EVENT-STVG: SPATIO-TEMPORAL VIDEO GROUNDING FOR LONG-FORM EVENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatio-Temporal Video Grounding (STVG) aims to jointly localize a target in space and time given a natural-language query. Despite continued advances, existing STVG benchmarks remain largely centered on localizing targets through instance-specific descriptions, leaving them largely decoupled from long-form reasoning increasingly emphasized in multimodal large language models (MLLMs). In this paper, we introduce Event-STVG, a task that shifts STVG from locating what is described toward event-level reasoning by unifying long-form context, repeated-event reasoning, event localization, role inference, and multi-target grounding. To facilitate research on Event-STVG, we construct a large-scale benchmark comprising 4,016 annotated event occurrences across soccer and seven additional domains. Experiments with state-of-the-art MLLMs reveal substantial difficulty, with the strongest model achieving only 5.4% vIoU at the 10-minute horizon. Across domains, we consistently observe confusion among repeated same-type events, spatial grounding bottlenecks, and rare success in jointly grounding all required roles. Event-STVG establishes a new challenge for visual understanding in MLLMs, calling for future models that can jointly reason about complex events across space and time.

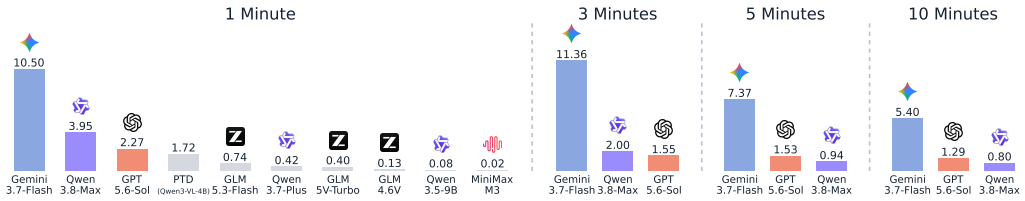


Figure 1: Model performance on Event-STVG across different video durations.

1 INTRODUCTION

Spatio-Temporal Video Grounding (STVG) aims to jointly localize a target of interest in space and time within an untrimmed video given a natural-language query (Zhang et al., 2020). By linking natural-language queries to precise spatiotemporal localization in video, STVG provides a concrete means of evaluating fine-grained multimodal grounding. Research in this area has advanced both grounding methods (Yang et al., 2022; Rasheed et al., 2026) and task formulations, with recent work extending STVG to multiple targets (Yao et al., 2026), long-form videos (Gu et al., 2026), and compositional queries (Wang et al., 2026b).

Despite their growing complexity, existing STVG tasks remain largely centered on description-guided localization. For example, the query in Figure 2 identifies a woman through her appearance, action, and relation to another person. As a result, current STVG benchmarks primarily test whether models can match a detailed linguistic description to the corresponding visual instance. This formulation is well suited to precise referent localization, but places limited demands on the richer long-form reasoning capabilities increasingly expected of MLLMs.

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.

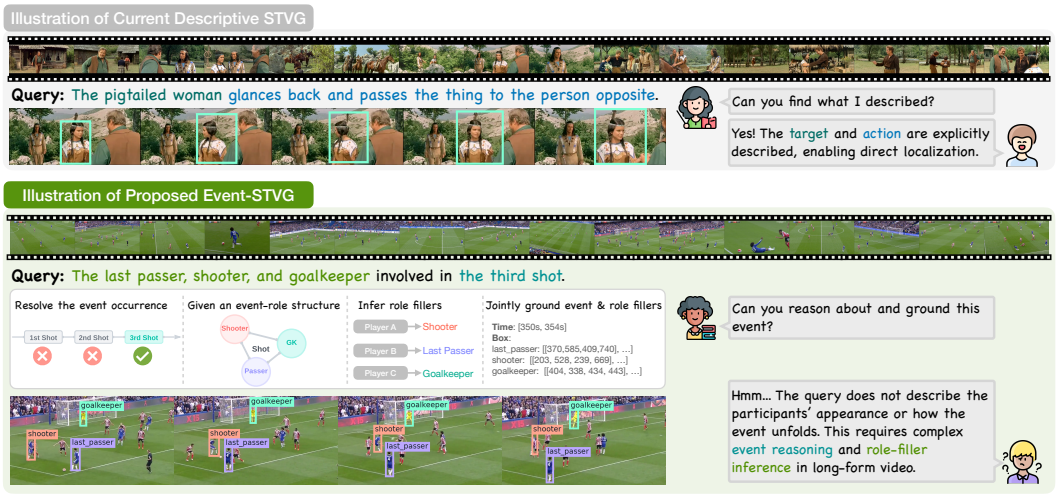


Figure 2: Comparison between Event-STVG and existing STVG. Existing STVG primarily performs description-guided localization, whereas Event-STVG requires complex event-level reasoning.

To move STVG beyond description-guided localization, we introduce **Event-STVG**, a new formulation that turns spatio-temporal grounding into a test of long-form event reasoning in MLLMs. Rather than describing the target instance directly, Event-STVG specifies an event category, its ordinal occurrence, and a set of semantic roles, requiring models to resolve the target occurrence, infer the corresponding role fillers, and jointly ground the event interval and its participants. This brings STVG substantially closer to the long-context temporal reasoning, repeated-event disambiguation, structured role inference, and multi-entity grounding increasingly emphasized in recent MLLM benchmarks (Cheng et al., 2026; Zou et al., 2026; Wang et al., 2026a; Aparcedo et al., 2026; Asgarov et al., 2026). To study this formulation at scale, we construct the accompanying benchmark, also named **Event-STVG**, comprising 4,016 annotated event occurrences spanning 8 domains, 26 event types, and 30 semantic roles. We use soccer as the primary domain because of its visually similar players, consistent scenes, and multi-participant interactions. Specifically, 3,516 soccer events in the 1-minute setting support broad evaluation across models, while 300-event subsets at 3, 5, and 10 minutes enable controlled analysis of how performance changes with increasing temporal context. The remaining 500 events span seven additional domains and are organized into 100 minute-scale, multi-occurrence challenges, allowing us to examine whether the observed difficulties generalize beyond soccer. Comprehensive experiments with state-of-the-art MLLMs and a specialized STVG model reveal that Event-STVG remains highly challenging. The strongest model achieves only 5.4% vIoU with 10-minute context. More importantly, our analyses reveal several recurring bottlenecks across frontier models. As context grows, models increasingly confuse the requested occurrence with other same-type events. Even when the target event is successfully localized, performance deteriorates sharply under role-grounding evaluation. Moreover, models rarely ground all required roles correctly within the same event. **These findings expose a persistent gap between resolving when an event occurs and grounding the structured participants that constitute it.**

Our main contributions are threefold:

(1) We rethink what STVG should evaluate for MLLMs and introduce **Event-STVG**, a task reframing spatio-temporal grounding from description-guided localization toward a test of long-form event reasoning in MLLMs.

(2) We construct a high-quality benchmark, also named **Event-STVG**, comprising 4,016 annotated event occurrences across 8 domains, and comprehensively evaluate frontier MLLMs together with a specialized STVG model, revealing substantial challenges in joint spatio-temporal grounding.

(3) We conduct systematic error analyses across models and domains, identifying **consistent failure patterns** in event-instance selection, spatial role grounding, and complete event grounding. These findings provide insights for improving both general-purpose MLLMs and STVG specialists.

2 Related Work

Table 1: Comparison of Event-STVG with related benchmarks. TG/SG: temporal/spatial grounding; Joint Grounding: joint spatio-temporal grounding; Event–Role Grounding: spatial grounding of event roles. M: manual annotation; A&M: automatic and manual annotation; NR: not reported. Duration reports mean video length. Videos count all evaluation settings.

Benchmark	Videos	Duration (s)	Anno	TG	SG	Joint Grounding	Event–Role Grounding
Video-MME	900	1,017.9	M	✗	✗	✗	✗
LVBench	103	4,101	M	✗	✗	✗	✗
VideoZeroBench	138	667.14	M	✓	✓	✗	✗
CaST-Bench	1,015	13.68	A&M	✓	✓	✓	✗
ST-Evidence	2,896	NR	A&M	✓	✓	✓	✗
VidEvent	1,110	82	M	✓	✗	✗	✗
NEST	1,005	5,872.2	A&M	✓	✗	✗	✗
MultiVENT-G	1,168	83.7	M	✓	✓	✗	✓
VidSTG	6,924	28.01	M	✓	✓	✓	✗
HC-STVG-v1	5,660	20	A&M	✓	✓	✓	✗
BOSTVG	10,018	36.5	M	✓	✓	✓	✗
LF-STVG	6,000	180	M	✓	✓	✓	✗
STVG-CompBench	2,000	NR	A&M	✓	✓	✓	✗
Event-STVG	**4,516**	**121.5**	M	✓	✓	✓	✓

Spatio-Temporal Video Grounding Early STVG benchmarks ground language-referred entities in short videos (Zhang et al., 2020; Tang et al., 2022), spurring continued methodological advances (Yang et al., 2022; Jin et al., 2022; Li et al., 2025; Zhang et al., 2026b; Gao et al., 2026; Rasheed et al., 2026). Meanwhile, benchmark extensions to STVG address multiple targets (Ji et al., 2024; Yao et al., 2026), compositional queries (Wang et al., 2026b), specialized domains (Otsubo et al., 2026; Xu et al., 2025), and minute-scale contexts (Gu et al., 2026). Yet these formulations remain largely description-guided, with instance-specific appearance, action, or relational cues specifying what should be grounded. **Existing STVG therefore primarily evaluates description-conditioned detection and localization rather than more complex event-level reasoning in MLLMs.**

Structured and Grounded Event Understanding Event understanding has long been modeled through predicates, semantic roles, and visual grounding, from imSitu to CineMEC (Yatskar et al., 2016; Pratt et al., 2020; Sadhu et al., 2021; Khan et al., 2022; Darur et al., 2026). These formulations operate on images or pre-segmented short events and thus do not require event temporal grounding. VidEvent and NEST introduce event localization and structured argument understanding but do not spatially ground role fillers (Liang et al., 2025; Asgarov et al., 2026); MultiVENT-G additionally grounds role-specific spatial evidence but separates temporal and spatial retrieval (Sanders et al., 2024). **To our knowledge, no prior benchmark jointly requires temporal grounding of a target event and spatial grounding of its constituent semantic-role fillers.**

Video QA with Spatio-Temporal Evidence General video QA benchmarks primarily evaluate answer correctness (Fu et al., 2025; Zhou et al., 2025; Wang et al., 2025), while recent benchmarks increasingly require visual evidence through temporal localization (Chen et al., 2025; Huang et al., 2026; Zou et al., 2026; Liu et al., 2026) or decomposed spatio-temporal grounding (Deng et al., 2025; Cheng et al., 2026; Meng et al., 2026). More recently, E-VQA jointly grounds objects or regions supporting an answer (Wang et al., 2026a), while CaST-Bench localizes evidence along a causal reasoning chain (Zhang et al., 2026a). **These benchmarks ground evidence supporting an answer or reasoning process, rather than jointly localizing an event and the semantic-role fillers that constitute its structure.**

3 EVENT-STVG

3.1 TASK FORMULATION

Distinct from prior formulations, Event-STVG jointly requires temporal localization of the specified event occurrence and spatial grounding of its constituent roles, testing event–role reasoning in long-form video. The task can be formulated as follows. Given a long-form video V, an event type E may occur repeatedly, forming an ordered sequence of event instances

$$\mathcal{E}(V, E) = \langle E_1, E_2, \ldots, E_n \rangle.$$

Each event type is associated with a semantic role schema

$$\mathcal{R}(E) = \{r_1, r_2, \ldots, r_K\},$$

which specifies the constituent roles involved in the event. Given an event type E, an ordinal index m, and a set of required roles $\mathcal{R}_q \subseteq \mathcal{R}(E)$, Event-STVG asks a model to identify the m-th occurrence E_m and jointly return its temporal interval

$$\tau_m = [t_m^s, t_m^e]$$

together with the spatial trajectory of each queried role within that interval,

$$\mathcal{B}_m = \{\mathbf{B}_{m,r} \mid r \in \mathcal{R}_q\},$$

where $\mathbf{B}_{m,r}$ denotes the sequence of bounding boxes grounding the entity that fills role r. Thus, the query specifies **what event**, **which occurrence**, and **which roles**, but provides no instance-specific description of the participants, their appearance, or how the event unfolds.

