	Qwen3.5-9B	102	12.58	11.47	18.63	**44.84**	**43.07**	45.87	7.94	7.25	8.12
16	**EvoSee (Ours)**	102	**31.97**	**32.15**	**33.71**	39.80	39.71	**45.92**	**43.73**	**43.63**	**50.46**
17	Qwen3.5-9B	62	7.69	6.86	12.68	**44.30**	**43.39**	45.33	8.47	7.80	8.67
17	**EvoSee (Ours)**	62	**29.54**	**29.37**	**31.27**	41.13	41.13	**47.49**	**45.97**	**45.97**	**53.08**
18	Qwen3.5-9B	65	4.73	5.58	8.37	**38.41**	**37.03**	39.12	4.27	4.36	4.35
18	**EvoSee (Ours)**	65	**33.33**	**34.14**	**35.15**	35.38	35.38	**40.86**	**45.38**	**45.38**	**52.41**
19	Qwen3.5-9B	58	6.95	6.13	12.46	38.64	37.64	40.75	6.06	6.03	6.39
19	**EvoSee (Ours)**	58	**24.65**	**23.28**	**26.24**	**46.55**	**46.55**	**53.75**	**47.41**	**47.41**	**54.75**
20	Qwen3.5-9B	72	8.45	7.32	12.63	36.00	32.78	37.03	16.00	13.80	16.46
20	**EvoSee (Ours)**	72	**25.54**	**25.35**	**26.49**	**48.61**	**48.61**	**56.13**	**48.61**	**48.61**	**56.13**
21	Qwen3.5-9B	67	12.21	11.29	17.25	30.00	27.26	31.90	9.33	8.01	9.92
21	**EvoSee (Ours)**	67	**30.32**	**30.25**	**31.97**	**47.01**	**47.01**	**54.29**	**50.00**	**50.00**	**57.74**
22	Qwen3.5-9B	116	7.49	6.92	10.81	36.14	31.87	37.27	17.89	15.78	18.45
22	**EvoSee (Ours)**	116	**23.73**	**23.73**	**24.45**	**49.14**	**49.14**	**56.74**	**49.57**	**49.57**	**57.24**
23	Qwen3.5-9B	153	5.47	4.72	8.29	34.90	30.70	35.88	20.67	17.63	21.25
23	**EvoSee (Ours)**	153	**19.63**	**19.53**	**20.53**	**49.02**	**49.02**	**56.60**	**49.35**	**49.35**	**56.98**
24	Qwen3.5-9B	130	4.57	3.79	7.05	18.78	16.31	19.59	18.45	16.26	19.25
24	**EvoSee (Ours)**	130	**19.59**	**19.62**	**20.65**	**47.31**	**47.31**	**54.63**	**48.46**	**48.46**	**55.96**
25	Qwen3.5-9B	166	3.55	3.12	5.42	30.67	27.63	31.46	18.40	16.12	18.88
25	**EvoSee (Ours)**	166	**18.86**	**18.88**	**20.14**	**45.78**	**45.78**	**52.87**	**49.40**	**49.40**	**57.04**
26	Qwen3.5-9B	51	8.22	7.44	10.84	30.04	28.04	30.72	17.39	14.38	17.79
26	**EvoSee (Ours)**	51	**20.13**	**20.05**	**21.47**	**38.24**	**38.24**	**44.15**	**45.10**	**45.10**	**52.07**
27	Qwen3.5-9B	76	4.11	3.86	5.51	33.33	34.12	34.81	16.95	14.65	17.70
27	**EvoSee (Ours)**	76	**28.07**	**28.40**	**29.77**	**42.11**	**42.11**	**48.62**	**46.71**	**46.71**	**53.94**
28	Qwen3.5-9B	68	10.59	11.68	13.44	30.43	31.47	31.58	11.18	9.02	11.60
28	**EvoSee (Ours)**	68	**25.06**	**25.40**	**26.50**	**38.97**	**38.97**	**45.00**	**44.85**	**44.85**	**51.79**
29	Qwen3.5-9B	130	8.18	7.98	9.71	31.24	32.10	31.55	18.69	16.56	18.87
29	**EvoSee (Ours)**	130	**26.64**	**26.77**	**27.58**	**42.69**	**42.69**	**49.30**	**48.08**	**48.08**	**55.51**
30	Qwen3.5-9B	130	7.39	7.06	8.99	30.34	30.51	31.02	15.17	13.10	15.51
30	**EvoSee (Ours)**	130	**21.68**	**21.90**	**22.69**	**43.08**	**43.08**	**49.74**	**48.46**	**48.46**	**55.96**
> 30	Qwen3.5-9B	1,387	11.30	11.42	13.27	38.01	38.39	39.07	10.71	8.77	11.01
> 30	**EvoSee (Ours)**	1,387	**20.01**	**20.49**	**21.23**	**42.14**	**42.14**	**48.66**	**49.86**	**49.86**	**57.57**

