LEARNING LIFETIME-AWARE MEMORY FOR STREAMING SPORTS COMMENTARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming sports commentary requires models to identify participants, track unfolding actions, and decide when to describe an event using only the video observed so far. These decisions depend on interrelated semantic states with different useful lifetimes. Identity evidence remains relevant across events, active event states evolve as actions unfold, and completed events provide historical context. The challenge is to maintain these states over their respective lifetimes while preserving the semantic dependencies needed for accurate and timely commentary. In this paper, we propose GroundedPlay, a lifetime-aware memory framework that preserves identity evidence across events, updates active event states, and retains completed outputs as episodic context. It learns reciprocal interactions between identity and action representations, using identity evidence to refine action predictions and action context to select relevant player observations. In addition, context from the completion predictor guides evidence accumulation, while updated event states help predict whether the current event has ended. Experiments on NBA_Streaming and MatchTime-SID demonstrate that GroundedPlay outperforms state-of-the-art baselines in event response timing and commentary quality. It also achieves the highest video processing throughput among the evaluated methods. These results suggest that lifetime-aware memory and reciprocal state interactions offer a promising design approach for streaming tasks that require persistent entity knowledge, evolving event understanding, and timely responses.

1 INTRODUCTION

Automatic sports commentary helps audiences follow rapidly unfolding play by translating match footage into natural language descriptions (Rao et al., 2024; Wu et al., 2026). Recent video language models increasingly support online video understanding by incrementally processing incoming frames, retaining historical context, and generating proactive responses (Chen et al., 2024; Huang et al., 2025; Yang et al., 2026; Zhang et al., 2026b). In streaming sports commentary (Wu et al., 2026), a system must identify participants, track unfolding actions, and decide when to describe an event using only the video observed so far. Because the evidence needed for these decisions is distributed over time, the system must integrate earlier observations with incoming frames to maintain an evolving understanding of play. An important question is therefore how memory should preserve and update the different kinds of semantic information needed as play unfolds.

Consider a player who is clearly visible before a shot but becomes occluded as the play unfolds. The earlier observation can help identify the shooter and recognize the same player in later events, giving identity evidence a useful lifetime that extends beyond the current shot. The active event state, by contrast, must be updated as the shot develops and cease to represent ongoing play once the event ends. The completed shot can then serve as historical context for describing a subsequent rebound. Discarding identity evidence at the event boundary can weaken later attribution, whereas carrying outdated action states into the next event can distort the interpretation of ongoing play. These differences in lifetime do not make the underlying semantic states independent. Identity evidence helps associate an action with its participants, while action context helps determine which player observations are relevant. The evolving event interpretation also provides evidence for deciding whether the event has ended. As illustrated in Figure 1, these requirements motivate memory that preserves semantic information over distinct lifetimes while supporting interactions between identity, event interpretation, and completion prediction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

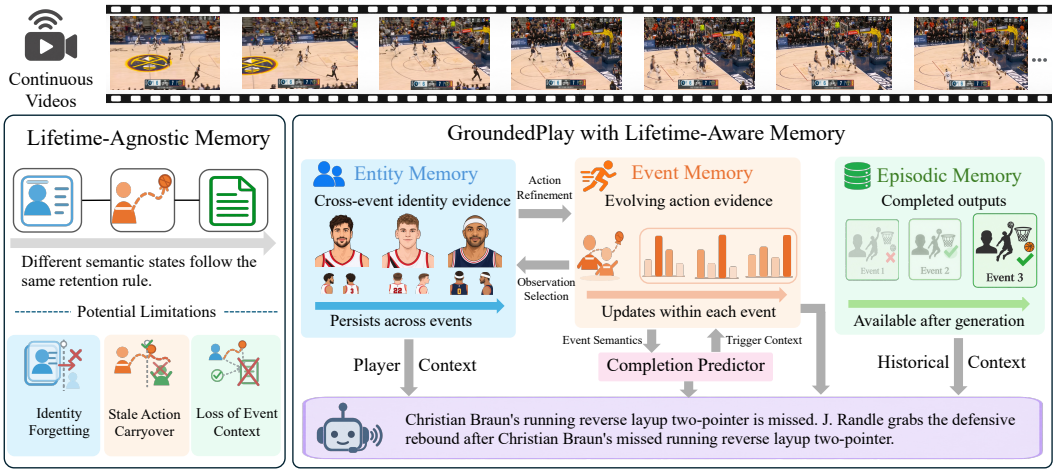


Figure 1: Streaming commentary requires preserving identity across events, updating ongoing actions, and retaining historical context while maintaining dependencies between these semantic states.

Prior work on streaming video understanding maintains temporal context through visual evidence compression and selective retention (Huang et al., 2025; Zhang et al., 2026b). Multimodal memory methods explore semantic organization, separating perceptual and cognitive information in MemoryVLA (Shi et al., 2026) or organizing episodic and semantic memory around entities in M3-Agent (Long et al., 2026). For streaming sports commentary, StreamSoccer models event lifecycles through active states and completed records (Shao et al., 2026), while NBA_Streaming performs detailed identity and action grounding after event localization (Wu et al., 2026). These advances motivate coordinating semantic states with different lifetimes and connecting their interactions to event completion prediction. We focus on reciprocal interactions between persistent identity evidence and evolving action representations before completion is decided, while coupling evidence accumulation with completion prediction. Predictor context guides event state updates, while updated states inform completion decisions. We investigate how learning these interactions within a lifetime-aware memory framework supports accurate and timely commentary.

To this end, we introduce GroundedPlay, a lifetime-aware memory framework for streaming sports commentary. GroundedPlay preserves reusable identity evidence in Persistent Entity Memory, tracks ongoing actions in Active Event Memory, and retains completed outputs in Episodic Memory. Historical identity predictions are retrieved and fused with current predictions through a learned gate, whereas active event states are recurrently updated and reset at predicted completion. To maintain action-participant associations, Reciprocal Semantic Grounding (RSG) uses role-specific player attention to refine action predictions and action context to guide the selection of player observations. Evidence-Completion Coupling (ECC) uses predictor context to guide visual evidence accumulation and updated event states to inform completion decisions. At predicted completion, GroundedPlay generates commentary from current visual and semantic evidence together with retrieved episodic context, while perception continues. We evaluate GroundedPlay on NBA_Streaming (Wu et al., 2026) and MatchTime-SID, a streaming version of MatchTime (Rao et al., 2024) with player identity. GroundedPlay achieves superior event response performance and commentary quality. On NBA_Streaming, it reduces response time difference by 0.655 s and improves CIDEr by 128.491 points over the state-of-the-art baseline. Ablation studies further show the benefits of distinct memory lifetimes and reciprocal interactions.

Our contributions are summarized as follows:

- We propose GroundedPlay, a lifetime-aware memory framework for streaming sports commentary that preserves identity evidence across events, updates active event states within events, and retains completed outputs as historical context, supporting distinct semantic lifetimes in a unified framework.
- We introduce Reciprocal Semantic Grounding (RSG) to maintain coherent action-identity associations through bidirectional interaction. To connect this grounding with response

timing, Evidence-Completion Coupling (ECC) uses trigger context to guide evidence accumulation and updated event states to inform completion prediction.

- Experiments on NBA_Streaming and MatchTime-SID demonstrate the benefits of heterogeneous memory lifetimes and reciprocal state interactions, improving event response and commentary quality while maintaining efficient streaming inference.