To reduce ambiguity in annotation and evaluation, we define event-specific temporal boundaries using canonical role actions, specifying only when an event begins and ends rather than how its constituent roles interact. For example, a soccer shot may be defined from possession by the last passer to the point at which the shot outcome becomes clear. Different occurrences may nevertheless realize the same event structure differently, such as through a cross followed by a header or a through ball followed by a shot. These instance-specific interaction patterns are neither encoded in the query nor prescribed by the event definition, and must therefore be inferred from the video to resolve the target occurrence and ground its constituent roles.

3.2 BENCHMARK CONSTRUCTION

Data Sources and Scope Under this formulation, Event-STVG naturally applies to diverse real-world scenarios in which events recur over time with shared role structures but different participants and realizations. Examples include grounding the participants in a particular restaurant order, the entities in a specific traffic accident, or the part produced in a given manufacturing cycle. Considering both the availability of existing video datasets and our emphasis on structured multi-participant activities, we instantiate our benchmark primarily with soccer, complemented by basketball, volleyball, ice hockey, animal hunting, fast-food packing, entertainment programs, and esports as shown in Figure 3. The videos are drawn from SoccerNet-v2 (Deliège et al., 2021), PL-NBA (Zhao et al., 2026), VNL-STES (Nguyen et al., 2025), VIP-HTD (Prakash et al., 2024), MammalNet (Chen et al., 2023), and publicly available YouTube videos.

Human Annotation The annotation of soccer follows the process illustrated in Figure 3. We first use SoccerNet-v2 event annotations to automatically retrieve candidate occurrences. Annotators then inspect each candidate segment and manually determine its start and end according to event-specific boundary definitions. To annotate role trajectories efficiently, we use SoccerMaster (Yang et al., 2026) together with tracking and re-identification tools to generate candidate detections and track identities, which are presented to annotators at 2 FPS. Annotators bind tracks to event-specific semantic roles such as *shooter*, *last passer*, and *goalkeeper*, while retaining full control to reassign

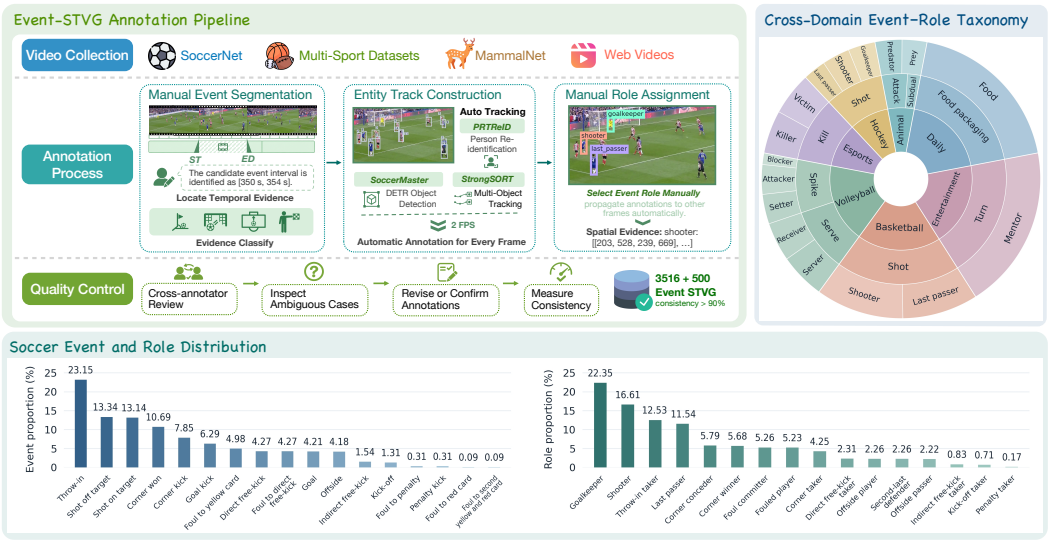


Figure 3: Overview of Event-STVG benchmark construction and composition. We construct Event-STVG through carefully designed human annotation with automated tracking assistance and multi-stage quality control. The resulting benchmark exhibits substantial diversity in event types and semantic roles, covering major soccer events and seven additional domains.

tracks, remove erroneous boxes, or manually add missing boxes. Thus, automatic tracking serves only as annotation assistance, while all ground-truth trajectories are human-verified. Starting from these annotations, we construct evaluation contexts ranging from oracle event windows to contexts of up to 10 minutes. Occurrences of each event type are temporally ordered within each context to derive ordinal indices, after which Event-STVG queries are generated from the event type, ordinal index, and required roles.

For the remaining domains, since source annotations and tracking availability vary across domains, we use a source-agnostic human annotation procedure. Annotators manually determine event boundaries and ground semantic roles either by assigning available tracking boxes or by manually drawing boxes on video frames sampled at 2 FPS. This yields 500 events with temporal intervals and role trajectories. We organize them into 100 minute-scale evaluation videos by extracting continuous intervals from source videos when possible or, otherwise, concatenating related clips from the same game or from videos of the same animal species. Repeated events are then enumerated to construct ordinal event-role queries following the task formulation.

Quality Control We independently cross-annotate approximately 10% of the data for both temporal segmentation and role grounding. The paired annotations achieve a mean temporal IoU of 0.9039 for event intervals and a mean spatial IoU of 0.9161 when averaged equally over sample–role pairs. We further sample approximately 5% of the data and inspect the corresponding annotations to verify the correctness of ordinal indices and generated prompts, with all the inspected annotations passing the checks. Appendix E provides further details on annotation and quality control.

3.3 Benchmark Statistics

As shown in Figure 3, our benchmark comprises **4,016 annotated event instances** across soccer and seven additional domains. Together, they contain **26 event types**, **30 semantic roles**, **7,323 event–role pairs** and **59,381 annotated bounding boxes**, providing dense temporal and spatial supervision for event occurrence selection and role grounding. Notably, **58.91%** of all event instances involve multiple semantic roles, making the joint grounding of multiple role fillers a common requirement. Ordinal disambiguation is also central to Event-STVG. The proportion of soccer contexts containing repeated same-type events increases from **13.68%** at one minute to **68.69%** at ten minutes, while **99.40%** of queries in the remaining domains contain repeated same-type occurrences.

4 EXPERIMENT

4.1 SETTINGS

Evaluated Models We evaluate ten vision-language models, including nine general-purpose VLMs and one STVG specialist: Gemini-3.7 Flash (Google DeepMind, 2026), GPT-5.6 Sol (OpenAI, 2026), Qwen-3.8 Max (Qwen Team, 2026c), Qwen-3.7 Plus (Qwen Team, 2026b), Qwen-3.5 9B (Qwen Team, 2026a), GLM-5.3 Flash (Z.ai, 2026), GLM-5V Turbo (GLM-V Team, 2026), GLM-4.6V (GLM-V Team, 2025), MiniMax-M3 (Lai et al., 2026), and PTD with its released Qwen3-VL-4B checkpoint (Rasheed et al., 2026). This selection spans leading proprietary models, open-weight models, locally deployable models, and a task-specialized STVG model.

Evaluation Protocol For soccer, we evaluate all 3,516 events under the oracle and 1-minute settings. Under the oracle setting, models are given the ground-truth event interval and evaluated only on role grounding, isolating spatial grounding from temporal retrieval. These oracle windows are therefore not counted as complete Event-STVG videos in Table 1. Building on the 1-minute setting, which provides the primary short-context evaluation of end-to-end Event-STVG, we further evaluate a fixed stratified subset of 300 events under each of the 3-, 5-, and 10-minute settings to assess long-context robustness. Finally, 100 cross-domain evaluation videos assess generalization beyond soccer. In total, this gives 4,516 non-oracle evaluation videos across the benchmark. Prompt templates and model evaluation details are provided in Appendices B and C, respectively.

Metrics We report mean temporal IoU (m-tIoU) (Zhang et al., 2020; Tang et al., 2022; Yang et al., 2022) for event intervals and mean video IoU (m-vIoU) for role trajectories. Event-level vIoU (Zhang et al., 2020) averages tube IoU over queried roles; m-vIoU then averages over events. We also report $\mathrm{vIoU@0.3}$ and $\mathrm{vIoU@0.5}$, the proportions of events above each event-level threshold. Missing or invalid predictions contribute zero. Appendix A gives the equations and scoring rules.

4.2 MAIN RESULTS

Table 2: Performance on the Event-STVG benchmark (3,516 events in soccer; %) under the Oracle and 1-minute settings. **Bold** indicates the best result in each column.

	Oracle			1 minute			
Model	m-vIoU	vIoU@0.3	vIoU@0.5	m-tIoU	m-vIoU	vIoU@0.3	vIoU@0.5
General-Purpose VLMs							
Gemini-3.7 Flash	**24.18**	**38.40**	**11.35**	43.39	**10.50**	**12.66**	**3.56**
GPT-5.6 Sol	13.38	13.99	5.49	**48.23**	2.27	0.43	0.11
Qwen-3.8 Max	11.42	12.43	2.84	46.46	3.95	1.34	0.11
Qwen-3.7 Plus	2.73	0.71	0.09	38.47	0.42	0.00	0.00
Qwen-3.5 9B	2.52	1.91	0.37	15.82	0.08	0.00	0.00
GLM-5.3 Flash	12.75	14.56	2.42	31.38	0.74	0.06	0.00
GLM-5V Turbo	10.32	10.17	2.89	17.28	0.40	0.03	0.00
GLM-4.6V	11.51	11.38	2.62	6.35	0.13	0.00	0.00
MiniMax-M3	0.26	0.00	0.00	4.18	0.02	0.00	0.00
STVG-Specialized VLMs							
PTD (Qwen3-VL-4B)	13.19	12.68	1.99	15.98	1.72	0.26	0.00

Short-context Event-STVG is already challenging for current models. Table 2 shows that even with the ground-truth event interval, performance remains far from saturated: grounding the event's constituent roles is difficult. The task is substantially harder in the 1-minute setting, where six of the nine general-purpose models obtain m-vIoU below 1%, and never achieve $\mathrm{vIoU@0.5}$ on any event.

The STVG-specialized baseline PTD also struggles, achieving only 1.72% m-vIoU at 1 minute. Moreover, model strengths differ markedly across the two components. GPT achieves the highest temporal localization performance at 1 minute, whereas Gemini is substantially stronger in joint spatio-temporal grounding. Given the near-floor performance of most models and limited frame budget of PTD, we restrict the subsequent long-context and multi-domain evaluations to Gemini, GPT, and Qwen.