C.4 Temporal Extent

Equal observation and future durations define each temporal group. Table A8 shows that EvoSee leads all trajectory metrics in every group. Its observed accuracy stays close across groups, while future accuracy decreases with longer intervals. Its Future Point-F1 advantage over Qwen3.5-9B reaches 15.14 points in the longest group, accompanied by recall and overlap gains.

Table A8. Entity Continuity across temporal extents.

Extent	Method	N	Observed P-F1	Observed P-R	Observed mIoU	Future P-F1	Future P-R	Future mIoU	Combined P-F1	Combined P-R	Combined mIoU
1–3 s	Qwen3.5-9B	11,001	22.28	21.30	22.32	17.11	16.31	18.45	19.86	18.93	20.48
1–3 s	**EvoSee (Ours)**	11,001	**30.25**	**31.02**	**30.12**	**26.99**	**28.21**	**27.22**	**28.70**	**29.67**	**28.74**
4–6 s	Qwen3.5-9B	1,616	14.08	13.95	19.32	10.21	9.90	15.15	12.21	11.96	17.29
4–6 s	**EvoSee (Ours)**	1,616	**31.90**	**31.69**	**30.28**	**26.58**	**26.51**	**26.03**	**29.28**	**29.10**	**28.18**
≥ 7 s	Qwen3.5-9B	1,009	11.97	11.80	17.15	6.34	5.92	11.59	9.14	8.81	14.40
≥ 7 s	**EvoSee (Ours)**	1,009	**31.47**	**30.99**	**29.96**	**21.48**	**20.90**	**23.08**	**26.49**	**25.92**	**26.54**

Table A9 shows that EvoSee also improves all semantic metrics in every temporal group. Its Event Global F1 margins narrow with longer intervals, while Visual Evidence Global F1 margins remain

above 24 points. The accompanying Scene-F1 gains show that the advantage also holds when scenes receive equal weight.

Table A9. Scene Evolution Reasoning across temporal extents.

			Entity			Event			Visual Evidence		
Extent	**Method**	N	**GF1**	**SF1**	**GPR**	**GF1**	**SF1**	**GPR**	**GF1**	**SF1**	**GPR**
1–3 s	Qwen3.5-9B	11,001	39.90	41.97	42.17	27.68	26.21	27.71	30.49	34.66	30.51
1–3 s	**EvoSee (Ours)**	11,001	**51.00**	**57.98**	**51.38**	**47.38**	**49.27**	**50.40**	**54.93**	**58.41**	**58.44**
4–6 s	Qwen3.5-9B	1,616	28.80	35.79	32.90	37.82	35.98	37.82	27.71	32.80	27.71
4–6 s	**EvoSee (Ours)**	1,616	**42.40**	**54.44**	**43.46**	**46.58**	**48.26**	**49.51**	**52.46**	**55.13**	**55.77**
$\geq$ 7 s	Qwen3.5-9B	1,009	26.79	35.20	31.47	39.47	37.35	39.47	26.33	31.78	26.33
$\geq$ 7 s	**EvoSee (Ours)**	1,009	**41.00**	**54.91**	**42.60**	**44.85**	**46.02**	**47.87**	**52.38**	**55.07**	**55.92**

Individual temporal extents. The individual temporal extents reveal variation within these groups. Table A10 shows that EvoSee leads all trajectory metrics at all ten extents. Its observed scores vary less than its future scores, which decrease over longer intervals. Identity matching remains fixed from observation, so these gains concern continuation of the same entities at each temporal extent.

Table A10. Entity Continuity at each temporal extent.