2 Related Work

2.1 Streaming Video Understanding

Streaming video understanding requires models to interpret continuously arriving observations without access to future frames. Causal captioning and dialogue frameworks incrementally incorporate visual evidence to support narration and conversational responses based on the observed video prefix (Zhou et al., 2024; Chen et al., 2024; Huang et al., 2025). Unified streaming instruction tuning extends these capabilities to narration, temporal grounding, and question answering (Xia et al., 2026). Other approaches improve online efficiency through selective model invocation and query-relevant evidence selection (Ding et al., 2025; Zhang et al., 2026b), while asynchronous and parallel architectures coordinate visual processing with language generation (Tang et al., 2026; Zhang et al., 2026a). Beyond generating response content, streaming models must determine when to respond. Existing methods address this decision by assessing evidence readiness or detecting event transitions (Azad et al., 2026; Guo et al., 2026). GroundedPlay couples event completion prediction with ongoing semantic evidence accumulation. Trigger context guides evidence selection, while updated event states inform completion decisions.

2.2 Memory for Video Understanding

Video memory research manages growing observation histories through evidence compression and semantic organization. Compression methods reduce redundant visual features and tokens (Huang et al., 2025; Wang et al., 2026), combine recent visual details with older textual summaries (Chatterjee et al., 2025), or compress and retrieve visual key-value caches (Chen et al., 2026). These methods control storage and retrieval costs, but observation age and redundancy do not fully capture the useful lifetimes of identity evidence and event interpretations. Structured memories explicitly organize retained information. OASIS (Liang et al., 2026) uses hierarchical events for retrieval, while MemoryVLA (Shi et al., 2026) separates perceptual and cognitive memories and M3-Agent (Long et al., 2026) maintains episodic and semantic knowledge around entities. GroundedPlay organizes memory according to distinct semantic lifetimes. Identity evidence persists across events, active event states are updated within events and reset at predicted completion, and completed events are consolidated into compact episodic records. These retention and transition rules preserve reusable evidence and historical context while limiting interference from outdated actions and outcomes.

2.3 Sports Commentary and Semantic Grounding

Sports commentary has evolved from clip-level descriptions to contextual narration over continuous matches. Early work models fine-grained actions and player interactions (Yu et al., 2018), while SoccerNet-Caption (Mkhallati et al., 2023) and MatchTime (Rao et al., 2024) provide timestamped commentary and improved video-text alignment. MatchAware (Sun et al., 2026) retrieves historical event context, while LLM-IAVC (Xi et al., 2025) combines player recognition with bidirectional interaction between player and clip features. However, this interaction remains within individual clips, without an explicit mechanism for updating identity memory across events. Streaming frameworks further address event localization and memory evolution. NBA_Streaming (Wu et al., 2026) applies ball-centered semantic grounding after localizing completed events, so its detailed player and action grounding does not inform completion decisions. StreamSoccer (Shao et al., 2026) learns active event updates and operational transitions, retaining completed records for commentary at multiple temporal scales. Its memory centers on ongoing and completed events. GroundedPlay couples persistent identity memory with active event memory throughout the stream. Identity evidence supports participant attribution, while event context guides identity updates, enabling reciprocal action-identity grounding before event completion is predicted.

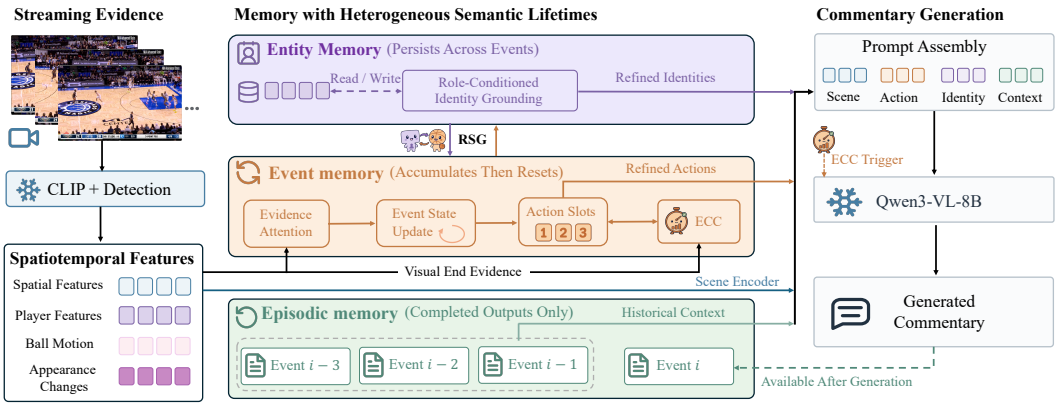


Figure 2: Overview of GroundedPlay. Three memories maintain semantic information over distinct lifetimes. RSG supports identity-action grounding, while ECC couples evidence accumulation with completion prediction. Current evidence and episodic context condition commentary generation.

3 METHOD

GroundedPlay predicts event endings and generates sports commentary using only the video observed so far. As shown in Figure 2, it maintains three memories with distinct semantic lifetimes (Section 3.1). Reciprocal Semantic Grounding (RSG) couples action interpretation with participant identity (Section 3.2), while Evidence-Completion Coupling (ECC) connects evidence accumulation with event completion prediction (Section 3.3). The resulting visual and semantic representations condition commentary generation while perception continues (Section 3.4).

3.1 LIFETIME-AWARE MEMORY

We preserve identity evidence across events, update action states within events, and retain completed outputs as episodic context.

Spatiotemporal Features. For each incoming frame, a frozen CLIP encoder (Radford et al., 2021) extracts a feature map from an adaptive ball-centered region, which is then pooled into spatial tokens $\mathbf{v}_t$. ROIAlign (He et al., 2017) extracts player appearance features from the same map, which are combined with box coordinates and ball-relative geometry to form player features $\mathbf{p}_t$. Backward differences of detected ball positions, detection confidence, and validity indicators form a motion descriptor $\mathbf{m}_t$. Together with inter-frame appearance changes $\mathbf{d}_t$, these features support causal memory updates and completion prediction.

Event Memory. Active event memory selectively accumulates evidence to track ongoing actions. Motion-conditioned attention uses a query derived from $\mathbf{m}_t$ to aggregate spatial tokens $\mathbf{v}_t$. To emphasize relevant participants, player attention scores each box using its ball-relative geometry and $\mathbf{m}_t$, then aggregates $\mathbf{p}_t$ with softmax-normalized scores after masking invalid boxes. The aggregated spatial and player representations are fused with $\mathbf{m}_t$ and $\mathbf{d}_t$ and processed by a two-layer GRU (Chung et al., 2014), yielding the event state $\mathbf{h}_t^A$. Three learnable action slots predict actions from $\mathbf{h}_t^A$ in commentary order, with an empty label for unused slots. To revise predictions as evidence accumulates, each slot maintains an action distribution updated through a learned gate:

$$\boldsymbol{\pi}_{t,k}^A = (1 - g_{t,k})\boldsymbol{\pi}_{t-1,k}^A + g_{t,k}\tilde{\boldsymbol{\pi}}_{t,k}^A, \qquad k \in \{1, 2, 3\}, \tag{1}$$

where $\tilde{\boldsymbol{\pi}}_{t,k}^A$ is the current prediction for slot k and $g_{t,k} \in [0, 1]$ controls its contribution. The state, prediction confidence, and distribution changes provide completion evidence to ECC. At predicted completion, the current action semantics are retained for commentary generation before active memory is reset to prevent outdated action states from carrying over into subsequent play.