Table 3: Performance on the fixed 300-query Event-STVG analysis subset (300 events in soccer; %) under the 3-, 5-, and 10-minute settings. **Bold** indicates the best result in each column.

	3 minutes		5 minutes		10 minutes	
Model	m-tIoU	m-vIoU	m-tIoU	m-vIoU	m-tIoU	m-vIoU
Gemini-3.7 Flash	33.21	**11.36**	25.14	**7.37**	21.54	**5.40**
GPT-5.6 Sol	**43.16**	1.55	**41.22**	1.53	**36.82**	1.29
Qwen-3.8 Max	39.00	2.00	32.67	0.94	24.57	0.80

Longer contexts further exacerbate the difficulty. Table 3 shows that both temporal and joint grounding performance decrease as context grows. At 10 minutes, the best m-vIoU among the three models is only 5.40%. GPT retains the strongest temporal localization, whereas Gemini remains substantially stronger in joint grounding. No model therefore dominates both components, revealing that stronger event localization does not translate directly into stronger end-to-end grounding. This persistent divergence suggests that long-form Event-STVG requires models to jointly maintain event-instance resolution and fine-grained role grounding rather than solving either component in isolation.

Table 4: Performance on Event-STVG (500 events in multiple domains; %). Domains are ranked by mean m-vIoU across the three models.

		Gemini-3.7 Flash		GPT-5.6 Sol		Qwen-3.8 Max		Mean
Domain	Proportion (%)	m-tIoU	m-vIoU	m-tIoU	m-vIoU	m-tIoU	m-vIoU	m-vIoU
Daily work	20	53.97	21.84	53.04	28.35	45.29	13.71	21.30
Entertainment	20	25.07	16.82	27.27	19.90	15.36	9.21	15.31
Animal activity	3	20.65	7.83	36.14	8.06	19.30	5.12	7.00
Volleyball	18	34.75	8.47	44.95	3.57	31.82	5.23	5.76
Esports	10	42.87	7.24	47.65	2.97	35.24	0.48	3.56
Basketball	20	13.50	2.73	23.04	1.61	18.38	1.27	1.87
Ice hockey	9	3.44	0.33	13.20	1.28	4.09	0.46	0.69

Event-STVG remains challenging across domains. Table 4 shows substantial variation across domains, yet performance remains limited overall. Averaged over the three models, even the best-performing domain reaches only 21.3% m-vIoU, while the lowest is just 0.69%. The relative ordering of domains also reveals a meaningful pattern, as daily work and entertainment, which primarily involve simpler single-target actions, are comparatively easier than domains dominated by rapid multi-agent interactions, such as volleyball, esports, basketball, and ice hockey. Despite these differences in visual content and event structure, none of the domains approaches saturated performance. These results indicate that the challenges exposed by Event-STVG are not specific to soccer, but recur across substantially different event types and visual environments. The consistently low performance across team-sport domains further positions soccer as a representative and demanding main testbed for structured event grounding.

4.3 Error Analysis

We analyze the low performance of end-to-end grounding through three hierarchical bottlenecks: (1) *event-instance resolution*, where the model must distinguish the queried occurrence from other instances of the same event type; (2) *spatial role grounding*, where the model must correctly localize the semantic role fillers after identifying the event temporally; and (3) *multi-role composition*, where the model must jointly ground all required role fillers to recover the complete event structure. These bottlenecks form increasingly demanding levels of Event-STVG, from resolving the target event occurrence, to grounding its constituent roles, and finally to composing them into complete event-level evidence. The first bottleneck is particularly pronounced in sports scenes with repeated same-type events, while the latter two remain prominent across domains.

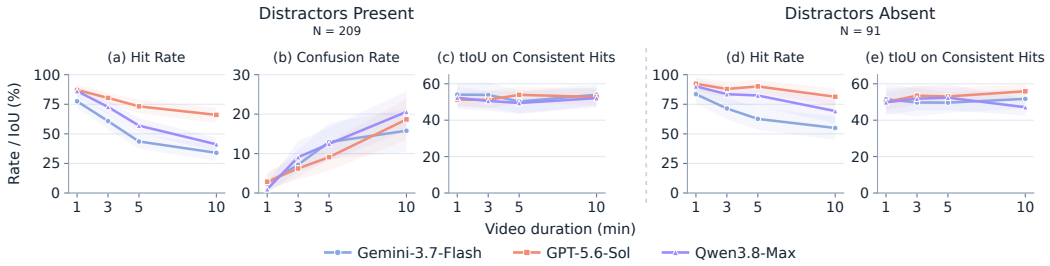


Figure 4: Bottleneck I: temporal instance resolution across video durations, with and without same-type distractors. The panels report target hit rate, instance confusion rate, and temporal intersection over union (tIoU) on predictions that consistently hit the target across all durations in soccer videos.

Bottleneck I: Resolving the target occurrence among repeated events Longer inputs primarily impair event-instance resolution rather than boundary localization after the target occurrence has been identified in soccer videos. As the input becomes longer, models increasingly select another valid occurrence of the same event type instead of the annotated target. Correspondingly, target hit rates decrease and instance-confusion rates increase significantly for all three models when same-type distractors are present (all Holm-adjusted $p < 0.0001$). Hit rates also decline when no same-type distractor is annotated, although the decline remains significant after correction only for Gemini and Qwen and is smaller than that in the distractor-present group. Intriguingly, among queries for which a model hits the target at every input duration, tIoU does not significantly decline as the video becomes longer. This suggests that models can still determine reasonable temporal boundaries once the intended occurrence has been resolved. The same bottleneck extends beyond soccer: for later targets in basketball, the average confusion rate reaches 41.81% across the three models, indicating substantial interference from repeated occurrences. Thus, the primary long-context bottleneck lies not simply in determining *when* an event occurs, but in resolving *which* occurrence the query refers to.

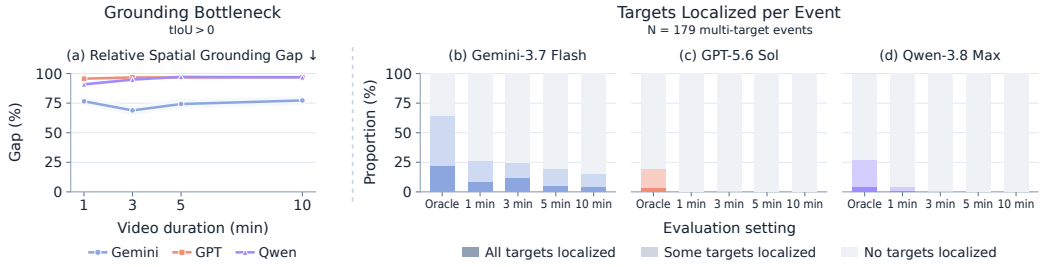


Figure 5: Bottleneck II and III: spatial and multi-target grounding bottlenecks. (a) Relative Spatial Grounding Gap among temporally overlapping predictions. (b–d) Proportions of multi-target events for which each model localizes all, some, or none of the required role fillers under Oracle and different video-duration settings, using per-role full-trajectory vIoU > 0.3 as the localization criterion.

Bottleneck II: Spatial role-grounding failure despite temporal localization Temporal overlap does not guarantee that the entities participating in the event are grounded correctly. This failure occurs across all domains. A model may identify when an event happens but assign boxes to the wrong entity, produce inaccurate role trajectories, or fail to track a role throughout the event. To isolate this spatial deficit from temporal localization, we define the Relative Spatial Grounding Gap as the relative shortfall from perfect spatial grounding under the same predicted temporal support:

$$\begin{aligned} \mathrm{tIoU}_{\mathrm{disc}} &= \frac{|I|}{|U|} \qquad \mathrm{vIoU}_{\mathrm{disc}} = \frac{1}{|U|}\sum_{t\in I}\mathrm{IoU}_t, \\ \mathrm{Gap}_{\mathrm{SG}} &= 1 - \frac{\mathrm{vIoU}_{\mathrm{disc}}}{\mathrm{tIoU}_{\mathrm{disc}}} = 1 - \frac{1}{|I|}\sum_{t\in I}\mathrm{IoU}_t. \end{aligned} \tag{1}$$

Here, I and U denote the intersection and union of the predicted and ground-truth temporal supports on the evaluation sampling grid, and IoU_t denotes the spatial role-box IoU at frame t. We compute the gap only for predictions with $\mathrm{tIoU}_{\mathrm{disc}} > 0$ and then average it across queries. As shown in Fig. 5(a), the spatial grounding gap in soccer remains large across input durations for all three models, showing that temporal overlap does not imply accurate role grounding. Across the remaining domains, the average gap reaches 71.12%, corresponding to only 28.88% mean spatial IoU within temporally overlapping frames. These results identify spatial role grounding as a distinct bottleneck in Event-STVG, even after the target event has been localized in time.

Bottleneck III: Complete multi-role composition A further bottleneck lies in jointly grounding all required role fillers. Figures 5(b–d) show that partial and failed role coverage remain common under both Oracle and long-video settings in soccer, and the same pattern extends across domains, with an average complete-success rate of only 0.73%. These results suggest that models can recover individual participants without successfully assembling the full event structure. This compositional bottleneck persists even after the correct event occurrence and its temporal extent have been identified, because end-to-end success requires every queried role to be correctly localized and tracked. Taken together, the three bottlenecks form a hierarchy from resolving the target occurrence, to grounding its constituent roles, and finally to composing them into complete event-level evidence.

5 Conclusion

In this paper, we introduce **Event-STVG**, a new formulation of spatio-temporal video grounding that shifts description-guided localization toward structured event reasoning. To this end, we construct a benchmark with 4,016 manually annotated event occurrences across eight domains, which requires models to resolve a target occurrence among repeated events, infer its semantic role fillers, and jointly ground them in space and time. Our experiments show that current models remain far from solving Event-STVG, with the best model achieving only 5.4% m-vIoU under 10-minute context. The analyses further reveal three key challenges: (1) longer contexts hinder event-instance resolution, especially with repeated same-type events; (2) temporal localization does not ensure accurate spatial role grounding; and (3) complete grounding becomes substantially harder with multiple semantic roles. Event-STVG thus moves beyond viewing MLLMs merely as multimodal alignment tools toward evaluating the deeper reasoning required for structured event understanding. More broadly, its verifiable grounding signals may also provide fine-grained feedback for reinforcement learning–based post-training toward stronger visual reasoning. We hope Event-STVG can serve as a testbed for advancing long-form event reasoning through verifiable spatio-temporal grounding.

AI Use Statement

Generative AI tools assisted with literature research, writing, and coding. They were not used to generate or annotate benchmark ground truth. All AI-assisted outputs were verified by the authors.

Reproducibility Statement

Detailed benchmark construction, annotation, prompting, and evaluation protocols are provided in the appendix. Code and annotations will be released.

REFERENCES

Alejandro Aparcedo, Akash Kumar, Aaryan Garg, Dalton Pham, Wen-Kai Chen, Anirudh Bharadwaj, Aman Chadha, and Yogesh Rawat. VISTA: A Benchmark for Spatio-Temporal Interaction Understanding in Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 8411–8422, 2026. URL https://openaccess.thecvf.com/content/CVPR2026W/PVUW/html/Aparcedo_VISTA_A_Benchmark_for_Spatio-Temporal_Interaction_Understanding_in_Videos_CVPRW_2026_paper.html.

