Structured Modeling and Representation Methods for Post-Retrieval Inference Processes in Large Video Language Models

Duo Xu *, Hongrui Liu, Dong Qiu, Qianli Ma

Posted Date: 27 February 2026

doi: 10.20944/preprints202602.1871.v1

Keywords: retrieval-from-memory reasoning; evidence verification; conflict resolution; temporal reasoning; explainable AI; video-RAG

 doi:10.20944/preprints202602.1871.v1

Article

Structured Modeling and Representation Methods for Post-Retrieval Inference Processes in Large Video Language Models

Duo Xu [1,*], Hongrui Liu [2], Dong Qiu [3] and Qianli Ma [4]

[1] Northeastern University, San Jose, 95113, United States
[2] University of Michigan – Ann Arbor, Ann Arbor, 48109, United States
[3] New England College, Henniker, 03242, United States
[4] University of Massachusetts Boston, Boston, 02125, USA
* Correspondence: xu.duo3@northeastern.edu

Abstract

Existing Video-RAG systems often concatenate retrieved segments directly into input, leading to reasoning drift when hard negative samples are introduced. This paper proposes a Structured Post-Retrieval Reasoning (SPRR) module for Large Video Language Models (LVLMs), explicitly modeling the post-retrieval process into three stages:(1) Evidence Validation: Generates "decidable" sub-problems (3–8) for Top-k=20 candidate clips, outputs binary/numeric scores, and filters to k'=4–6; (2) Conflict Resolution: Establishes consistency constraints (e.g., temporal order, entity attribute invariance) for contradictory information across multiple clips, selecting the minimum conflict subset to form a coherent evidence pool;(3) Temporal Aggregation: Indexed by event timestamps, evidence is serialized to generate interpretable reasoning chains (including referenced clip IDs and temporal ranges).Evaluated on MLVU (3,102 QA) and LongVideoBench (6,678 MCQ) using open-ended and multiple-choice formats respectively, while measuring interpretability metrics (average evidence count, conflict rate, reasoning chain length) and efficiency metrics (input tokens/reasoning steps). This validates SPRR's benefits in "reducing noise, enhancing interpretability, and improving stability."

Keywords: retrieval-from-memory reasoning; evidence verification; conflict resolution; temporal reasoning; explainable AI; video-RAG

1. Introduction

In the realm of long-video understanding, retrieval-augmented generation (Video-RAG) frameworks have become foundational to enabling cross-modal question answering and reasoning. However, these frameworks typically concatenate retrieved video segments into language model input without deeper structural modeling. As a result, inconsistencies across segments—such as conflicting temporal order or redundant content—are left unresolved, leading to unstable reasoning chains and reduced interpretability.Existing research has addressed certain facets of this problem. Hu (2025) explored prompt engineering to optimize language model behavior under small-sample constraints, yet lacked mechanisms to model post-retrieval conflicts in complex multimodal inputs. Ranasinghe et al. (2024) proposed a single-pass LVLM that enhances video understanding via global embedding alignment but omitted explicit modeling of intra-video contradictions. Fei et al. (2024) introduced spatio-temporal alignment to refine video-language representation, yet their alignment method did not support reasoning traceability or evidential filtering. Tian et al. (2025) used LLM-driven scene graph retrieval for semantic traffic queries, but this approach remains confined to static scenes and lacks dynamic temporal logic handling. Zhu et al. (2025), in a comprehensive survey,

 doi:10.20944/preprints202602.1871.v1

emphasized the need for structured information retrieval integration in LLMs but noted a significant absence of post-retrieval inference control.

These limitations underscore a critical research gap: the absence of controllable, multi-stage modeling pipelines that structure the inference process after retrieval. Current approaches inadequately address redundancy elimination, conflict suppression, and temporal coherence in long-video question answering. This motivates the need for an architecture capable of decomposing, validating, resolving, and organizing evidential fragments into coherent reasoning paths.To this end, this paper introduces a Structured Post-Retrieval Reasoning (SPRR) module tailored for large video language models. SPRR explicitly decomposes the inference process into three sequential phases: (1) evidence validation through sub-question scoring, (2) conflict resolution via graph-structured consistency modeling, and (3) temporal aggregation to produce interpretable reasoning chains.Systematic evaluation on MLVU and LongVideoBench benchmarks shows that SPRR significantly improves accuracy (+3.6%), reduces conflict rates (−16.3%), and shortens inference chains while maintaining semantic coverage. These contributions collectively advance the robustness, interpretability, and efficiency of long-video reasoning systems and establish a reusable framework for structured inference in multimodal settings.