Entity Memory. To preserve identity evidence across events, a bounded, game-specific memory $\mathcal{M}^E$ stores player appearance representations, original identity distributions, role confidence, attention concentration, and timestamps. Role-conditioned queries aggregate recent player features into

participant representations, from which identity distributions are predicted. At each trigger, participant representations retrieve historical entries by appearance similarity, excluding entries whose observation windows overlap the current window. Retrieval weights also account for reliability estimated from identity uncertainty, role confidence, and attention concentration. During retrieval, stored identity distributions are temporarily refined using compatible, non-overlapping entries without modifying the original distributions in memory. A weighted aggregation of these refined distributions yields a historical identity distribution $\boldsymbol{\pi}_{t,r}^{E,\mathrm{mem}}$ for role r, which is fused with the current identity distribution $\boldsymbol{\pi}_{t,r}^{E}$:

$$\boldsymbol{\pi}_{t,r}^{E,+} = (1-\beta_{t,r})\boldsymbol{\pi}_{t,r}^{E} + \beta_{t,r}\boldsymbol{\pi}_{t,r}^{E,\mathrm{mem}}, \tag{2}$$

where the learned gate $\beta_{t,r} \in [0,1]$ controls reliance on history based on confidence, appearance matching, and prediction agreement. After all roles have read the memory, eligible entries containing appearance representations, original identity distributions, reliability statistics, and temporal metadata are written with redundancy filtering and capacity control. Memory-refined predictions are never written back, preventing repeated storage of memory-derived corrections. Historical identity evidence is also retrieved before triggering to support RSG.

Episodic Memory. To contextualize subsequent events, a bounded episodic memory $\mathcal{M}^C$ stores completed captions, event timestamps, and associated player appearance representations and original identity distributions. Records become accessible only after commentary generation finishes. At each trigger, we retrieve up to three completed outputs from the preceding 30 seconds of the same stream. Their captions provide textual context, while stored identity evidence refines current player representations through learned gates. Identity retrieval uses appearance similarity, confidence, and recency, retaining only identity entries whose predicted names appear in the corresponding captions.

3.2 Reciprocal Semantic Grounding

Identity-Guided Action Refinement. To distinguish action-relevant participants in multi-player scenes, we use role-specific evidence from the identity branch to refine action predictions. Each action slot computes matching scores with role-specific participant representations and uses these scores together with role presence probabilities to aggregate the corresponding player-box attention maps into an attention prior. Matching scores between each action slot and candidate player boxes are combined with this prior and normalized by softmax to aggregate player features. A multilayer perceptron (MLP) combines these features, the slot representation, and role matching information to predict an action-logit correction. A learned gate scales the correction using action certainty, role presence, attention concentration, and visual compatibility, limiting the influence of unreliable role evidence. This refinement is applied after the gated update in Eq. 1. The corrected logits are normalized by softmax to obtain the refined action distribution $\boldsymbol{\pi}_{t,k}^{A,+}$.

Action-Guided Identity Grounding. Identifying action participants requires selecting relevant player observations across time. We therefore use action semantics to guide where and when the identity branch attends. Each action slot's representation and refined action distribution $\boldsymbol{\pi}_{t,k}^{A,+}$ are projected into keys and values for role queries. Attention weighted by each slot's non-empty probability aggregates action evidence to update role queries and presence probabilities. The updated queries attend again to player observations, using player features, ball motion, and relative time to weight candidate boxes across time steps. The aggregated features form role-specific participant representations for predicting identity distributions over player names.

3.3 Evidence-Completion Coupling

Visual Completion Prediction. To detect event endings from changes in the observed play, a visual trigger integrates scene, player, and motion evidence. Motion-conditioned attention separately aggregates spatial tokens $\mathbf{v}_t$ and player features $\mathbf{p}_t$. These summaries are fused with MLP encodings of $\mathbf{m}_t$ and $\mathbf{d}_t$. After adding positional encodings, a four-layer causal Transformer integrates the spatially aggregated frame representations over time to obtain a trigger representation $\mathbf{h}_t^T$. A linear head maps $\mathbf{h}_t^T$ to a visual completion logit z_t^V, representing visual evidence for event completion.

Trigger-Event Interaction. Completion prediction depends on the accumulated action evidence, while trigger context helps determine which observations should contribute to that evidence. ECC

couples these processes through bidirectional interaction. An MLP combines the trigger representation $\mathbf{h}_t^T$ with the previous event state $\mathbf{h}_{t-1}^A$ to form a query. Its matching scores with spatial tokens and player features are added to the corresponding attention logits in the action branch. The resulting softmax weights aggregate visual evidence, which is fused with motion and appearance-change encodings to update $\mathbf{h}_t^A$ through the GRU. In the reverse direction, a completion head combines the updated state with action confidence, temporal stability, distribution changes, and update gates to predict an action-based completion logit z_t^A. A separate MLP uses both completion logits and action statistics to produce a correction Δ_t, yielding

$$z_t = \tfrac{1}{2} z_t^V + \tfrac{1}{2} z_t^A + \Delta_t, \qquad p_t^{\text{end}} = \sigma(z_t), \tag{3}$$

where σ denotes the sigmoid function. This sequence allows trigger context to guide evidence accumulation before the updated action state informs the response decision.

At inference, the system initially permits a response and triggers immediately when $p_t^{\text{end}} \geq \tau$, with τ selected on the validation set. After each trigger, further responses are suppressed until the probability falls to 0.75τ or below, preventing repeated responses during sustained high scores without requiring future-frame confirmation.

3.4 Commentary Generation and Training

Memory-Conditioned Commentary Generation. At each trigger, a scene encoder summarizes spatial tokens from the latest L frames through spatial pooling, temporal modeling, and learnable scene queries. Retrieved episodic identity evidence refines current player representations and identity predictions. Action labels and role-name pairs form textual prompts, while scene, action, and player representations are projected into continuous prompts. A frozen Qwen3-VL-8B (Bai et al., 2025) generates commentary from these prompts, a generation instruction, and previously completed captions as historical context. After generation finishes, the caption, event timestamp, player appearance representations, and original identity distributions become available in $\mathcal{M}^C$.

Training Objectives and Optimization. Training proceeds in three stages. First, we independently pretrain the visual trigger, the action module underlying active event memory, and the identity prediction module using event-ending, action, and weak identity supervision, respectively. Second, we jointly optimize these modules, RSG, ECC, and the learnable retrieval and fusion mechanisms of entity memory. The objectives supervise completion, action and identity prediction, and memory refinement, while response alignment encourages one response per annotated ending and penalizes unmatched or mistimed responses. Finally, we freeze the upstream modules and Qwen3-VL-8B and train the scene encoder, visual projectors, and episodic identity fusion modules using cross-entropy over target caption tokens. The first epoch of this stage uses ground-truth action and identity labels as textual prompts without historical context. Subsequent epochs use predicted labels and generated captions, accessing only outputs completed before the current trigger. Detailed loss formulations, weights, and training procedures are provided in Appendix A.3.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate GroundedPlay on continuous basketball and soccer video streams using NBA_Streaming and MatchTime-SID. NBA_Streaming (Wu et al., 2026) provides temporally aligned commentary with player identities, fine-grained actions, and event-chain annotations. Its 152 games are split into 121 for training, 15 for validation, and 16 for testing. We use event endings for response supervision without event-start annotations. For soccer, we adapt the version of MatchTime (Rao et al., 2024) that includes player names to streaming commentary, and refer to the resulting dataset as MatchTime-SID. We retain the original split of 373 training, 49 validation, and 49 test games, using manually aligned test commentary timestamps as reference response times.