Ali Asgarov, Kaushik Narasimhan, Najibul Haque Sarker, Hani Alomari, Chia-Wei Tang, Anushka Sivakumar, Zaber Ibn Abdul Hakim, Shaurya Mallampati, and Chris Thomas. NEST: Narrative Event Structures in Time for Long Video Understanding. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2026. URL https://arxiv.org/abs/2606.19706.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. CG-Bench: Clue-grounded Question Answering Benchmark for Long Video Understanding. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2412.12075.

Jun Chen, Ming Hu, Darren J. Coker, Michael L. Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, and Mohamed Elhoseiny. MammalNet: A large-scale video benchmark for mammal recognition and behavior understanding. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13052–13061. IEEE, June 2023. doi: 10.1109/CVPR52729.2023.01254. URL https://doi.org/10.1109/CVPR52729.2023.01254.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-STaR: Benchmarking Video-LLMs on Video Spatio-Temporal Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9155–9164, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Cheng_V-STaR_Benchmarking_Video-LLMs_on_Video_Spatio-Temporal_Reasoning_CVPRF_2026_paper.html.

Balaji Darur, Amanmeet Garg, and Makarand Tapaswi. One Identity, Many Roles: Multimodal Entity Coreference for Enhanced Video Situation Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8268–8279, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Darur_One_Identity_Many_Roles_Multimodal_Entity_Coreference_for_Enhanced_Video_CVPRF_2026_paper.html.

Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4508–4519, June 2021. doi: 10.1109/CVPRW53098.2021.00508. URL https://openaccess.thecvf.com/content/CVPR2021W/CVSports/html/Deliege_SoccerNet-v2_A_Dataset_and_Benchmarks_for_Holistic_Understanding_of_Broadcast_CVPRW_2021_paper.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*, pp. 8625–8636, 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.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Hong Gao, Xiangkai Xu, Bin Zhong, Junjie Yin, Fangyu Kang, Yutong Xu, Xiugang Dong, Xurui Gao, and Min-Ling Zhang. SARL-STG: A Spatially Aware Reinforcement Learning Framework for Refining MLLMs in Spatio-Temporal Video Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24630–24639, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Gao_SARL-STG_A_Spatially_Aware_Reinforcement_Learning_Framework_for_Refining_MLLMs_CVPR_2026_paper.html.

GLM-V Team. GLM-4.5v and GLM-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

GLM-V Team. GLM-5v-turbo: Toward a native foundation model for multimodal agents, 2026. URL https://arxiv.org/abs/2604.26752.

Google DeepMind. Gemini 3.7 Flash: Model card. Technical report, Google DeepMind, August 2026. URL https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-7-Flash-Model-Card.pdf.

Xin Gu, Bing Fan, Jiali Yao, Zhipeng Zhang, Yan Huang, Cheng Han, Heng Fan, and Libo Zhang. Towards Long-Form Spatio-Temporal Video Grounding. In *European Conference on Computer Vision*, 2026. URL https://arxiv.org/abs/2602.23294.

Linpeng Huang, Weixing Chen, Zexin Chen, Yang Liu, and Liang Lin. EG-VQA: Benchmarking Verifiable Video Question Answering with Grounded Temporal Evidence. arXiv preprint arXiv:2606.24797, 2026. URL https://arxiv.org/abs/2606.24797.

Wei Ji, Xiangyan Liu, Yingfei Sun, Jiajun Deng, You Qin, Ammar Nuwanna, Mengyao Qiu, Lina Wei, and Roger Zimmermann. Described Spatial-Temporal Video Detection. arXiv preprint arXiv:2407.05610, 2024. URL https://arxiv.org/abs/2407.05610.

Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing Consistency: A One-Stage Approach for Spatio-Temporal Video Grounding. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2209.13306.

Zeeshan Khan, C. V. Jawahar, and Makarand Tapaswi. Grounded Video Situation Recognition. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/36474dd150d970be6fcb17035ecc4dce-Abstract-Conference.html.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, Jinkai Hu, Jiayao Li, Rui Gao, Zekun Li, Songquan Zhu, Jingkai Zhou, and Pengyu Zhao. MiniMax sparse attention, 2026. URL https://arxiv.org/abs/2606.13392.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. LLaVA-ST: A Multimodal Large Language Model for Fine-Grained Spatial-Temporal Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2501.08282.

Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. VidEvent: A Large Dataset for Understanding Dynamic Evolution of Events in Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5128–5136, 2025. doi: 10.1609/aaai.v39i5.32544. URL https://ojs.aaai.org/index.php/AAAI/article/view/32544.

Pengyiang Liu, Junbo Niu, Xiaoyang Hu, Zhongyue Shi, Zitian Wang, Linjiang Huang, and Si Liu. TRACE: Temporal Retrieval with Anchored and Convergent Evidence for Long-Horizon Video Understanding. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2608.22516.

Jiahao Meng, Yue Tan, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Haodong Duan, and Yunhai Tong. VideoZeroBench: Probing the Limits of Video MLLMs with Spatio-Temporal Evidence Verification. arXiv preprint arXiv:2604.01569, 2026. URL https://arxiv.org/abs/2604.01569.

Hoang Quoc Nguyen, Ankhzaya Jamsrandorj, Vanyi Chao, Yin May Oo, Muhammad Amrulloh Robbani, Kyung-Ryoul Mun, and Jinwook Kim. VNL-STES: A benchmark dataset and model for spatiotemporal event spotting in volleyball analytics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5852–5861. IEEE, June 2025. doi: 10.1109/CVPRW67362.2025.00584. URL https://doi.org/10.1109/CVPRW67362.2025.00584.

OpenAI. GPT-5.6 system card. Technical report, OpenAI, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6/gpt-5-6.pdf.

Rintaro Otsubo, Ryo Fujii, Reina Ishikawa, Taiki Kanaya, Kanta Sawafuji, Hiroki Kajita, Shigeki Sakai, Hideo Saito, and Ryo Hachiuma. AnyGroundBench: A Specialized-Domain Benchmark for Video Grounding in Vision-Language Models. arXiv preprint arXiv:2607.02269, 2026. URL https://arxiv.org/abs/2607.02269.

Harish Prakash, Yuhao Chen, Sirisha Rambhatla, David A. Clausi, and John Zelek. VIP-HTD: A public benchmark for multi-player tracking in ice hockey. *Journal of Computational Vision and Imaging Systems*, 9(1):22–25, April 2024. doi: 10.15353/jcvis.v9i1.10006. URL https://doi.org/10.15353/jcvis.v9i1.10006.

Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded Situation Recognition. In *Computer Vision – ECCV 2020*, 2020. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123490307.pdf.

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

Qwen Team. Qwen3.7-Plus, 2026b. URL https://www.qwencloud.com/models/qwen3.7-plus. Official QwenCloud model card.

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

Hanoona Rasheed, Haania Siddiqui, Ming-Hsuan Yang, Fahad Shahbaz Khan, and Salman Khan. Locate Anything in Videos: Rethinking Efficient Generative Spatio-Temporal Video Grounding. arXiv preprint arXiv:2608.28192, 2026. URL https://arxiv.org/abs/2608.28192.

Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual Semantic Role Labeling for Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5589–5600, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Sadhu_Visual_Semantic_Role_Labeling_for_Video_Understanding_CVPR_2021_paper.html.

Kate Sanders, Reno Kriz, David Etter, Hannah Recknor, Alexander Martin, Cameron Carpenter, Jingyang Lin, and Benjamin Van Durme. Grounding Partially-Defined Events in Multimodal Data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15905–15927. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.934. URL https://aclanthology.org/2024.findings-emnlp.934/.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric Spatio-Temporal Video Grounding With Visual Transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2022. doi: 10.1109/TCSVT.2021.3085907. URL https://arxiv.org/abs/2011.05049.

Shijie Wang, Honglu Zhou, Ziyang Wang, Ran Xu, Caiming Xiong, Silvio Savarese, Chen Sun, and Juan Carlos Niebles. Evidence-Backed Video Question Answering. In *European Conference on Computer Vision*, 2026a. URL https://arxiv.org/abs/2607.11862.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. LVBench: An Extreme Long Video Understanding Benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wang_LVBench_An_Extreme_Long_Video_Understanding_Benchmark_ICCV_2025_paper.html.

Xingjian Wang, Shijian Wang, Yibo Wang, Zihao Yu, Runhao Fu, Xuelian Cheng, and Zongyuan Ge. Learning Compositional Spatio-Temporal Video Grounding with Synthetic Curriculum. In *Findings of the Association for Computational Linguistics: EMNLP*, 2026b. URL https://arxiv.org/abs/2608.30584.

Qi'ao Xu, Tianwen Qian, Yuqian Fu, Kailing Li, Yang Jiao, Jiacheng Zhang, Xiaoling Wang, and Liang He. ToG-Bench: Task-Oriented Spatio-Temporal Grounding in Egocentric Videos. arXiv preprint arXiv:2512.03666, 2025. URL https://arxiv.org/abs/2512.03666. Version 2, revised April 2026.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tube-DETR: Spatio-Temporal Video Grounding With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16442–16453, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_TubeDETR_Spatio-Temporal_Video_Grounding_With_Transformers_CVPR_2022_paper.html.

Haolin Yang, Jiayuan Rao, Haoning Wu, and Weidi Xie. SoccerMaster: A vision foundation model for soccer understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21549–21560, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_SoccerMaster_A_Vision_Foundation_Model_for_Soccer_Understanding_CVPR_2026_paper.html.

Jiali Yao, Xin Gu, Xinran Deng, Mengrui Dai, Bing Fan, Zhipeng Zhang, Yan Huang, Heng Fan, and Libo Zhang. OmniSTVG: Toward Spatio-Temporal Omni-Object Video Grounding. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=azcQJtcYTE.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation Recognition: Visual Semantic Role Labeling for Image Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5534–5542, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Yatskar_Situation_Recognition_Visual_CVPR_2016_paper.html.

Z.ai. GLM-5.3-flash, 2026. URL https://huggingface.co/zai-org/GLM-5.3-Flash. Official model card.

Mingfang Zhang, Jingjing Pan, Ashutosh Kumar, Rajat Saini, Mustafa Erdogan, Hsuan-Kung Yang, Caixin Kang, Yifei Huang, Yoichi Sato, and Quan Kong. CaST-Bench: Benchmarking Causal Chain-Grounded Spatio-Temporal Reasoning for Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026a. URL https://arxiv.org/abs/2605.23216.

Xiaowen Zhang, Zhi Gao, Licheng Jiao, Lingling Li, and Qing Li. STVG-R1: Incentivizing Instance-Level Reasoning and Grounding in Videos via Reinforcement Learning. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ee6630dcbcff857026e474fc857aa9f0-Abstract-Conference.html.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where Does It Exist: Spatio-Temporal Video Grounding for Multi-Form Sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. URL https://arxiv.org/abs/2001.06891.

Yunhao Zhao, Haoying Sun, Jiarui Li, Zhuming Wang, Ya Jing, Xiangbo Shu, Lifang Wu, and Changwen Chen. PL-NBA: A possession-level universal basketball video dataset supporting multiple visual understanding tasks, 2026. URL https://arxiv.org/abs/2608.19646.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking Multi-task Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_MLVU_Benchmarking_Multi-task_Long_Video_Understanding_CVPR_2025_paper.html.

Yuanhao Zou, Arthad Kulkarni, Lucas Toñanez, Lincoln Spencer, Guangyu Sun, Tianxingjian Ding, Andong Deng, Yi Li, Shuangjun Liu, Yuan Li, Dashan Gao, Ning Bi, Taotao Jing, Shuai Zhang, and Chen Chen. Conditional Multi-Event Temporal Grounding in Long-Form Video. arXiv preprint arXiv:2606.15320, 2026. URL https://arxiv.org/abs/2606.15320.