			Observed			Future			Combined		
Extent	**Method**	N	**P-F1**	**P-R**	**mIoU**	**P-F1**	**P-R**	**mIoU**	**P-F1**	**P-R**	**mIoU**
1 s	Qwen3.5-9B	7,294	27.12	25.91	26.85	20.99	20.00	22.35	24.24	23.10	24.70
1 s	**EvoSee (Ours)**	7,294	**30.59**	**31.50**	**30.49**	**26.83**	**28.48**	**27.60**	**28.79**	**30.04**	**29.11**
2 s	Qwen3.5-9B	2,498	18.01	17.25	18.93	13.43	12.81	15.23	15.88	15.17	17.19
2 s	**EvoSee (Ours)**	2,498	**29.53**	**30.16**	**29.43**	**27.41**	**27.79**	**26.60**	**28.56**	**29.03**	**28.12**
3 s	Qwen3.5-9B	1,209	1.92	1.84	2.02	1.27	1.25	1.54	1.61	1.55	1.79
3 s	**EvoSee (Ours)**	1,209	**29.71**	**29.91**	**29.31**	**27.08**	**27.45**	**26.15**	**28.46**	**28.71**	**27.81**
4 s	Qwen3.5-9B	690	14.84	14.81	20.11	10.90	10.73	16.17	12.94	12.81	18.20
4 s	**EvoSee (Ours)**	690	**31.75**	**31.64**	**30.51**	**28.11**	**28.15**	**26.99**	**29.98**	**29.92**	**28.79**
5 s	Qwen3.5-9B	520	14.25	13.99	19.11	10.60	10.20	15.00	12.47	12.12	17.09
5 s	**EvoSee (Ours)**	520	**31.73**	**31.38**	**29.90**	**26.31**	**26.10**	**25.61**	**29.04**	**28.72**	**27.77**
6 s	Qwen3.5-9B	406	12.58	12.45	18.27	8.55	8.10	13.62	10.65	10.30	15.99
6 s	**EvoSee (Ours)**	406	**32.36**	**32.16**	**30.39**	**24.33**	**24.25**	**24.94**	**28.38**	**28.21**	**27.69**
7 s	Qwen3.5-9B	329	13.33	13.07	18.67	7.63	7.03	12.99	10.50	10.02	15.86
7 s	**EvoSee (Ours)**	329	**32.89**	**32.49**	**30.27**	**22.91**	**22.45**	**23.90**	**27.90**	**27.44**	**27.11**
8 s	Qwen3.5-9B	268	11.80	11.65	16.90	6.54	6.06	11.60	9.13	8.78	14.28
8 s	**EvoSee (Ours)**	268	**30.84**	**30.25**	**29.66**	**21.72**	**20.93**	**23.11**	**26.33**	**25.58**	**26.41**
9 s	Qwen3.5-9B	222	10.34	10.26	15.63	4.62	4.45	10.39	7.42	7.27	13.02
9 s	**EvoSee (Ours)**	222	**31.37**	**30.77**	**29.85**	**21.36**	**20.73**	**22.61**	**26.33**	**25.65**	**26.22**
10 s	Qwen3.5-9B	190	11.75	11.61	16.63	5.81	5.54	10.53	8.80	8.57	13.64
10 s	**EvoSee (Ours)**	190	**30.03**	**29.72**	**29.98**	**18.81**	**18.40**	**22.15**	**24.47**	**24.06**	**26.13**

Table A11 shows that EvoSee also leads every semantic metric at every temporal extent. Its Event and Visual Evidence Global F1 scores vary less than Entity Global F1, and Scene-F1 gains accompany the pooled improvements. Both continuity and semantic evaluations therefore show gains across the individual temporal settings.

Table A11. Scene Evolution Reasoning at each temporal extent.