Evaluation Metrics. We evaluate event response, commentary quality, and streaming efficiency. Event response is measured by response time difference (Δt), recall, and precision. Commentary quality is assessed using CIDEr (Vedantam et al., 2015) and BLEU-4 (Papineni et al., 2002), together with player identity and action accuracies to evaluate semantic correctness. Streaming efficiency is

Table 1: Comparison on the NBA_Streaming and MatchTime-SID benchmarks. Δt and Lat. denote the response time difference and output latency, both measured in seconds. Rec. and Prec. denote response recall and response precision, respectively. C and B-4 denote CIDEr and BLEU-4. Id. and Act. denote identity and action accuracies. Rec., Prec., Id., and Act. are reported as percentages. "-" denotes baselines not evaluated due to missing required event-boundary annotations.

Method	NBA_Streaming									MatchTime-SID								
	Event Response			Commentary Quality				Efficiency		Event Response			Commentary Quality				Efficiency	
	$\Delta t \downarrow$	Rec. ↑	Prec. ↑	C ↑	B-4 ↑	Id. ↑	Act. ↑	FPS ↑	Lat. ↓	$\Delta t \downarrow$	Rec. ↑	Prec. ↑	C ↑	B-4 ↑	Id. ↑	Act. ↑	FPS ↑	Lat. ↓
General-purpose VLMs																		
Qwen3-VL-8B	7.602	29.055	29.527	14.209	3.529	2.860	5.683	12.49	2.05	15.323	17.714	7.898	6.816	1.151	5.374	14.446	12.06	1.76
Video-R2	18.709	15.814	16.615	10.423	1.504	0.812	2.505	9.57	1.72	13.277	15.139	9.896	4.849	0.735	2.099	10.408	15.57	1.34
Streaming VLMs																		
VideoLLM-Online	5.663	53.289	44.434	12.236	7.965	0.110	15.071	52.39	**0.32**	23.694	7.416	7.437	7.242	2.142	3.534	7.358	86.92	**0.40**
StreamMind	31.867	45.485	27.218	14.902	4.423	0.375	4.933	158.51	0.71	17.662	15.078	13.595	2.884	1.036	0.111	3.320	150.32	0.85
Sports Commentary Models																		
MATR + MatchVoice	4.983	61.711	45.426	74.967	18.869	3.317	25.230	78.96	2.06	-	-	-	-	-	-	-	-	-
MATR + LLM-IAVC	4.983	61.711	45.426	51.229	17.336	1.488	23.244	75.74	6.16	-	-	-	-	-	-	-	-	-
NBA_Streaming	2.046	77.715	70.240	191.525	34.295	17.697	41.299	230.65	1.15	-	-	-	-	-	-	-	-	-
GroundedPlay	**1.391**	**80.802**	**78.403**	**320.016**	**45.924**	**36.592**	**52.659**	**239.94**	1.06	**12.561**	**31.964**	**19.318**	**27.155**	**4.291**	**5.556**	**21.471**	**234.40**	1.26

measured by video processing throughput (FPS) and output latency. Detailed event-matching rules and metric definitions are provided in Appendix A.3.

Implementation Details. We sample videos at 6 FPS and pool frozen CLIP features into 3×3 spatial tokens. The trigger uses a four-layer causal Transformer with hidden size 256. The event memory uses a two-layer GRU with hidden size 192 and three action slots. Entity memory retains up to 128 entries per game, while episodic retrieval accesses at most three completed outputs from the preceding 30 seconds. The scene encoder uses the latest $L = 120$ frames for commentary generation. Joint perception training uses AdamW on 8 GPUs for up to 15 epochs with a per-GPU batch size of 128. Generation training runs for up to 12 epochs with a per-GPU batch size of 2, keeping the upstream modules and Qwen3-VL-8B frozen. Inference uses greedy decoding with at most 96 new tokens. The trigger threshold is set to $\tau = 0.4$, selected on the validation set. Further architecture and optimization details are provided in Appendix A.3.

Baseline Methods. We compare GroundedPlay with three groups of baselines to assess general video understanding, streaming response, and sports-specific commentary. General-purpose VLMs include Qwen3-VL-8B (Bai et al., 2025) and Video-R2 (Maaz et al., 2025), testing whether general multimodal capabilities suffice for streaming commentary. VideoLLM-Online (Chen et al., 2024) and StreamMind (Ding et al., 2025) provide comparisons with models designed for online video understanding. For sports commentary, we pair MATR (Song et al., 2024) with MatchVoice (Rao et al., 2024) and LLM-IAVC (Xi et al., 2025), combining event localization with soccer commentary and identity-aware basketball captioning, respectively. We also include NBA_Streaming (Wu et al., 2026) as a dedicated streaming basketball commentary baseline. These sports commentary baselines are evaluated only on NBA_Streaming, as their training setups require complete event-boundary annotations that are unavailable in MatchTime-SID. Except for Qwen3-VL-8B and Video-R2, all baselines are retrained on the training split of each dataset on which they are evaluated.

4.2 Main Results

Event Response and Commentary Quality. As shown in Table 1, GroundedPlay achieves the best event response and commentary quality across both benchmarks. On NBA_Streaming, it reduces Δt by 0.655 s and improves recall and precision by 3.087% and 8.163%, respectively, compared with the SOTA method. These gains support incorporating evolving action evidence into completion prediction rather than grounding events only after localization. GroundedPlay also achieves 320.016 CIDEr, 36.592% identity accuracy and 52.659% action accuracy, outperforming VideoLLM-Online (Chen et al., 2024) and StreamMind (Ding et al., 2025). While these methods model streaming context through language modeling or event-gated perception, GroundedPlay preserves player identities across events while resetting event states at event completion, retaining reusable identity evidence and limiting interference from outdated actions for more accurate commentary. On MatchTime-SID, GroundedPlay achieves a Δt of 12.561 s and a CIDEr score of 27.155. Despite leading the com-

pared baselines, it exhibits a larger response time difference and lower commentary quality than on NBA_Streaming. The timing gap may partly reflect differences in event complexity and reference timestamp annotations, while the larger player vocabulary poses additional challenges for player identification. These results support the applicability of our approach across sports while highlighting the challenges in both timely response and fine-grained soccer commentary.

To control for differences in event coverage, we further evaluate commentary quality on the intersection of reference events successfully matched by all methods. As shown in Table 5 in Appendix A.4, GroundedPlay also outperforms the compared baselines on this shared event set, indicating that its commentary quality gains are not solely due to differences in event coverage.

Table 2: Ablation of memory components, lifetimes, and historical context on NBA_Streaming. GroundedPlay preserves entity memory across events, resets event memory at predicted event completion, and retrieves the most recent K completed events. Earlier-event retrieval skips these events and reads the preceding K events under the same memory and context token budgets.