APPENDICES

A EVALUATION METRICS

For event i among the N evaluation events, let $\hat{\mathcal{T}}_i$ and $\mathcal{T}_i$ denote the predicted and ground-truth intervals. We compute

$$\mathrm{tIoU}_i = \frac{|\hat{\mathcal{T}}_i \cap \mathcal{T}_i|}{|\hat{\mathcal{T}}_i \cup \mathcal{T}_i|}, \qquad \text{m-tIoU} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{tIoU}_i. \tag{2}$$

For role r, let $S^I_{i,r}$ and $S^U_{i,r}$ be the intersection and union of the sampled-frame supports of the predicted and ground-truth tubes, and let $\hat{b}^t_{i,r}$ and $b^t_{i,r}$ be their boxes at frame t. We extend single-target vIoU (Zhang et al., 2020; Tang et al., 2022) to multiple roles as

$$\begin{aligned} \mathrm{vIoU}_{i,r} &= \frac{1}{|S^U_{i,r}|}\sum_{t\in S^I_{i,r}} \mathrm{IoU}(\hat{b}^t_{i,r}, b^t_{i,r}), \\ \mathrm{vIoU}_i &= \frac{1}{R_i}\sum_{r=1}^{R_i} \mathrm{vIoU}_{i,r}, \qquad \text{m-vIoU} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{vIoU}_i, \end{aligned} \tag{3}$$

Here R_i is the number of queried roles. Missing or invalid predictions contribute zero, frames absent from both tubes are excluded, roles are equally weighted within each event, and events are equally weighted overall. The $\mathrm{vIoU@0.3}$ and $\mathrm{vIoU@0.5}$ scores are the proportions of events whose event-level vIoU exceeds the respective thresholds.

B PROMPT TEMPLATES AND OUTPUT FORMAT

We query each model with an event type, an ordinal occurrence, event-specific temporal boundaries, and the roles to ground. The following is the prompt template for a 10-minute soccer video and the *Shot on Target* event. The ordinal placeholder is replaced by the requested occurrence index.

> You will see a 10-minute soccer match video. Find the {ordinal} occurrence of Shot on Target. The event starts when the last passer first gains possession during the target attack, and ends when the outcome of the shot becomes clear. Draw bounding boxes for the last passer, the shooter, and the defending goalkeeper.
>
> **Instructions:**
>
> - Report the event's start and end times in seconds relative to the input video.
> - Starting at the event's start time, provide role boxes every 0.5 seconds within the event window.
> - For each sampled timestamp, include all three roles: `last_passer`, `shooter`, and `goalkeeper`.
> - Use tight bounding boxes in `[xmin, ymin, xmax, ymax]` format, normalized to $[0, 1000]$, with the origin at the top-left of the frame.
> - Track the same person for each role throughout the event. If that person is not reliably visible, output `null`; resume tracking the same person when visible again.
> - Return only JSON with `temporal_window` (`start_sec`, `end_sec`) and `role_localization` (`time_sec` and each role's box or `null`).

The required output is a JSON object with one temporal window and a list of role localizations. The example below illustrates the structure and coordinate format; its values are schematic.

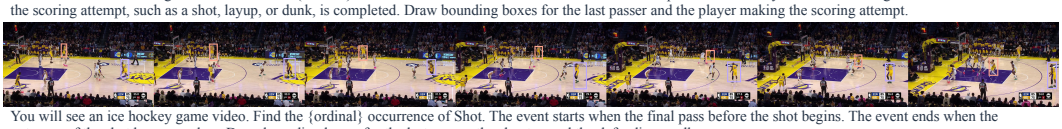
```
{
  "temporal_window": {
    "start_sec": 123.0,
    "end_sec": 123.5
  },
  "role_localization": [
    {
      "time_sec": 123.0,
      "last_passer": [120, 240, 180, 500],
      "shooter": null,
      "goalkeeper": [810, 210, 860, 470]
    },
    {
      "time_sec": 123.5,
      "last_passer": [125, 240, 185, 500],
      "shooter": [420, 220, 475, 490],
      "goalkeeper": [805, 210, 855, 470]
    }
  ]
}
```

Figure 6 shows event-specific prompt examples and role-box annotations across all eight benchmark domains.

You will see a soccer match video. Find the {ordinal} occurrence of Shot on Target. The event starts when the last passer first gains possession during the target attack. The event ends when the outcome of the shot becomes clear. Draw bounding boxes for the last passer, the shooter, and the defending goalkeeper.



You will see a basketball game video. Find the {ordinal} occurrence of Shot. The event starts when the final pass made directly to the shooter begins. The event ends when the scoring attempt, such as a shot, layup, or dunk, is completed. Draw bounding boxes for the last passer and the player making the scoring attempt.



You will see an ice hockey game video. Find the {ordinal} occurrence of Shot. The event starts when the final pass before the shot begins. The event ends when the outcome of the shot becomes clear. Draw bounding boxes for the last passer, the shooter, and the defending goalkeeper.



You will see a volleyball match video. Find the {ordinal} occurrence of Spike. The event starts when the final set before the attack begins. The event ends when the block is completed. Draw bounding boxes for the setter, the attacker, and the opposing blocker.



You will see a League of Legends gameplay video. Find the {ordinal} occurrence of Kill. Count kills in the order of the victims' deaths. The event starts when the victim first loses health in the combat that leads to its death. The event ends when the victim dies. Draw bounding boxes for the killer and the victim in the main gameplay view.



You will see a wildlife hunting video. Find the {ordinal} occurrence of Final Subdual. The event starts when the final phase of the attack begins. The event ends when the prey stops visibly struggling and becomes immobilized. Draw bounding boxes for the predator that subdues the prey and the prey being subdued.



You will see a food-packaging video. Find the {ordinal} occurrence of Food Packaging. The event starts when the process of packaging and sealing the prepared food begins. The event ends when the packaging is sealed. Draw bounding boxes for the food item, including its enclosing wrapper or container.



You will see a performance video from The Voice. Find the {ordinal} occurrence of Turn. The event starts when the mentor's chair begins turning for the contestant. The event ends when the turn is completed. Draw bounding boxes for the mentor who turns, rather than the chair alone.



Figure 6: Cross-scenario prompt examples in Event-STVG. Each row shows an event-specific query and sampled frames with role bounding boxes, spanning soccer, basketball, ice hockey, volleyball, esports, wildlife hunting, food packaging, and entertainment.

C Evaluation Protocol and Implementation Details

This appendix provides additional details on the evaluation protocol, model input configurations, failure handling, and supplementary experiments used in Event-STVG. Unless otherwise stated, all metrics are reported as percentages. For paired comparisons, we use 10,000 paired bootstrap resamples with a fixed seed of 21 and report two-sided 95% percentile confidence intervals.

C.1 Evaluation Settings

Soccer benchmark. The main soccer benchmark contains 3,516 target event queries with 6,497 required role instances. The Oracle and 1-minute settings use the same target events. In the Oracle setting, the model is directly given the ground-truth event clip and is evaluated only on grounding the required semantic roles, thereby isolating spatial role grounding from temporal event retrieval. In the 1-minute setting, the model receives a frozen one-minute context and must identify the queried occurrence according to its ordinal index, predict the event interval, and ground all required role fillers.

For the long-context evaluation, we use a fixed stratified subset of 300 target events and evaluate the same targets under 3-, 5-, and 10-minute contexts. Because enlarging the temporal context may introduce additional occurrences of the same event type, ordinal indices are recomputed independently for each context duration. The corresponding frozen query for each duration is used directly during inference.

Multidomain evaluation. The multidomain extension contains 500 queries over 100 videos and is evaluated independently from the soccer benchmark. It uses the same Event-STVG formulation but domain-specific event and role definitions.

Aggregation. The basic evaluation unit is a query. For a query requiring multiple semantic roles, role-level grounding scores are first averaged equally across the required roles, after which query-level scores are averaged over the evaluation set. Different experimental subsets, such as the 300-query long-context set and the supplementary robustness subsets described below, are therefore reported separately rather than pooled.

C.2 Model Inputs and Inference Configuration

Different models expose different native video interfaces. We therefore follow each model's supported interface while keeping the semantic task specification fixed whenever possible. The main configurations are summarized in Table C.1.

Model / interface	Video input	Inference setting
GPT-5.6 Sol	2 FPS JPEG frames, 512×288, quality 80	`detail=low`; provider generation defaults
Gemini-3.7 Flash (1 min)	2 FPS JPEG frames, 512×288, quality 85	Google GenAI interface; provider generation defaults
Qwen / GLM / MiniMax API models	Native MP4 video input	Provider generation defaults
Qwen3.5-9B API baseline	Native MP4 video input	Fixed OpenRouter/SiliconFlow route; provider generation defaults
PTD (Qwen3-VL-4B)	Uniformly sampled 64 frames	FP16, SDPA, greedy decoding

Table C.1: Main video input and inference protocols. Provider-native video models receive the original MP4 rather than a client-side reconstruction of the 2 FPS frame sequence.

For all hosted API models, optional generation parameters such as temperature, top-p, and random seed are left unspecified. Thus, "provider defaults" indicates that these values are determined by the serving platform and should not be interpreted as temperature $= 0$. The seed used for dataset sampling and bootstrap analysis is not sent to the models.

For GPT-5.6 Sol, videos are converted to ordered JPEG frames at 2 FPS and 512×288 resolution. The main soccer experiments use at most 1,200 frames, which covers the full 10-minute context at 2 FPS. The multidomain evaluation permits up to 1,300 frames to accommodate the longest videos. Each frame is accompanied by its corresponding timestamp.

All inference outputs are retained before parsing. Failed requests, malformed responses, and missing predictions remain part of the predefined evaluation set rather than being removed post hoc.

C.3 Parsing and Scoring Rules

We use the same frozen parser and evaluator for all general-purpose models. A response is considered parseable when a unique complete JSON prediction can be extracted from the returned text. We allow a single JSON object enclosed in a code block or embedded in surrounding text, but do not repair malformed brackets, choose among multiple alternative answers, reorder coordinates, or clip invalid coordinates.

A valid bounding box must contain four finite values in `xyxy` order, lie within $[0, 1000]$, and have positive width and height. A valid predicted temporal interval must satisfy

$$0 \leq t_s \leq t_e \leq T,$$

where T is the duration of the input video.

Importantly, temporal and spatial components are evaluated independently whenever possible. For example, a prediction with a valid temporal interval but missing role boxes can still receive a non-zero temporal IoU while receiving zero spatial credit. Conversely, an invalid temporal interval receives zero temporal and joint spatio-temporal scores. This component-wise evaluation avoids discarding otherwise valid portions of partially structured predictions.

Temporal IoU is computed between the predicted and ground-truth event intervals. For role grounding, the joint vIoU is computed over the union of the ground-truth and predicted temporal supports, such that missing boxes, spatial errors, and temporal over-extension are all penalized. Scores for multiple required roles are averaged equally within each query.

Failed API requests and unparseable responses are assigned zero for the affected components and remain in the original evaluation denominator. We never select predictions according to their ground-truth score.

C.4 Repeatability under Default Inference Settings

Hosted API models may exhibit run-to-run variation even when the request configuration is unchanged. To quantify this effect, we sample a fixed set of 352 soccer queries and repeat inference twice in addition to the original run. The same query IDs are used for Oracle and 1-minute evaluation. We perform this experiment for GLM-4.6V, MiniMax-M3, Qwen3.7-Plus, and Qwen3.8-Max. PTD is evaluated on the same subset as a deterministic local reference.