			Entity			Event			Visual Evidence		
Extent	**Method**	N	**GF1**	**SF1**	**GPR**	**GF1**	**SF1**	**GPR**	**GF1**	**SF1**	**GPR**
1 s	Qwen3.5-9B	7,294	47.43	49.49	49.76	25.40	24.11	25.44	30.49	34.45	30.52
1 s	**EvoSee (Ours)**	7,294	**52.32**	**58.62**	**52.62**	**47.56**	**49.52**	**50.64**	**54.70**	**58.15**	**58.25**
2 s	Qwen3.5-9B	2,498	35.39	37.71	37.75	30.28	28.50	30.30	30.26	34.85	30.28
2 s	**EvoSee (Ours)**	2,498	**49.53**	**57.10**	**49.99**	**47.25**	**49.10**	**50.22**	**55.62**	**59.26**	**59.11**

Continued on next page

Table A11 (continued)

Extent	Method	N	Entity			Event			Visual Evidence		
			GF1	SF1	GPR	GF1	SF1	GPR	GF1	SF1	GPR
3 s	Qwen3.5-9B	1,209	5.37	5.38	5.87	35.93	34.20	35.93	30.95	35.54	30.95
3 s	**EvoSee (Ours)**	1,209	**46.44**	**55.98**	**47.16**	**46.55**	**48.13**	**49.36**	**54.89**	**58.16**	**58.20**
4 s	Qwen3.5-9B	690	28.25	34.56	32.27	36.87	34.69	36.87	28.93	33.49	28.93
4 s	**EvoSee (Ours)**	690	**42.72**	**54.30**	**43.64**	**46.69**	**48.43**	**49.59**	**52.45**	**54.90**	**55.71**
5 s	Qwen3.5-9B	520	29.57	37.14	33.57	38.55	36.93	38.55	26.92	32.48	26.93
5 s	**EvoSee (Ours)**	520	**42.25**	**54.36**	**43.35**	**46.51**	**48.24**	**49.44**	**52.99**	**56.09**	**56.33**
6 s	Qwen3.5-9B	406	28.70	36.14	33.10	38.51	36.95	38.51	26.63	32.03	26.64
6 s	**EvoSee (Ours)**	406	**42.08**	**54.77**	**43.33**	**46.48**	**47.99**	**49.47**	**51.82**	**54.27**	**55.16**
7 s	Qwen3.5-9B	329	28.90	37.55	33.40	39.52	37.65	39.53	25.95	31.67	25.95
7 s	**EvoSee (Ours)**	329	**41.00**	**54.51**	**42.40**	**44.00**	**45.24**	**46.94**	**51.54**	**53.85**	**54.97**
8 s	Qwen3.5-9B	268	25.96	34.81	30.54	36.33	33.94	36.33	26.62	32.82	26.62
8 s	**EvoSee (Ours)**	268	**41.10**	**54.79**	**42.67**	**45.87**	**47.01**	**48.96**	**52.67**	**55.53**	**56.22**
9 s	Qwen3.5-9B	222	26.11	34.35	30.72	41.55	39.50	41.58	27.70	32.03	27.72
9 s	**EvoSee (Ours)**	222	**40.61**	**54.79**	**42.34**	**44.80**	**46.17**	**47.83**	**52.42**	**55.26**	**55.95**
10 s	Qwen3.5-9B	190	25.24	32.68	30.38	41.26	39.12	41.26	24.91	30.19	24.91
10 s	**EvoSee (Ours)**	190	**41.28**	**55.90**	**43.16**	**44.90**	**45.79**	**48.00**	**53.40**	**56.32**	**57.09**

C.5 Video Source

The source analysis covers 32 APIDIS, 11,048 OVT-B, and 2,546 TAO test windows. Table A12 separates localization in observation from continuation into the future. OVT-B has the highest future and combined trajectory scores. APIDIS has the highest observed MeanIoU alongside lower future accuracy, and TAO also shows a decline from observation to future. Spatial continuation therefore varies across sources even when observed localization is strong.

Table A12. Entity Continuity across video sources for EvoSee.

Source	N	Observed			Future			Combined		
		P-F1	P-R	mIoU	P-F1	P-R	mIoU	P-F1	P-R	mIoU
APIDIS	32	13.38	25.79	**34.04**	2.09	3.44	10.15	7.59	14.31	21.70
OVT-B	11,048	**32.97**	**32.74**	31.09	**29.55**	**30.09**	**28.70**	**31.29**	**31.42**	**29.92**
TAO	2,546	20.19	24.04	25.88	13.77	16.39	18.60	17.24	20.42	22.49

Table A13 reveals a different source profile for semantic prediction. TAO has the highest Entity scores, while APIDIS leads all Event and Visual Evidence metrics. On APIDIS, strong category recognition coexists with lower future geometric accuracy. Evaluating both predictions on the same source exposes this difference between recognition of change and spatial continuation.