Memory Configuration	Event Response			Commentary Quality			
	$\Delta t \downarrow$	Rec. ↑	Prec. ↑	C ↑	B-4 ↑	Id. ↑	Act. ↑
GroundedPlay	**1.391**	**80.802**	78.403	**320.016**	**45.924**	**36.592**	**52.659**
Memory Removal							
w/o Entity Memory	1.415	80.639	77.755	93.127	28.383	0.388	52.210
w/o Event Memory	1.904	78.067	68.334	185.262	28.835	23.845	27.803
w/o Episodic Memory	1.391	80.802	78.403	312.693	45.876	36.163	52.130
Memory Lifetimes and Context							
Reset Entity After Each Event	1.431	79.935	77.766	273.662	43.108	30.207	52.388
Retain Event Across Events	1.727	79.096	**78.564**	292.883	43.967	33.601	50.860
Retrieve Earlier Events	1.391	80.802	78.403	303.628	43.722	34.161	50.655

Streaming Efficiency. Table 1 reports video processing throughput and output latency. GroundedPlay keeps processing video frames while generating commentary, which helps maintain high FPS. On NBA_Streaming, it achieves the highest throughput among the evaluated methods. Compared with NBA_Streaming, it improves FPS from 230.65 to 239.94 and reduces latency from 1.15 to 1.06 s. Although VideoLLM-Online (Chen et al., 2024) and StreamMind (Ding et al., 2025) achieve lower output latencies, GroundedPlay incorporates explicit semantic memory updates and interactions before commentary generation, introducing additional processing. GroundedPlay achieves substantially higher CIDEr scores and identity and action accuracies than these streaming baselines while maintaining higher video processing throughput, suggesting a favorable trade-off between commentary quality and output latency. On MatchTime-SID, GroundedPlay achieves 234.40 FPS with an output latency of 1.26 s.

4.3 Ablation and Analysis

Effect of Memory Design. Table 2 examines the contributions of memory components and their retention policies on NBA_Streaming. Removing entity memory reduces CIDEr from 320.016 to 93.127 and identity accuracy from 36.592% to 0.388%. This suggests that persistent identity evidence primarily benefits participant attribution and commentary quality. Removing event memory reduces action accuracy from 52.659% to 27.803% and increases Δt from 1.391 to 1.904 s, highlighting the importance of accumulated event evidence for both action recognition and completion prediction. Without episodic memory, response metrics remain unchanged, while CIDEr decreases to 312.693, indicating a smaller but complementary contribution from completed-event context.

Beyond the presence of memory, we examine whether its retention should follow semantic lifetimes. Resetting entity memory after each event reduces identity accuracy to 30.207% and CIDEr to 273.662, supporting the value of preserving identity evidence across events. Retaining event memory across events reduces action accuracy to 50.860% and increases Δt to 1.727 s, despite a slight increase in precision. These results are consistent with the need to clear outdated event states when transitioning to new play. Retrieving earlier events instead of the most recent completed events reduces CIDEr to 303.628 under the same memory and context token budgets. This suggests that the relevance of historical context, rather than its quantity alone, contributes to commentary quality.

Effect of Bidirectional Interactions. Table 3 shows the contributions of ECC and RSG. For ECC, $E \rightarrow T$ provides the larger gain, reducing Δt from 1.995 to 1.400 s and increasing precision from 66.714% to 77.682%. Adding $T \rightarrow E$ further improves timing, recall, precision, and action accuracy, supporting reciprocal interaction between event-state updates and completion prediction. For RSG, $A \rightarrow I$ raises identity accuracy from 31.018% to 35.730%, highlighting the value of action context

Table 3: Ablation of bidirectional interactions on NBA_Streaming. E and T denote the active event state and visual trigger, while I and A denote the identity and action branches. Arrows indicate the direction of information flow. The last row in each panel corresponds to GroundedPlay.

(a) Evidence-Completion Coupling

Interaction		Event Response			Action
$E \to T$	$T \to E$	$\Delta t \downarrow$	Rec. ↑	Prec. ↑	Act. ↑
×	×	1.995	77.119	66.714	48.233
✓	×	1.400	80.774	77.682	51.928
×	✓	1.986	77.254	66.535	49.668
✓	✓	**1.391**	**80.802**	**78.403**	**52.659**

(b) Reciprocal Semantic Grounding

Interaction		Event Response			Commentary Quality			
$I \to A$	$A \to I$	$\Delta t \downarrow$	Rec. ↑	Prec. ↑	C ↑	B-4 ↑	Id. ↑	Act. ↑
×	×	1.412	80.720	77.772	277.781	44.414	31.018	51.464
✓	×	1.398	80.774	77.601	283.491	43.776	31.585	51.777
×	✓	1.421	80.720	**78.406**	308.678	45.818	35.730	52.216
✓	✓	**1.391**	**80.802**	78.403	**320.016**	**45.924**	**36.592**	**52.659**



NBA_Streaming:
23min00.33s L. Black misses a jump shot three-pointer. B. Carrington assists. B. Portis grabs the defensive rebound after L. Black's missed three-pointer.
23min21.33s C. Anthony misses a jump shot three-pointer. B. Lopez grabs the defensive rebound after C. Anthony's missed three-pointer.
23min31.83s C. Anthony makes a floating jump shot two-pointer, assisted by K. Kuzma.

w/o Entity Memory:
23min00.33s J. Allen misses a pullup three-pointer; J. Allen grabs the defensive rebound after missing his jump shot three-pointer.
23min21.67s J. Allen misses a jump shot three-pointer; J. Allen grabs the defensive rebound after missing his jump shot three-pointer.
23min32.00s D. Mitchell makes a driving floating jump shot two-pointer, assisted by J. Allen.

Ours:
23min00.33s B. Carrington misses a pullup three-pointer; B. Portis grabs the defensive rebound after B. Carrington's missed jump shot three-pointer.
23min21.67s B. Portis misses a jump shot three-pointer; T. Vukcevic grabs the defensive rebound after B. Portis's missed three-pointer.
23min32.00s T. Vukcevic makes a driving finger roll two-pointer, assisted by B. Carrington.

Groundtruth:
23min00.50s B. Carrington misses a pullup three-pointer; B. Portis grabs the defensive rebound after B. Carrington's missed jump shot three-pointer.
23min22.33s B. Portis misses a jump shot three-pointer; T. Vukcevic grabs the defensive rebound after B. Portis's missed three-pointer.
23min32.50s T. Vukcevic makes a driving finger roll two-pointer, assisted by B. Carrington.

Figure 3: Qualitative comparison of GroundedPlay, NBA_Streaming, and a variant without entity memory on three consecutive events from an NBA video stream.

for participant identification. Adding $I \to A$ further improves CIDEr to 320.016, identity accuracy to 36.592%, and action accuracy to 52.659%, with nearly unchanged precision. These results support the complementary benefits of bidirectional interactions for event response and semantic grounding.

Qualitative Analysis. Figure 3 shows that GroundedPlay matches the reference actions and participants across three consecutive events, whereas NBA_Streaming misidentifies players and the final scoring action. Removing entity memory preserves response timing but introduces repeated identity errors, highlighting its role in maintaining participant consistency across events.

Further analysis of trigger threshold sensitivity, additional qualitative results, and extended analyses are provided in the Appendix A.4 and Appendix A.5.

5 Conclusion

We presented GroundedPlay, a lifetime-aware memory framework for streaming sports commentary. By maintaining persistent entity memory, active event memory, and compact episodic memory, the framework preserves reusable identity evidence while separating ongoing play from completed events. RSG supports reciprocal identity and action grounding, while ECC couples evidence accumulation with event completion prediction. Together, these mechanisms connect participant attribution, event interpretation, and response timing using only the observed video prefix. Experiments demonstrate improved event response performance and commentary quality. These findings suggest that aligning memory retention with semantic lifetimes and learning interactions between semantic states can benefit streaming sports commentary.

REFERENCES

Shehreen Azad, Vibhav Vineet, and Yogesh S Rawat. Streamready: Learning what to answer and when in long streaming videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40494–40504, 2026.

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

Dibyadip Chatterjee, Edoardo Remelli, Yale Song, Bugra Tekin, Abhay Mittal, Bharat Bhatnagar, Necati Cihan Camgoz, Shreyas Hampali, Eric Sauser, Shugao Ma, et al. Streaming videollms for real-time procedural video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22586–22598, 2025.

Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18407–18418, 2024.

Tao Chen, Kun Zhang, Qiong Wu, Xiao Chen, Chao Chang, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Scaling the long video understanding of multimodal large language models via visual memory mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31877–31888, 2026.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13448–13459, 2025.

Zhenghui Guo, Yuanbin Man, Junyuan Sheng, Bowen Lin, Ahmed Ahmed, Bo Jiang, Boyuan Zhang, Miao Yin, Sian Jin, Omprakash Gnawal, and Chengming Zhang. Event-vstream: Event-driven real-time understanding for long video streams. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3060–3069, 2026.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2961–2969, 2017.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3328–3338, 2025.

Zhijia Liang, Jiaming Li, Weikai Chen, Yanhao Zhang, Haonan Lu, and Guanbin Li. Oasis: On-demand hierarchical event memory for streaming video reasoning. *arXiv preprint arXiv:2604.17052*, 2026.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, pp. 146197–146246, 2026.

Muhammad Maaz, Hanoona Rasheed, Fahad Shahbaz Khan, and Salman Khan. Video-r2: Reinforcing consistent and grounded reasoning in multimodal language models. *arXiv preprint arXiv:2511.23478*, 2025.

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5074–5085, 2023.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. Matchtime: Towards automatic soccer game commentary generation. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1671–1685, 2024.

Chenxi Shao, Bozhong Wang, Jiaxin Huang, Zhao Liu, Sunwei Zhu, Tianxin Hang, Gaoqi He, Yang Li, and Changbo Wang. Streamsoccer: Event-driven memory for streaming soccer commentary. *arXiv preprint arXiv:2608.19723*, 2026.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. Memoryvla: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. In *International Conference on Learning Representations*, pp. 18567–18602, 2026.

Youngkil Song, Dongkeun Kim, Minsu Cho, and Suha Kwak. Online temporal action localization with memory-augmented transformer. In *European Conference on Computer Vision*, pp. 74–91, 2024.

Haoran Sun, Natthawut Kertkeidkachorn, and Kiyoaki Shirai. Visual and memory-augmented soccer commentary generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 10614–10629, 2026.

Yolo Y Tang, Chao Huang, Susan Liang, Jing Bi, Yicheng Wang, Daiki Shimada, and Chenliang Xu. Asynchronous temporal modeling with two-agent framework for streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2799–2810, 2026.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575, 2015.

Yiyu Wang, Xuyang Liu, Xiyan Gui, Xinying Lin, Boxue Yang, Chenfei Liao, Tailai Chen, and Linfeng Zhang. Accelerating streaming video large language models via hierarchical token compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18523–18533, 2026.

Lifang Wu, Yuyang Wu, Yangdong Gao, Fengyu Liu, Ya Jing, and Liang Wang. Nba_streaming: A large-scale benchmark for fine-grained basketball commentary generation in continuous streams. *arXiv preprint arXiv:2608.09200*, 2026.

Zeyu Xi, Haoying Sun, Yaofei Wu, Junchi Yan, Haoran Zhang, Lifang Wu, Liang Wang, and Changwen Chen. Player-centric multimodal prompt generation for large language model based identity-aware basketball video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24330–24339, 2025.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31219–31229, 2026.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. Livestar: Live streaming assistant for real-world online video understanding. In *Advances in Neural Information Processing Systems*, pp. 31266–31304, 2026.

Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6006–6015, 2018.

Jialiang Zhang, Junlong Tong, Junyan Lin, Hao Wu, Yirong Sun, Yunpu Ma, and Xiaoyu Shen. Think-as-you-see: Streaming chain-of-thought reasoning for large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11998–12008, 2026a.

Kairui Zhang, Zhenyu Yang, Bing Wang, Shengsheng Qian, and Changsheng Xu. Querystream: Advancing streaming video understanding with query-aware pruning and proactive response. In *International Conference on Learning Representations*, 2026b.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18243–18252, 2024.

A Appendix

A.1 AI Use Statement

During the preparation of this manuscript, we used large language models (LLMs) solely to assist with language editing, including grammar correction, sentence refinement, and improvements to clarity and readability. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content of this paper.

A.2 Reproducibility Statement

We have made every effort to ensure the reproducibility of our work. GroundedPlay is built upon publicly available models and standard neural network components. For visual representation, we employ a frozen CLIP encoder to extract frame-level visual features. For commentary generation, we use a frozen Qwen3-VL-8B as the language model.

The main paper provides a detailed description of the proposed architecture, including Lifetime-Aware Memory, Reciprocal Semantic Grounding (RSG), and Evidence-Completion Coupling (ECC). Specifically, we describe the construction and update mechanisms of event memory, entity memory, and episodic memory, the bidirectional interaction between action semantics and participant identity, and the causal completion prediction and response mechanism for online commentary generation. The inference procedure, including the event-ending threshold and response reactivation rule, is also explicitly specified in the main text.

A.3 Implementation Details

A.3.1 Architecture and Optimization

Additional settings for the three training stages are summarized in Table 4.

A.3.2 Training Objectives

As described in Section 3.4, GroundedPlay is trained in three stages. We first independently pretrain the visual trigger, the action module underlying evolving event memory, and the identity prediction module using event-ending, action, and weak identity supervision, respectively. We then jointly optimize these modules together with RSG, ECC, and the learnable retrieval and fusion mechanisms of entity memory. Finally, we freeze the upstream modules and Qwen3-VL-8B and train the scene encoder, visual projectors, and modules that integrate retrieved episodic evidence with current representations for commentary generation. Below, we provide the detailed objectives, loss weights, response-alignment formulation, and supervision used in each stage.

Table 4: Additional settings of GroundedPlay. LR denotes learning rate.

Setting	Configuration
Independent Module Pretraining	
Trigger initial LR	2×10^{-4}
Action initial LR	3×10^{-4}
Identity initial LR	3×10^{-4}
Joint Perception Training	
Trigger / action LR	5×10^{-6}
Identity LR	2×10^{-5}
ECC LR	1×10^{-4}
RSG LR	2×10^{-4}
Entity memory retrieval and fusion LR	3×10^{-4}
Memory entries per predicted identity	4
Commentary Generation Training	
Spatial / temporal Transformer layers	1 / 3
Frame / scene queries	4 / 32
Scene-query decoder layers	2
Base LR	5×10^{-5}
Episodic identity fusion LR	3×10^{-4}
Weight decay	0.01

Stage I: Independent Module Pretraining. We independently pretrain the visual trigger, action module, and identity prediction module. The visual trigger is supervised with Focal Loss for event-ending prediction. The action module is trained with weighted cross-entropy over fine-grained action classes, binary cross-entropy over action attributes, and a margin loss for similar action categories. Focal Loss is additionally applied to the action completion and response predictions, together with a temporal ranking loss on response scores. The identity prediction module is trained with binary cross-entropy for role presence and weakly supervised identity classification on both role-query aggregation and candidate-player aggregation. We further apply a within-event margin loss to player representations and a cross-event contrastive loss across events of the same game.

Stage II: Joint Perception Training. We initialize the model with the independently pretrained visual trigger, action module, and identity prediction module, and jointly optimize these modules together with RSG, ECC, and the learnable retrieval and fusion mechanisms of entity memory.