Table C.2 reports results on the full 352-query set, including failed requests.

	Oracle m-vIoU			**1-min m-tIoU**			**1-min m-vIoU**		
Model	R0	R1	R2	R0	R1	R2	R0	R1	R2
GLM-4.6V	10.889	7.592	6.481	7.550	4.658	4.190	0.159	0.119	0.071
MiniMax-M3	0.262	0.269	0.198	3.552	5.154	3.998	0.019	0.026	0.003
Qwen3.7-Plus	2.518	2.654	2.235	38.095	38.346	41.125	0.383	0.418	0.616
Qwen3.8-Max	11.548	12.208	11.389	47.726	46.905	44.702	3.384	3.057	3.756
PTD-64f	13.736	13.736	13.736	15.719	15.719	15.719	1.659	1.659	1.659

Table C.2: Repeated evaluation on a fixed 352-query subset. R0 denotes the original run and R1–R2 denote two additional runs under the same inference configuration. All values are percentages and failures remain in the denominator.

The two Qwen models show relatively small Oracle variation, although their 1-minute results are not identical across runs. For example, Qwen3.8-Max changes from 47.726% to 44.702% m-tIoU between R0 and R2, corresponding to a paired difference of -3.024 pp with a 95% bootstrap interval of $[-5.600, -0.439]$.

The larger variation of GLM-4.6V is primarily associated with service-side input failures rather than changes among successfully parsed predictions. Its Oracle runs contain 0, 135, and 152 API errors, respectively, with the additional failures reported as HTTP 400 image-input parsing errors. On the subset successfully parsed in all three runs, Oracle m-vIoU varies only from 12.521% to 13.439%. We therefore retain failures in the end-to-end statistics instead of interpreting the full-score decrease solely as stochastic decoding variation.

PTD produces identical outputs across all three runs under the fixed 64-frame greedy configuration. This establishes repeatability for the evaluated configuration and subset.

C.5 PTD Baseline and Interface Adaptation

PTD is designed around a referring-object localization interface rather than the structured Event-STVG JSON interface. We therefore use its native inference format and convert the resulting special-token representation into the common Event-STVG evaluation format.

For each semantic role, PTD receives a role-specific referring expression derived only from the frozen event and role definition. For example, for a shot-on-target event, the three role queries are:

> `last_passer`: the attacking player who makes the final effective pass to the shooter before the shot
>
> `shooter`: the attacking player who takes the shot
>
> `goalkeeper`: the goalkeeper defending the goal against the shot

Each role is inferred independently using the same 64 uniformly sampled video frames. Oracle clips containing fewer than 64 distinct source frames retain all available frames; otherwise, 64 frames are uniformly sampled. The model uses FP16 inference, SDPA attention, a single GPU worker, a maximum of 1,024 generated tokens, and greedy decoding.

PTD time tokens denote sampled temporal positions rather than absolute seconds. We map these positions back to source-video time using the sampled frame indices and convert the sparse predicted boxes to the 0.5-second evaluation grid by interpolation only between valid neighboring predictions; no temporal extrapolation is performed. Failed roles are treated as missing roles rather than causing the complete query to be discarded.

A semantic difference remains between the native PTD interface and the full Event-STVG query. In particular, the PTD referring expression does not explicitly preserve the ordinal event index or the complete event-boundary specification, and multi-role queries are decomposed into independent role requests. This distinction is immaterial in the Oracle setting, where the target event has already been cropped, but can affect the 1-minute setting when several same-type events occur in the context. To quantify the effect of this interface adaptation, we additionally apply the PTD-style role-by-role query and output protocol to Qwen3.8-Max on the same fixed subset, as analyzed below.

C.6 Controlled Comparison with the PTD-Style Query and Output Adaptation

On the same fixed, event-stratified subset of 100 Event-STVG queries, we compare three conditions. *Qwen-full-prompt* uses the original Event-STVG prompt, including the target ordinal, event-boundary definition, and all required roles, and returns the standard unified JSON. *Qwen-PTD-style* uses the same role-specific referring queries as PTD, queries each role independently, and returns simplified role-level bbox/tube JSON that is converted to the unified Event-STVG evaluation format. This adapts the PTD-style query/output interface without PTD-native decoding: Qwen generates no PTD special tokens. In both Qwen conditions, the original MP4 is sent through Qwen's native video API with default generation settings; Qwen does not use PTD's 64-frame sampling. PTD uses its native 64-frame input and special-token output.

Setting	Condition	N	m-tIoU	m-vIoU
Oracle	Qwen3.8, Event-STVG	100	–	9.448
	Qwen3.8, PTD-style	100	–	8.532
	PTD	100	–	13.040
1 minute	Qwen3.8, Event-STVG	100	53.022	4.214
	Qwen3.8, PTD-style	100	31.341	2.184
	PTD	100	16.165	2.104

Table C.3: PTD-style interface control on the same 100 target queries.

In the paired Qwen comparison, Oracle m-vIoU changes from 9.448% to 8.532% (paired difference -0.917 pp; 95% CI $[-3.491, +1.910]$). At 1 minute, m-tIoU changes from 53.022% to 31.341% (-21.682 pp; $[-27.789, -15.792]$), m-vIoU from 4.214% to 2.184% (-2.031 pp; $[-4.007, -0.278]$), and m-vIoU-Int by -1.692 pp ($[-5.017, +1.546]$). The adaptation has little detectable effect on Oracle spatial grounding, where the target event is isolated, but substantially reduces 1-minute temporal and end-to-end grounding when the requested occurrence must be resolved among repeated events.

On the same subset, PTD reaches 13.040% Oracle m-vIoU versus 8.532% for Qwen3.8 under the PTD-style adaptation. At 1 minute, their m-vIoU is similar (2.104% versus 2.184%), although m-tIoU differs substantially (16.165% versus 31.341%).

This is a PTD-style interface control, not a strictly identical-input model comparison. The adaptation jointly changes explicit ordinal and event-boundary information, joint multi-role output versus role-by-role requests, and output representation and conversion. Qwen and PTD also retain different visual sampling and decoding protocols; these differences should not be attributed solely to model capability.

C.7 Effect of GPT Visual `detail=low`

GPT-5.6 Sol uses `detail=low` for individual video frames in the main experiments. We evaluate whether this choice materially changes Event-STVG performance using a paired experiment on 300 1-minute queries. The two conditions use identical query IDs, prompts, frame sampling, resolution, JPEG quality, and provider route; the only difference is whether `detail=low` is explicitly specified or the `detail` field is omitted. The omitted condition should not be interpreted as explicitly setting `detail=high`.

Detail	N	m-tIoU	m-vIoU	m-vIoU-Int
low	300	45.141	2.211	3.914
omit	300	45.318	2.021	3.349

Table C.4: Paired comparison of GPT visual detail settings on 300 1-minute queries. Both conditions achieve 300/300 parseable responses.

Taking `omit` minus `low`, the paired differences are $+0.176$ pp for m-tIoU (95% CI $[-1.905, +2.224]$), -0.190 pp for m-vIoU ($[-0.619, +0.249]$), and -0.564 pp for m-vIoU-Int ($[-1.229, +0.086]$). We therefore do not observe evidence that `detail=low` reduces performance on this paired 1-minute subset. These confidence intervals quantify uncertainty across sampled queries rather than repeated decoding variance.

We additionally test whether the two configurations can process a complete 10-minute frame sequence. With 1,200 frames sampled at 2 FPS, the `detail=low` request completes successfully, whereas omitting `detail` exceeds the provider-reported context limit in the tested request. Reducing the omitted-detail input to 900 frames still exceeds the limit, while a 500-frame request succeeds but covers only the beginning of the video. We therefore use `detail=low` to retain complete long-video coverage rather than truncating the temporal context.

D Additional Error Analysis

We analyze Event-STVG failures along three recurrent paths: (1) *event-instance confusion*, where the model recognizes the event type but selects the wrong occurrence; (2) *spatial role-grounding failure*, where the target event is localized temporally but its role fillers are grounded poorly; and (3) *incomplete multi-role composition*, where only a subset of the required roles is successfully localized. Together, these analyses cover both the soccer benchmark and the multidomain evaluation.

Unless otherwise specified, results are reported in the order **Gemini-3.7 Flash / GPT-5.6 Sol / Qwen-3.8 Max**.

D.1 Event-Instance Confusion

D.1.1 Soccer

Longer contexts substantially increase the difficulty of resolving the requested event occurrence. We focus on 209 soccer queries for which another qualified occurrence of the same event type appears in at least one context setting. From 1 to 10 minutes, target hit rates decrease by **43.54 / 21.05 / 44.98 percentage points**, while same-type instance mismatch rates increase by **13.40 / 15.79 / 19.62 percentage points**. All six changes remain significant after multiple-comparison correction.

This degradation is driven primarily by losing the target occurrence rather than by progressively worse temporal boundaries once the target is found. Among queries successfully hit at all four durations, mean tIoU shows no consistent decline. For the 209-query group, the 1-to-10-minute changes are only -0.05 / $+1.45$ / -0.21 **percentage points**. This suggests that long contexts primarily challenge *which occurrence to select*, rather than boundary refinement after successful retrieval.

Among queries hit at 1 minute but missed at 10 minutes, **26.04% / 55.36% / 38.61%** overlap another qualified occurrence of the same type, providing further evidence that a substantial fraction of long-context failures involve confusion with another plausible event instance.

D.1.2 Multidomain Evaluation

Instance-selection errors also occur beyond soccer. Among 500 multidomain queries, **106 / 104 / 88** predictions are more consistent with another annotated occurrence of the same event type than with the requested target under our candidate criterion. Such cases appear in all seven evaluated domains.

Repeated events are particularly challenging in several sports domains. In basketball, mean tIoU decreases from **27.48 to 7.50**, **37.57 to 16.79**, and **27.79 to 14.20** when comparing the first three occurrences with the fourth or later occurrences. Over the same split, wrong-instance candidate rates increase from **23.33% to 42.98%**, **16.67% to 39.52%**, and **13.33% to 42.92%**, respectively. Volleyball and ice hockey show similar reductions in later-occurrence tIoU, whereas daily work and entertainment do not exhibit the same consistent pattern.

These results indicate that **event-instance resolution is a recurring difficulty across domains**. However, the soccer duration experiment and the multidomain ordinal analysis capture different effects: later occurrences also covary with video position, event difficulty, and available visual evidence.

D.2 Temporal Retrieval Does Not Ensure Spatial Role Grounding

Successfully locating an event in time does not imply that its participants are correctly grounded.

In soccer, among temporally hit queries, mean $\mathrm{Gap}_{\mathrm{SG}}$ remains high. At 1 minute, the gaps are **76.53% / 95.65% / 90.86%**, and at 10 minutes they remain **77.25% / 97.01% / 96.77%**. Restricting evaluation to queries jointly hit by all three models produces the same overall pattern. Across all four durations, Gemini exhibits a significantly smaller spatial gap than GPT and Qwen on these matched queries.

The phenomenon generalizes beyond soccer. On multidomain queries individually hit by each model, mean $\mathrm{Gap}_{\mathrm{SG}}$ is **67.52% / 71.42% / 74.43%**. When restricted to the same 206 queries jointly hit by all three models, the gaps remain **67.59% / 69.12% / 73.54%**. The effect is especially pronounced in esports: on 30 jointly hit League of Legends queries, mean gaps reach **84.43% / 95.73% / 98.41%**.