Table A13. Scene Evolution Reasoning across video sources for EvoSee.

Source	N	Entity			Event			Visual Evidence		
		GF1	SF1	GPR	GF1	SF1	GPR	GF1	SF1	GPR
APIDIS	32	48.75	49.90	52.88	**65.63**	**65.63**	**65.63**	**81.25**	**81.25**	**81.25**
OVT-B	11,048	46.99	56.55	47.68	48.47	50.52	51.79	54.82	58.15	58.58
TAO	2,546	**59.45**	**60.81**	**59.49**	40.42	41.72	42.29	52.43	55.81	54.85

C.6 Scene Description Results

Table 3 reports BERTScore F1 for complete scene descriptions. Table A14 separates the present, future, and evidence content of those descriptions. EvoSee improves all three parts and gains 0.28 points overall over Qwen3.5-9B, with the largest gain in the present description.

Table A14. BERTScore F1 for the complete scene description and its present, future, and evidence parts.

Method	Overall	Present	Future	Evidence
Qwen3.5-9B	95.98	95.59	96.65	96.31
EvoSee (Ours)	**96.26**	**96.06**	**96.81**	**96.52**

D Qualitative Comparisons

The examples compare ground truth (GT), Qwen3.5-9B, and EvoSee under the same observation and future durations. The common question is "Given the observed H-second video window, its fixed keyframes, and the camera view, what scene evolution follows within the next H seconds?" The accompanying instruction requests observed trajectories, future trajectories, and a description relating the forecast to observed evidence. Reference close-ups from observation and future appear in the right panel.

Each row uses the same four timestamps. For each entity, the displayed centers accumulate from one to four points, with solid connections in observation and dashed connections into the future. Colors follow the fixed observed identity correspondence, and labels retain each method's output identities. The right panel places reference close-ups above the two descriptions. Gains displayed in each figure are EvoSee minus Qwen3.5-9B on that sample, measured in percentage points.

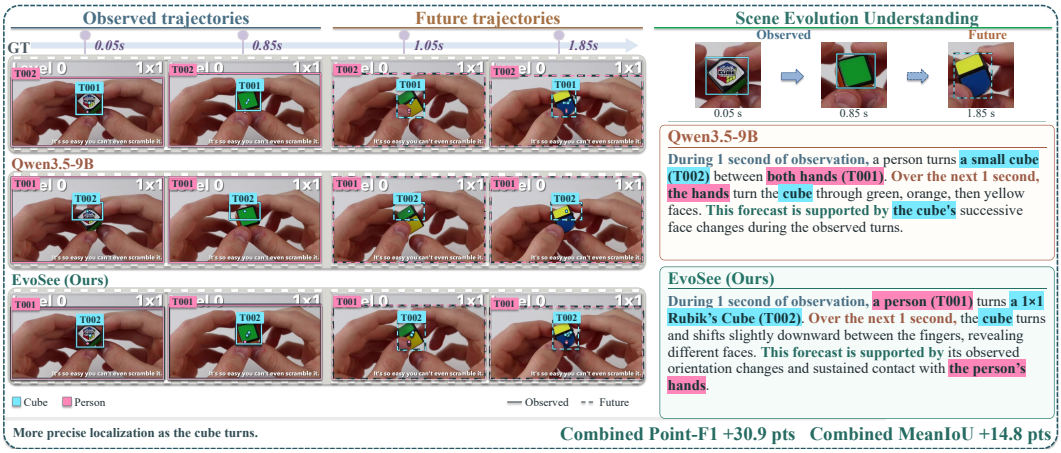


Figure A1. Cube rotation. EvoSee follows the changing spatial extent of the handheld cube, improving Combined Point-F1 by 30.9 points and MeanIoU by 14.8 points.