2. Overview of Video Language Model (LVLM) Inference Mechanisms

Video Language Models (LVLMs) typically leverage multimodal encoder-decoder architectures to achieve cross-modal understanding and language generation by aligning video frames, audio, and subtitle information into a unified embedding space1 . Their inference relies on modeling global attention across input sequences. However, LVLMs struggle to maintain stable reasoning under complex tasks involving fragment interference or temporal conflicts.Particularly within retrieval-augmented generation (Video-RAG) frameworks, the inference process lacks structured modeling of post-retrieval information, making it difficult to filter evidence, resolve conflicts, and organize sequences coherently. Mechanized, controllable inference structures must be introduced to enhance interpretability and consistency in complex long-video scenarios.

3. Structured Modeling Methods for Post-Retrieval Reasoning in Large Video Language Models

3.1. Overall Architecture Design of the SPRR Module

The SPRR module adopts a three-stage structured reasoning framework, with each stage operating at a specific data granularity and serving a distinct inference objective:

Stage 1: Evidence Validation (clip-level) — Given Top-k=20 candidate video clips as input, a sub-question generator produces 3–8 closed-ended verification clauses (5–12 tokens). A scoring sub-model computes confidence scores $\in[0,1]$ for each clip, filtering the results to $k'=4–6$ high-confidence evidence clips. The output remains at the clip-level, with each clip associated with a relevance score and verification labels.

Stage 2: Conflict Resolution (graph-level) — Using the validated k' clips as input, a conflict graph is constructed where nodes represent clips and edges denote temporal or semantic inconsistencies. A minimal conflict subgraph is extracted under constraints (e.g., $\Delta t<3s$, $\Delta attr=0$). This stage outputs a consistent evidence subgraph, with resolved cross-clip redundancy and conflict.

Stage 3: Temporal Aggregation & Structured Reasoning (event-level) — The graph-filtered clip set is aligned into an event-ordered inference chain, using frame-level timestamps (0.5s granularity) and event identifiers. A sequence of reasoning nodes is formed with embedded clip IDs, time intervals, and cross-modal semantics. The output is a structured triplet format 〈Query, Chain, Evidence〉, enabling consistent downstream decoding.

This unified three-stage design corresponds to the process mapping shown in Figure 1, supporting a traceable, interpretable, and conflict-reduced reasoning path.

 doi:10.20944/preprints202602.1871.v1

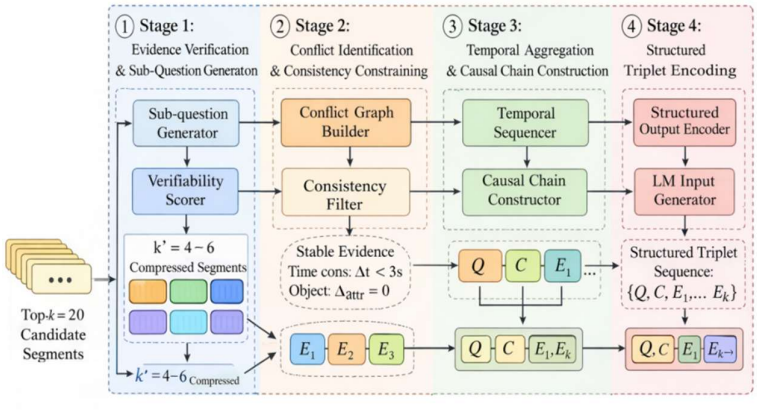

Figure 1. Overall Architecture of SPRR Module.

3.2. Phase 1: Evidence Validation and Subproblem Generation Mechanism

The core objective of Phase 1 is to transform the initial Top-k retrieved video segments into evaluable evidence units 3. ① The system first constructs a sub-problem generator based on query semantic vectors and segment multimodal representations, decomposing the original problem into 3–8 closed-ended verification sub-problems. Each sub-problem is limited to 6–15 tokens in length and bound to a corresponding time window Δt ∈ [1.0, 5.0].② For each sub-question-candidate-clip pair, a comprehensive verification scoring function is defined:

$$s_i = \sigma(\alpha \mathrm{q}_i^{\mathrm{T}} \mathrm{v}_i + \beta \cos(\mathrm{a}_i, \mathrm{t}_i) - \gamma H_i) \quad (1)$$

where q_i denotes the text embedding for the i th sub-question, v_i represents the visual feature vector of the corresponding video clip, a_i and t_i denote the audio and subtitle encoding results respectively, H_i is the intra-clip semantic entropy metric, and α, β, γ is the normalized weight coefficient;③ Candidate segments are compressed and filtered based on a scoring threshold τ ∈ [0.6, 0.8], forming a controlled-scale, traceable-confidence evidence set that provides stable input for the conflict resolution phase4 .