The same pattern remains under stronger temporal overlap. On the 26 multidomain multi-role queries for which all three models achieve $\mathrm{tIoU} > 0.5$, mean gaps remain **78.13% / 91.07% / 88.08%**.

Therefore, **temporal retrieval and spatial role grounding constitute distinct bottlenecks**. A model may identify when the requested event occurs while still failing to determine who fills its constituent semantic roles.

D.3 Incomplete Multi-Role Composition

Event-STVG further requires all queried role fillers to be grounded jointly, creating an additional compositional bottleneck beyond recovering individual participants.

In soccer, we evaluate 179 queries requiring two or three roles. Even under the Oracle setting, where the ground-truth event interval is provided, only **40 / 6 / 7** queries achieve $\mathrm{vIoU} > 0.3$ for every required role, corresponding to complete-success rates of **22.35% / 3.35% / 3.91%**. By contrast, at least one role succeeds in **64.25% / 18.99% / 27.37%** of the same queries. Thus, providing the correct temporal window does not remove the difficulty of assembling the complete event-role structure.

The challenge is even more apparent in the multidomain set. Among 274 multi-role queries, at least one role is successfully localized in **20 / 6 / 9** cases, whereas all required roles succeed in only **4 / 2 / 0** cases. The corresponding complete-success rates are only **1.46% / 0.73% / 0.00%**. None of the 52 three-role queries achieves complete success for any model.

The gap persists after conditioning on strong temporal localization. Among the 26 multi-role queries jointly satisfying $\mathrm{tIoU} > 0.5$, Gemini succeeds on at least one role in **10** cases but on all roles in only **2**; GPT achieves **3 versus 2**; and Qwen achieves **5 versus 0**.

These results show that recovering individual participants is not sufficient. Event-STVG requires models to **compose all required role fillers into a complete, spatially grounded event structure**, which remains rare even when the event has already been localized in time.

D.4 Robustness Checks

Several additional analyses confirm that the main observations are not artifacts of a single diagnostic criterion.

First, the increase in soccer instance mismatches remains significant under stricter definitions requiring overlap with another event interval or $\mathrm{tIoU} > 0.3$ with the alternative occurrence. Second, among valid predictions that miss the target at 10 minutes, the observed rates of overlapping another same-type occurrence are **23.91% / 54.93% / 38.05%**, compared with only **2.86% / 2.31% / 3.50%** under a random-position reference preserving prediction-window length. All three differences are significant after correction ($p < 0.0001$).

The temporal degradation also remains when restricting soccer evaluation to queries with valid predicted time windows at every duration. From 1 to 10 minutes, hit rates still decrease by **43.27 / 21.05 / 44.10 percentage points**, and mean tIoU decreases by **21.35 / 11.77 / 25.93 percentage points**.

Finally, the multi-role pattern is robust to the success threshold. Lowering the per-role criterion from $\mathrm{vIoU} > 0.3$ to $\mathrm{vIoU} > 0.1$ increases success rates, while raising it to $\mathrm{vIoU} > 0.5$ decreases them, but complete role coverage remains substantially harder than partial coverage.

D.5 Interpretation

The three error paths should be viewed as complementary diagnostics rather than independent probabilistic factors. Likewise, $\mathrm{Gap}_{\mathrm{SG}}$ captures multiple sources of spatial error, including localization quality, role correspondence, visibility, missing predictions, and output validity, and should not be attributed to a single internal mechanism.

Taken together, the evidence reveals a consistent cascade of difficulties: models must first **resolve the requested occurrence among repeated events**, then **ground its semantic role fillers after temporal retrieval**, and finally **localize all required roles jointly**. Failures arise at every stage, explaining why successful end-to-end Event-STVG remains difficult even for current frontier models.

E Dataset Construction, Annotation, and Quality Control

This appendix provides additional details on the construction, annotation, and quality control of Event-STVG. We first describe the common annotation target and workflow, then detail the annotator-facing guidelines and domain-specific annotation procedures, and finally report the quality-control protocol and quantitative annotation audits. Evaluation metrics are provided in Appendix A, and prompt examples and the output format in Appendix B.

E.1 Annotation Overview

Each Event-STVG instance is defined around a specific event occurrence and contains three forms of ground truth:

1. the temporal interval of the event;
2. the semantic roles involved in the event; and
3. the spatial trajectory of the entity filling each role during the event.

For long-context evaluation, occurrences of the same event type are additionally ordered within the model input to determine the target ordinal.

Across all domains, annotation follows the same high-level workflow:

1. define the event semantics, temporal boundaries, required roles, and occurrence-counting rules;
2. identify the target occurrence and annotate its temporal boundaries;
3. annotate the entities filling the required semantic roles at approximately 2 FPS;
4. validate temporal, spatial, identity-consistency, and visibility annotations;
5. construct the final evaluation context and recompute the target ordinal within that context; and
6. generate the model query and corresponding ground truth from the frozen annotation.

A semantic role is defined by its relation to the target event rather than by the identity of the entity itself. For example, `last passer` refers to the player who directly passes to the shooter for the target shot, rather than to an arbitrary player who happens to make the last visible pass in the surrounding video.

Within an event, the entity assigned to a role must remain identity-consistent across frames. When a role filler cannot be reliably localized in a sampled frame, the spatial annotation is recorded as invisible rather than being assigned to another nearby entity.

Throughout the construction pipeline, we retain mappings between the source video, source frames, timestamps, spatial coordinates, and the corresponding frames in the final model input. This distinction is important because, after cropping, concatenation, or re-encoding, the source-video timeline, the final-input timeline, and the nominal 2-FPS sampling timestamps need not coincide exactly.

E.2 Annotator-Facing Guidelines

To reduce annotation ambiguity, annotators follow a task-specific operational guide that separates the annotation process into three stages: **event-window annotation, role-trajectory annotation, and context/ordinal verification**. Automatic timestamps, detections, and trajectories are treated only as annotation aids; final event boundaries, semantic-role assignments, role trajectories, occurrence sequences, and ordinals must be supported by the video itself.

E.2.1 Stage 1: Event-Window Annotation

Annotators first verify that a candidate clip contains the intended event. A retained event must contain the correct event type and key action, be primarily observable in the standard broadcast view, and avoid unnecessary replay, close-up, or adjacent-event content.

We do not apply a universal temporal-boundary rule across event types. Instead, each event is annotated according to its semantic structure. Table E.1 summarizes the principal soccer rules exposed to annotators.

Table E.1: Soccer event-specific roles and temporal-boundary guidelines.

Event	Roles	Start	End
Throw-in	throw-in player	When the standard view returns and the thrower and ball are visible	When another player first touches or controls the thrown ball
Corner Won	corner winner, corner conceder	When the attacking player creating the corner gains possession	When the defender makes the final touch/block and the ball goes out for a corner
Corner Kick	corner taker	When the corner kick begins in the standard broadcast view	When another player first touches the ball, the ball is cleared/goes out, the view changes, or the role becomes invisible
Goal Kick	goalkeeper	When the goalkeeper begins the goal kick in the standard view	When another player first touches or controls the ball
Offside	offside passer, offsider, last defender	When the passer first gains possession	When the pass to the offside player is completed and the offside decision becomes clear
Indirect Free Kick after Offside	indirect free-kick taker	When the restart begins in the standard view	When another player first touches or controls the ball
Foul leading to Direct Free Kick / Penalty	fouler, fouled player	When the fouled player gains possession	When the foul contact occurs and the decision becomes clear
Direct Free Kick / Penalty	taker, goalkeeper	When the kick begins in the standard view	When the outcome becomes clear
Card-producing Foul	fouler, fouled player	When the fouled player gains possession	When the foul occurs and the decision becomes clear
Shot off Target / Shot on Target / Goal	last passer, shooter, goalkeeper	Usually when the last passer gains possession; if no clear final pass exists, from the shooter's possession or immediate preparation	When the shot outcome becomes clear
Kick-off	kick-off taker	When the player and ball are visible and the restart begins	When another player first touches or controls the ball

If a camera cut, replay, close-up, or loss of the relevant role occurs before the normal semantic endpoint, annotators terminate the interval at the last valid moment in the standard broadcast view.

For shot events, the endpoint may correspond to a goal, miss, goalkeeper save or control, defensive clearance, another player taking control, or a broadcast cut. The `last passer` role is optional when no clear final pass exists.

For card-related events, annotators label the **foul that produces the card**, rather than the later close-up of the referee displaying the card.

For events requiring a spatial event-location attribute, annotators record where the event itself occurs rather than where an earlier participant first receives possession. For example, a shot is localized at the shooting location and a foul at the contact location.

E.2.2 Stage 2: Semantic-Role Trajectory Annotation

Stage 2 begins from the temporal window finalized in Stage 1. Videos are sampled at approximately 2 FPS, and the interface provides automatically generated detections and player tracks without semantic-role assignments.

Before assigning roles, annotators re-check that:

- the event type is correct;
- the key action is present;
- the event is primarily visible in the standard broadcast view; and
- the temporal window is not seriously shifted.

Minor boundary differences of a few frames are not treated as failures when the event semantics, key action, broadcast view, and role structure remain intact.

For each required semantic role, annotators follow four operations:

Select. Assign the candidate player corresponding to the role.

Inspect. Review the sampled frames to ensure that the trajectory continues to follow the same individual.

Correct. Reassign a role when the automatic track switches identity, drifts to another person, or otherwise becomes incorrect.

Complete. Manually add a bounding box when the role is clearly visible but no usable automatic detection exists, and remove incorrect boxes when necessary.

A stable track ID is not itself evidence of correct identity. Conversely, a change in track ID does not imply that the role filler has changed. Semantic identity is determined from the event and the visual evidence.

Annotators draw a box only when the role filler is visible and can be reliably identified. They do not infer a location while the person is occluded or outside the frame. Optional roles, such as `last passer`, remain absent when no reliable role filler exists.

E.2.3 Context and Ordinal Verification

After temporal and role annotations are finalized, the target event is placed into a longer evaluation context.

For each query, annotators or reviewers:

1. identify all occurrences in the context that satisfy the query-specific event definition;
2. order them chronologically;
3. locate the target event in that sequence; and
4. assign its ordinal.

The target is the first occurrence if no qualifying event precedes it and the n-th occurrence if exactly $n-1$ qualifying events appear earlier.

Occurrence counting follows the **query definition**, rather than simply counting all available source labels. A single real-world event can belong to several semantic categories. For example, a direct free kick that scores may also qualify as a shot on target and a goal. Such an event is counted once within each query-specific occurrence sequence, while its ordinal is computed separately for each sequence.

Similarly, different foul queries define different occurrence sets. A query for the n-th foul leading to a penalty counts a different subset from a query for the n-th foul in general.

Not every source occurrence must correspond to a retained Event-STVG target annotation. A source event may be excluded as a target because it is insufficiently visible, lacks the required key action, or fails manual review. However, if it remains a valid occurrence under the current query definition and can be verified in the video, it may still contribute to the occurrence sequence used for ordinal assignment.

E.2.4 Final Annotator Verification

After the context and ordinal are generated, the final sample is manually reviewed.

A sample is marked **Pass** only when:

- the target event is correctly mapped;
- the target event is complete;
- the query-specific occurrence set is correct;
- there are no missing, duplicated, or incorrectly counted occurrences; and
- the target ordinal is correct.

If any of these conditions fails, the sample is marked **Fail**.

The resulting end-to-end workflow is:

candidate retrieval → event verification and temporal annotation → event attributes → automatic detection/tracking → manual semantic-role assignment → frame-level trajectory correction → final event/role confirmation → context construction → occurrence and ordinal generation → manual final verification → dataset finalization.