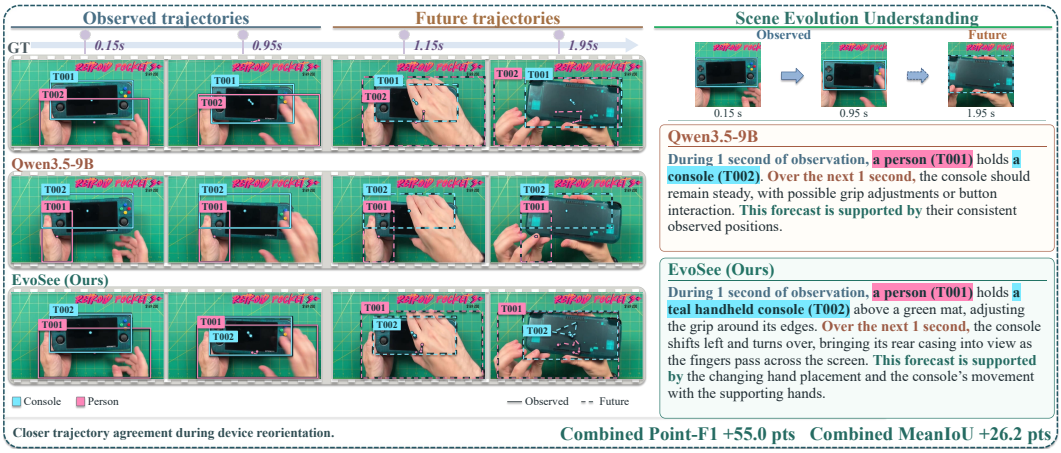


Figure A2. Device manipulation. Observed and future tracks follow a person and handheld console during reorientation. Combined Point-F1 and MeanIoU gains are 55.0 and 26.2 points.

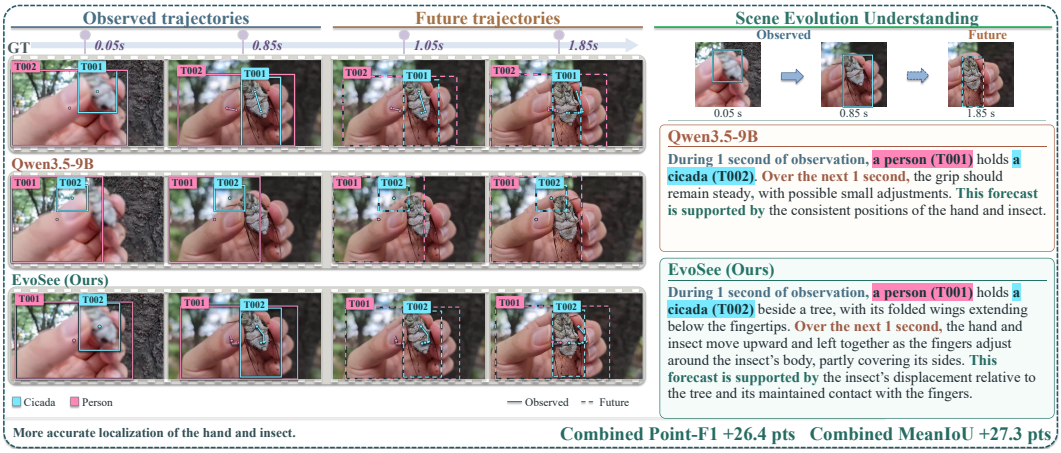


Figure A3. Hand and insect motion. The comparison shows entity localization at different spatial scales. Combined Point-F1 and MeanIoU improve by 26.4 and 27.3 points.