For ECC, the event-ending loss supervises the completion prediction after interaction, while the final action predictions after identity-guided refinement are supervised by the action classification loss. We retain the action-attribute, class-margin, and action-completion objectives, and additionally use a temporal ranking loss, a CTC loss for action-sequence alignment, and a response-alignment loss. The corresponding objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{TA}} = \mathcal{L}_{\mathrm{end}} + 0.25\mathcal{L}_{\mathrm{rank}} + \mathcal{L}_{\mathrm{act}} + 0.35\mathcal{L}_{\mathrm{attr}} \\ +0.25\mathcal{L}_{\mathrm{margin}} + 0.5\mathcal{L}_{\mathrm{comp}} + 0.5\mathcal{L}_{\mathrm{seq}} + 0.1\mathcal{L}_{\mathrm{align}}.\end{aligned} \tag{4}$$

Here, $\mathcal{L}_{\mathrm{end}}$ and $\mathcal{L}_{\mathrm{comp}}$ use Focal Loss, $\mathcal{L}_{\mathrm{act}}$ is weighted cross-entropy, $\mathcal{L}_{\mathrm{attr}}$ is binary cross-entropy, and $\mathcal{L}_{\mathrm{seq}}$ is the CTC loss. The margins for $\mathcal{L}_{\mathrm{rank}}$ and $\mathcal{L}_{\mathrm{margin}}$ are set to 0.75 and 0.5, respectively.

For response alignment, we use a differentiable approximation of the triggering and reactivation process to encourage the model to produce a single, temporally aligned response for each annotated event ending while suppressing spurious responses. Each soft response is assigned to the nearest annotated ending within six sampled time steps. The response-alignment objective is

$$\mathcal{L}_{\mathrm{align}} = \mathcal{L}_{\mathrm{count}} + \mathcal{L}_{\mathrm{false}} + 0.1\mathcal{L}_{\mathrm{time}}, \tag{5}$$

where $\mathcal{L}_{\mathrm{count}}$ encourages one response per annotated ending, $\mathcal{L}_{\mathrm{false}}$ penalizes unmatched responses, and $\mathcal{L}_{\mathrm{time}}$ penalizes temporal deviation from the assigned ending. This differentiable formulation is used only during training. Inference follows the causal threshold rule described in the main paper.

For RSG, the identity objective is

$$\begin{aligned}\mathcal{L}_{\text{id}} = \mathcal{L}_{\text{name}} + 0.5\mathcal{L}_{\text{role}} + 0.5\mathcal{L}_{\text{cand}} \\ +0.1\mathcal{L}_{\text{intra}} + 0.1\mathcal{L}_{\text{inter}} + \eta\mathcal{L}_{\text{aux}},\end{aligned} \tag{6}$$

where η decreases linearly from 0.5 to 0 during the first half of joint training.

Entity memory is trained with

$$\begin{aligned}\mathcal{L}_{\text{mem}} = \mathcal{L}_{\text{refine}} + 0.2\mathcal{L}_{\text{assoc}} \\ +0.1\mathcal{L}_{\text{qual}} + 0.2\mathcal{L}_{\text{retr}},\end{aligned} \tag{7}$$

covering refined identity prediction, observation association, prediction quality, and retrieved identity prediction.

The overall Stage II objective is

$$\begin{aligned}\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{TA}} + \mathcal{L}_{\text{id}} + \mathcal{L}_{\text{mem}} \\ +0.01\ \text{mean}(\delta^2),\end{aligned} \tag{8}$$

where δ denotes the correction applied to the action-class logits by identity-guided action refinement.

Stage III: Commentary Generation Training. We freeze the upstream modules and Qwen3-VL-8B and train the scene encoder, visual projectors, and episodic identity fusion modules. Commentary generation is supervised with cross-entropy over the target caption tokens:

$$\mathcal{L}_{\text{gen}} = -\sum_{n=1}^{N} \log p_\theta\left(w_n \mid w_{<n}, \mathcal{C}\right), \tag{9}$$

where $\mathbf{w} = (w_1, \ldots, w_N)$ denotes the target caption and $\mathcal{C}$ denotes the conditioning input to Qwen3-VL-8B.

Historical captions are generated during the first epoch and reused as context in subsequent epochs, subject to the causal availability constraint described in Section 3.4.

A.3.3 Evaluation Protocols

We evaluate event response using recall (Rec), precision (Prec), and response time difference (Δt), and assess commentary quality using CIDEr, BLEU-4, player identity accuracy, and fine-grained action accuracy.

Event Matching. Within each video stream, predicted event times and annotated event-ending times are matched one-to-one using the Hungarian algorithm. A prediction and an annotation form a valid match when their absolute time difference is no greater than 2 s. Unmatched predictions are treated as false responses, while unmatched annotations are treated as missed responses. This matching criterion is used to compute recall and precision and to select caption pairs for commentary evaluation. The response time difference is computed separately over all annotated events using the nearest predicted event in the same stream, without applying the 2-second tolerance.

Response Metrics. Recall and precision are defined as

$$\text{Rec} = \frac{N_m}{N_g} \times 100\%, \qquad \text{Prec} = \frac{N_m}{N_p} \times 100\%, \tag{10}$$

where N_m denotes the number of matched event pairs, N_g the total number of annotated events, and N_p the total number of predicted responses. The counts are accumulated over all video streams before computing the final ratios.

The response time difference is defined as

$$\Delta t = \frac{1}{N_g} \sum_{i=1}^{N_g} \min_{p_j \in \mathcal{P}_{s(i)}} \left| t_i^g - t_j^p \right|, \tag{11}$$

where $\mathcal{P}_{s(i)}$ denotes the set of predicted events in the same video stream as the i-th annotated event.

Table 5: Commentary quality on NBA_Streaming benchmark. Each baseline is compared with GroundedPlay using only ground-truth events successfully matched by both methods.

Method	C ↑	B-4 ↑	Id. ↑	Act. ↑
Common with NBA_Streaming				
NBA_Streaming	195.694	35.238	18.289	42.488
GroundedPlay	**331.775**	**46.987**	**37.547**	**54.588**
Common with MATR + LLM-IAVC				
MATR + LLM-IAVC	51.168	17.616	1.487	23.432
GroundedPlay	**360.163**	**48.579**	**39.356**	**59.734**

Table 6: Commentary quality on MatchTime-SID benchmark. Each baseline is compared with GroundedPlay using only ground-truth events successfully matched by both methods.

Method	C ↑	B-4 ↑	Id. ↑	Act. ↑
Common with Qwen3-VL-8B				
Qwen3-VL-8B	6.804	1.166	6.471	17.957
GroundedPlay	**28.090**	**4.987**	**6.689**	**25.075**
Common with VideoLLM-Online				
VideoLLM-Online	7.260	1.840	2.752	6.504
GroundedPlay	**55.379**	**3.842**	**5.310**	**16.667**

Commentary Evaluation. Commentary metrics are computed only between successfully matched predicted captions and reference captions. CIDEr and BLEU-4 are used to evaluate overall caption quality.

Semantic Accuracy. Player names in the predicted and reference captions are compared according to their sentence positions and within-sentence order. A comparison is counted as correct only when both captions contain the same player identity at the corresponding position. Missing, additional, and mismatched identities are counted as errors:

$$\mathrm{Acc}_{\mathrm{id}} = \frac{N_{\mathrm{id}}^{\mathrm{correct}}}{N_{\mathrm{id}}^{\mathrm{union}}} \times 100\%. \tag{12}$$

For fine-grained action accuracy, a shared rule-based parser maps action expressions in the predicted and reference captions to normalized action labels. These labels are compared according to their action category families and occurrence order within each family. A comparison is counted as correct only when both captions contain the same normalized action label at the corresponding position. Missing, additional, and mismatched action labels are counted as errors:

$$\mathrm{Acc}_{\mathrm{act}} = \frac{N_{\mathrm{act}}^{\mathrm{correct}}}{N_{\mathrm{act}}^{\mathrm{union}}} \times 100\%. \tag{13}$$

For each metric, the numerator counts correct comparisons, and the denominator counts all comparisons, including missing, additional, and mismatched predictions. These counts are accumulated over all successfully matched caption pairs before computing the final accuracies.