The central annotation principle is that automatic outputs are treated as suggestions, while the final Event-STVG ground truth must be supported by evidence that can be directly verified from the video.

E.3 Soccer Annotation

E.3.1 Temporal Annotation

For soccer, we first use existing event annotations to retrieve candidate occurrences from full broadcast matches. These annotations are used only for candidate retrieval and occurrence association and are not directly adopted as the final temporal intervals.

Annotators inspect the surrounding video and manually determine the start and end of each event according to the event-specific rules above. Unrelated replays, celebrations, close-ups, and subsequent phases of play are excluded when applicable.

Some source events are further decomposed according to the event semantics used in Event-STVG. For example, the action that results in a corner and the subsequent corner kick are treated as separate events. Likewise, a foul leading to a free kick and the execution of the free kick are annotated independently. The resulting temporal interval therefore corresponds to the semantics of the Event-STVG event rather than directly inheriting a point timestamp from the source annotation.

We additionally audit the mapping from locally annotated candidate clips back to the full-match timeline. All 3,516 soccer targets in the final benchmark have valid temporal mappings.

E.3.2 Spatial Role Annotation

After temporal annotation, candidate player trajectories are generated using SoccerMaster DETR, PRTReID, and StrongSORT. These trajectories serve only as annotation assistance and do not directly constitute ground truth.

Annotators identify the players corresponding to the semantic roles from the event context and inspect candidate trajectories at approximately 2 FPS. They retain full control over the trajectories, including reassigning tracks, removing erroneous detections, resolving track-ID switches, and manually adding missing boxes.

When a role filler cannot be reliably localized in a sampled frame, that frame is marked as invisible.

For 25-FPS source videos, half-second sampling is realized by rounding the corresponding frame positions, e.g.,

```
0, 13, 25, 38, 50, ...
```

rather than by sampling at a fixed 12-frame interval, which would introduce accumulated temporal drift.

All ground-truth boxes are first retained in the original image coordinate system together with their source-frame identities and are transformed only when adapting them to the final model input.

E.3.3 Context Construction and Ordinal Assignment

After the target events are frozen, we construct the Oracle and 1-, 3-, 5-, and 10-minute evaluation settings.

The same target events and semantic roles are retained across context durations. Each target interval is mapped into the corresponding context, while qualified same-type occurrences are deduplicated and temporally ordered to determine the ordinal of the target within that context.

Consequently, the underlying target event remains unchanged across durations, while its ordinal index may change as the context expands.

E.4 Multidomain Annotation

The multidomain evaluation set contains 500 target events organized into 100 final evaluation videos.

We retain the same event–role annotation abstraction across domains, while adapting the annotation interface to the form of the available source data. Depending on the domain, annotators either draw boxes directly on original frames, select from existing tracking trajectories, or annotate against frame-exact JPEG sequences.

Table E.2: Event types and semantic roles in the multidomain evaluation.

Domain	Event	Semantic roles
Basketball	Shot	last passer, shooter
Ice hockey	Shot	last passer, shooter, goalkeeper
Volleyball	Serve	server, receiver
Volleyball	Spike	setter, attacker, blocker
Esports	Kill	killer, victim
Animal activity	Initial Attack	predator, prey
Animal activity	Final Subdual	predator, prey
Daily activity	Food Packaging	packaged food
Entertainment	Turn	mentor

Only roles with valid spatial evidence under the corresponding query specification are included in the final model query.

E.4.1 Sports Domains

Basketball Shot events are annotated on the original possession clips. By default, the event begins with the direct pass to the shooter and ends when the shot, layup, or dunk is completed. Spatial annotations for the `last passer` and `shooter` are mapped to the nearest real source frame at each nominal half-second sampling point.

For ice hockey, annotators first define the temporal interval of the Shot and then assign semantic roles using existing player trajectories as candidates. Selecting a candidate initializes the corresponding role tube, after which the complete trajectory is reviewed for identity switches, box drift, and incorrectly missing frames.

For volleyball, annotations are performed against the original 25-FPS JPEG frames. A rally may contain multiple independently annotated Serve or Spike events, including temporally overlapping event windows when semantically appropriate. For Spike events, the `blocker` role is additionally annotated when reliable spatial evidence is available.

E.4.2 Esports, Animal Activity, and Daily Scenarios

For League of Legends, a Kill event covers the combat leading to the target death. The `killer` and `victim` must be localized in the main game view rather than through minimap icons, portraits, kill notifications, or player-camera overlays. When multiple kills occur within the same fight, occurrence order is determined by victim death time.

Animal annotations distinguish the Initial Attack stage from the Final Subdual stage of a hunt. Each event grounds the `predator` and `prey`, while the predator responsible for the initial attack need not be the same individual as the one responsible for the final subdual.

For Food Packaging, the spatial target is the food together with its active packaging rather than the worker performing the action. The temporal interval focuses on packaging and closure and excludes food preparation and subsequent transportation after packaging is complete.

For The Voice, each Turn event grounds the mentor performing the chair turn. Multiple turns within the same performance are ordered temporally.

E.5 Quality-Control Protocol

Event-STVG uses multiple levels of quality control because structurally valid annotations may still be semantically incorrect.

E.5.1 Structural Validation

We first perform automatic checks to verify that:

- annotation files are parseable;
- sample identifiers are unique;
- temporal boundaries are correctly ordered and lie within the source media;
- sampled frame indices are valid, ordered, and non-duplicated;
- bounding boxes have positive area and lie within the corresponding image coordinate system;
- frozen annotations contain no stale or unfinished spatial records; and
- frozen annotations, final-video manifests, query specifications, and ground-truth files are mutually consistent.

E.5.2 Temporal and Media Alignment

For all constructed videos, source-frame provenance is retained through cropping, concatenation, resizing, padding, and re-encoding.

Event boundaries and sampled role frames are mapped to actual output frames or timestamps rather than being inferred solely from nominal frame rates.

We explicitly distinguish three notions of time:

- the annotation time in the source video;
- the mapped time in the final input video; and
- the nominal half-second sampling time.

This prevents re-encoding or frame-rate conversion from silently changing the visual evidence referenced by the annotation.

E.5.3 Semantic Human Review

Automatic validation can verify that a box is structurally valid, but it cannot establish that the box contains the correct semantic role filler.

Human review therefore checks whether:

- the target event actually occurs;
- the temporal boundaries follow the event definition;
- each selected entity satisfies the required semantic role;
- role identity remains consistent throughout the event;
- invisible annotations correspond to genuinely unlocalizable frames rather than missed boxes; and
- annotation guidelines, interface instructions, and final query text are semantically consistent.

E.5.4 Occurrence and Ordinal Validation

Ordinal grounding introduces an additional source of annotation error.

Even if the temporal interval and role trajectories of an event are correct, the corresponding query is still incorrect if the event is assigned the wrong occurrence number.

We therefore separately verify same-type occurrences and their temporal ordering when constructing the final evaluation contexts. Replay or duplicate representations of the same physical event are not intended to create new occurrences.

E.5.5 Cross-Review and End-to-End Human Audit

For soccer, we additionally conduct paired re-annotation:

- 355 samples for temporal annotation; and
- 355 samples for spatial role annotation.

The temporal review interface displays the original interval as a read-only reference, while the role review uses the same tracking candidates as the original annotation.

We therefore interpret this experiment as measuring annotation consistency under the current interface and workflow rather than as independent double-blind annotation.

In addition, we perform a final human audit of the multidomain evaluation set. We sample 50 cases from the 500 multidomain instances according to the scene distribution and manually inspect each frozen benchmark instance as a whole.

This audit does not involve redrawing bounding boxes or re-annotating temporal boundaries. Instead, annotators directly judge whether the existing annotation and final query are jointly correct with respect to three components:

1. temporal localization of the target event;
2. spatial localization of the required semantic roles; and
3. the occurrence ordinal of the target event in the final video.

Each sampled instance receives a binary `Pass/Fail` judgment and is considered a Pass only when all three components are correct.

E.6 Annotation Agreement and Final Audits

The paired soccer annotation results are summarized below.

Table E.3: Soccer paired-annotation agreement.

Metric	Result
Paired temporal annotations	355
Mean temporal IoU	0.904
Median temporal IoU	1.000
Temporal IoU ≥ 0.5	96.06%
Temporal IoU ≥ 0.7	85.07%
Mean start-time difference	0.194 s
Mean end-time difference	0.349 s
Matched role–frame boxes	5,083
Mean bbox IoU, jointly visible	0.945
Mean bbox IoU, including one-sided missing	0.919
One-sided missing boxes	2.70%

Temporal annotations achieve a mean tIoU of 0.904 and a median of 1.000.

Among role–frame pairs for which both annotations contain a visible box, the mean bounding-box IoU is 0.945. When one-sided missing boxes are additionally included and assigned zero overlap, the mean decreases to 0.919, with 2.70% of role–frame pairs containing a box on only one side.

We report both quantities because evaluating only jointly visible frames would obscure disagreements in visibility annotation.

Beyond paired annotation, we conduct several construction-level audits.

For soccer, the final temporal-correction manifest contains all 3,516 unique target events, each with a valid temporal mapping.

We additionally perform a fixed final review of 176 targets across the constructed temporal settings, and all 176 pass the final review.

The Oracle and 1-, 3-, 5-, and 10-minute manifests each contain the same set of 3,516 target IDs, providing an additional consistency check that context construction does not change the evaluated target set.

For the multidomain data, the frozen release contains:

- 500 target events;
- 100 final input videos;
- 6,639 role-sampling records;
- 5,777 visible bounding-box annotations; and
- 862 invisible records.

No frozen target contains a stale spatial annotation.

If a role contains no valid spatial evidence over the entire event, that role is omitted from the corresponding final model query while its original annotation is retained in the provenance. Roles that are invisible only in a subset of frames remain part of the query.

On top of these construction checks, we sample 50 instances from the 500 multidomain cases according to the scene distribution and perform an end-to-end human quality audit. Each instance is evaluated with a binary `Pass/Fail` criterion that jointly considers temporal localization, spatial role localization, and the correctness of the target occurrence ordinal. **All 50 sampled instances pass the audit, corresponding to a sample-level pass rate of 100%.**

This audit is complementary to the paired annotation study above. It does not re-estimate temporal or spatial annotations and therefore is not used to measure inter-annotator tIoU or bounding-box IoU. Instead, it directly verifies whether a frozen benchmark instance remains jointly correct in temporal localization, spatial grounding, and ordinal assignment under the final video and query presented to the model.

We also audit the multidomain media-construction pipeline by decoding the final videos and checking the mapping between annotated source frames and output frames. According to the recorded construction audits, the maximum source-to-output temporal mapping discrepancy is approximately 0.017 s.

Finally, we run regression tests covering:

- role filtering;
- occurrence handling;
- ordinal assignment;
- sampling provenance;
- coordinate transformation; and
- query/ground-truth formatting.

These checks address complementary failure modes.

Automatic validation establishes structural and media-mapping consistency; paired re-annotation measures temporal and spatial consistency under the annotation workflow; annotator-facing operational rules constrain how temporal boundaries, semantic roles, and occurrence sequences are defined; and the final end-to-end human audit directly evaluates whether frozen instances are jointly correct in temporal localization, spatial localization, and occurrence ordinal.

We therefore report structural consistency, annotation agreement, operational annotation guidelines, and final sample-level human audit as complementary forms of quality evidence rather than treating any single check as equivalent to complete independent re-annotation.