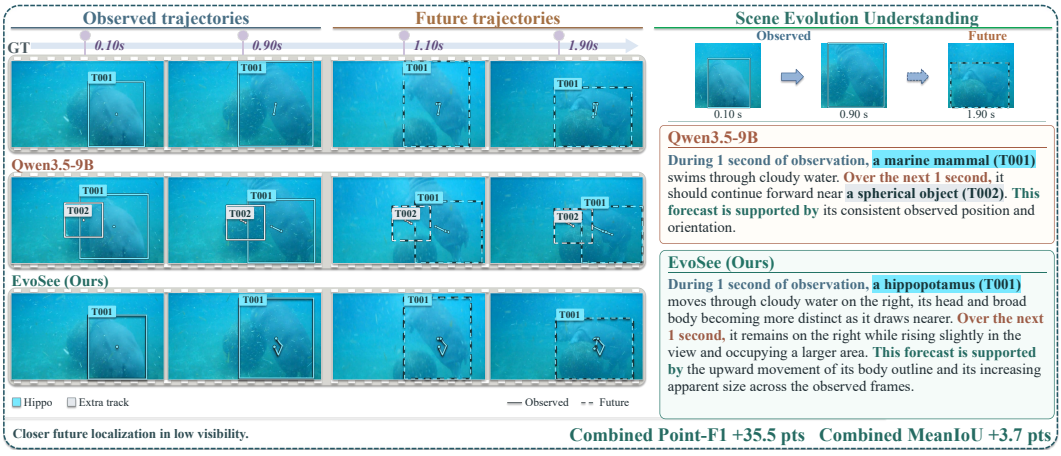


Figure A4. Underwater motion. EvoSee more closely localizes the animal through low visibility, with Combined Point-F1 and MeanIoU gains of 35.5 and 3.7 points.

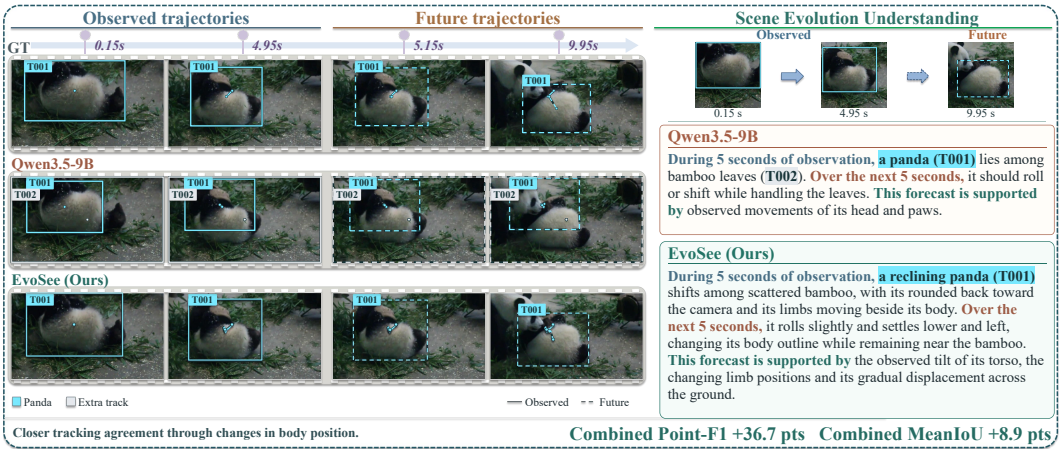


Figure A5. Extended body motion. Five seconds of observation precede a forecast of five seconds. Combined Point-F1 and MeanIoU gains are 36.7 and 8.9 points.

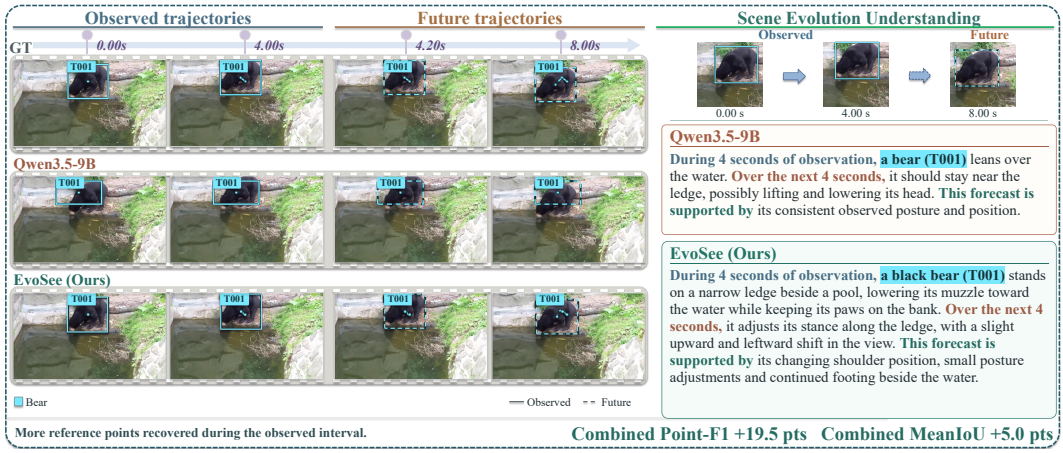


Figure A6. Motion near the water edge. Intervals of four seconds show localization around a stationary scene structure. Combined Point-F1 and MeanIoU improve by 19.5 and 5.0 points.