Streaming Efficiency. Throughput (FPS) is measured on three identical GPUs, with each GPU independently processing one video stream using a batch size of 1. The timing includes video decoding, preprocessing, online feature extraction, event triggering, and complete caption generation. The total number of sampled frames processed by the three streams is divided by the total wall-clock time to obtain the aggregate throughput. Three input streams at 6 FPS correspond to a real-time processing requirement of 18 FPS.

Table 7: Memory ablations on NBA_Streaming benchmark. Each variant is compared with GroundedPlay using only ground-truth events successfully matched by both models.

Memory Configuration	C ↑	B-4 ↑	Id. ↑	Act. ↑
Memory Removal				
w/o Entity Memory	92.997	28.463	0.391	52.413
GroundedPlay	**320.726**	**45.925**	**36.707**	**52.599**
w/o Event Memory	190.146	29.509	24.259	28.511
GroundedPlay	**323.694**	**46.348**	**37.053**	**53.555**
w/o Episodic Memory	312.693	45.876	36.163	52.130
GroundedPlay	**320.016**	**45.924**	**36.592**	**52.659**
Memory Lifetimes and Context				
Reset Entity Per Event	277.034	43.548	30.589	53.075
GroundedPlay	**325.468**	**46.369**	**37.255**	**53.421**
Retain Event Across Events	295.137	44.248	33.808	51.234
GroundedPlay	**323.576**	**46.253**	**37.030**	**53.199**
Retrieve Earlier Events	303.628	43.722	34.161	50.655
GroundedPlay	**320.016**	**45.924**	**36.592**	**52.659**

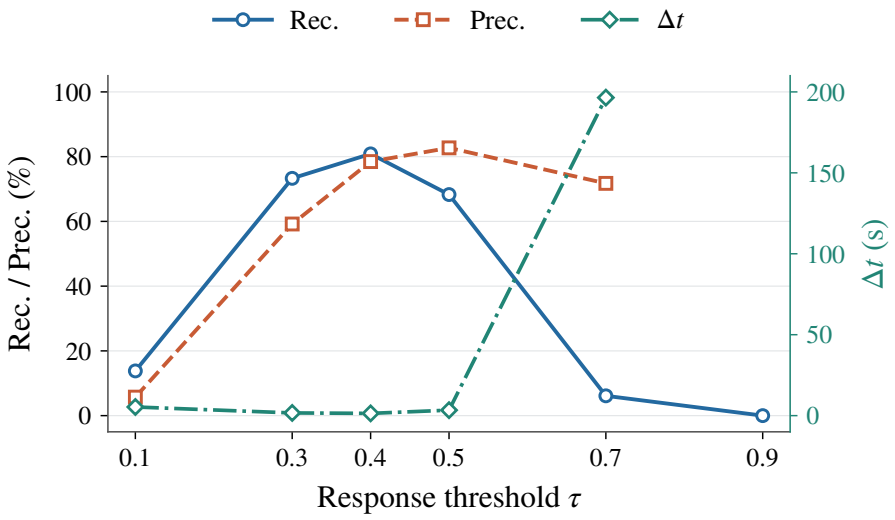


Figure 4: Sensitivity of GroundedPlay to the trigger threshold on NBA_Streaming.

Full-caption latency (Lat) is measured from the arrival of the video frame corresponding to a trigger until the final caption token is generated. It includes the feature extraction, trigger computation, memory processing, and language generation required for that response.

A.4 Quantitative Results

Comparison on Common Matched Events. To control for differences in event coverage, we compare GroundedPlay with each baseline using only ground-truth events successfully matched by both methods. On NBA_Streaming benchmark, GroundedPlay outperforms NBA_Streaming and MATR + LLM-IAVC on all four commentary metrics (Table 5), showing that its gains persist when each pair is evaluated on the same events.



NBA_Streaming:
14min47.33s J. Giddey's driving layup two-pointer is blocked by J. Giddey. J. Giddey blocks J. Giddey's layup two-pointer. A. Black grabs the defensive rebound after J. Giddey's driving layup two-pointer is blocked.
14min52.33s A. Black makes a dunk two-pointer, assisted by T. da Silva.
15min06.33s P. Williams commits a personal foul.

w/o Event Memory
14min47.00s M. Buzelis' driving layup two-pointer is blocked by G. Bitadze; G. Bitadze blocks M. Buzelis' layup two-pointer; G. Bitadze grabs the defensive rebound after M.Buzelis' driving layup two-pointer is blocked.
14min50.17s A. Black is called for a bad pass out-of-bounds turnover.
15min07.17s K. Huerter makes a running layup two-pointer, assisted by M. Buzelis.

w/o Entity Memory
14min46.50s J. Allen's driving layup two-pointer is blocked by J. Allen; J. Allen blocks J. Allen's layup two-pointer; J. Allen grabs the defensive rebound after J. Allen's driving layup two-pointer is blocked.
14min50.17s D. Bane is called for a bad pass out-of-bounds turnover.
15min07.00s J. Allen commits a personal foul.

w/o Episodic Memory
14min46.50s J. Allen's driving layup two-pointer is blocked by J. Black; J. Black blocks J. Allen's layup two-pointer; G. Bitadze grabs the defensive rebound after J. Allen's driving layup two-pointer is blocked.
14min50.17s A. Black is called for a bad pass out-of-bounds turnover.
15min07.00s K. Huerter commits a personal foul.

Ours:
14min46.50s J. Giddey's driving layup two-pointer is blocked by P. Banchero. P. Banchero blocks J. Giddey's layup two-pointer. A. Black grabs the defensive rebound after J. Giddey's driving layup two-pointer is blocked.
14min50.17s A. Black is called for a bad pass out-of-bounds turnover.
15min07.00s F. Wagner commits a personal foul.

Groundtruth:
14min46.00s J. Giddey's driving layup two-pointer is blocked by P. Banchero. P. Banchero blocks J. Giddey's layup two-pointer. A. Black grabs the defensive rebound after J. Giddey's driving layup two-pointer is blocked.
14min51.00s A. Black is called for a bad pass out-of-bounds turnover.
15min07.00s F. Wagner commits a personal foul.

Figure 5: Additional qualitative comparisons on consecutive events from an NBA video stream.

We further conduct the same pairwise comparison on MatchTime-SID benchmark with Qwen3-VL-8B and VideoLLM-Online. As shown in Table 6, GroundedPlay also achieves higher scores across all four metrics on the common matched events, extending these findings to football commentary.

We future compare GroundedPlay with six memory variants on NBA_Streaming benchmark, using only ground-truth events successfully matched by both the full model and the corresponding variant. As shown in Table 7, GroundedPlay outperforms every variant on all four commentary metrics, supporting the benefits of its memory design when event coverage is controlled.

Sensitivity to the Trigger Threshold. We examine how the trigger threshold affects response timing, recall, and precision on NBA_Streaming in Figure 4. Lowering the threshold reduces both recall and precision and increases the response time difference. Raising the threshold improves precision but reduces recall and increases the response time difference. These results suggest that the selected threshold $\tau = 0.4$ balances event coverage, response precision, and temporal accuracy.

A.5 Qualitative Analysis

We present additional qualitative examples in Figure 5.