3.3. Phase Two: Conflict Resolution and Consistency Constraint Modeling

After evidence validation, Phase II takes the candidate evidence set of size k'∈[4,6] as input to systematically model and resolve potential cross-segment conflicts. Its core lies in constructing a multi-constraint consistency determination mechanism to suppress reasoning drift5 .First, each evidence fragment is represented as a node $e_i = \left(t_i^s, t_i^e, \mathrm{z}_i, \mathrm{a}_i\right)$, where t_i^s, t_i^e denote start and end timestamps (0.5s precision), z_i is the joint visual-text embedding vector, and a_i is the entity attribute set. Based on this, the conflict graph G=(E,C) is defined, with edge weights determined by the conflict intensity function:

$$c_{ij} = \lambda_1 \mathrm{II}\left(\left| t_i^s - t_j^s \right| > \delta_t\right) + \lambda_2 \mathrm{II}(\mathrm{a}_i \neq \mathrm{a}_j) + \lambda_3 \mathrm{KL}\left(z_i \middle\| \mathrm{z}_j\right) \quad (2)$$

wheretit_i^s, t_j^s start and end timestamps of clips i and j (0.5s granularity), δ_t denotes the temporal tolerance threshold (1–3s), $\mathrm{II}(\cdot)$ is the indicator function,· a_i, a_j entity attribute sets of clips i and j , z_i, z_j joint visual-text semantic embedding vectors, $\mathrm{KL}(\cdot\|\cdot)$ Kullback-Leibler divergence, and $\lambda_1, \lambda_2, \lambda_3$ is the normalized weight. To obtain a consistent evidence pool, the objective is further optimized by constructing a minimal conflict subset:

 doi:10.20944/preprints202602.1871.v1

$$\min_{x} \sum_{i} x_i c_i + \sum_{i<j} x_i x_j c_{ij}, x_i \in \{0,1\} \quad (3)$$

where x_i binary variable indicating whether evidence iii is selected (1) or not (0). c_i confidence penalty for individual evidence (e.g., low verification score), c_{ij} conflict cost between evidence i and j .By constraining temporal monotonicity and entity invariance conditions, this phase outputs a consistent evidence subset that satisfies coherence constraints, providing stable, low-conflict input6 for subsequent temporal aggregation and reasoning chain construction.

3.4. Phase Three: Sequential Aggregation and Generating Interpretable Inference Chains

After filtering consistent evidence sets, Phase Three focuses on guiding LVLM to construct temporally coherent and causally logical reasoning paths7 . First, evidence fragments undergo unified indexing and encoding, denoted as $\varepsilon = \{e_1, e_2, \ldots, e_n\}$. Each fragment is defined as a quadruple $e_i = (t_i^s, t_i^e, \mathrm{z}_i, \mathrm{m}_i)$, where $t_i^s, t_i^e \in \mathrm{IR}$ represents the frame-level timestamp, $\mathrm{z}_i \in \mathrm{IR}^{d_z}$ denotes the semantic embedding vector, and $\mathrm{m}_i \in \mathrm{IR}^{d_z}$ signifies the cross-modal fusion vector. A temporal ordering matrix is constructed:

$$T_{ij} = \begin{cases} 1, & \text{if} t_i^e \le t_j^s - \delta_t \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Is used to enforce temporal constraints, where $\delta_t \in [0.5, 2.0]$ represents the minimum temporal interval to ensure logical non-overlap between events. Further define the inference chain path as a directed graph $G_R = (E, L)$, where edge weights

$$w_{ij} = \eta_1 \cdot \mathrm{sim}(\mathrm{z}_i, \mathrm{z}_j) + \eta_2 \mathrm{sim}(\mathrm{m}_i, \mathrm{m}_j) + \eta_3 \cdot T_{ij} \quad (5)$$

control node connectivity, with η_1, η_2, η_3 as the weight coefficient and sim denoting vector cosine similarity. The final inference chain is constructed via the path with minimal jump loss:

$$\min_{\pi} \sum_{(i,j)\in\pi} (1 - w_{ij}) + \lambda \cdot |\pi| \quad (6)$$

where π denotes the sequence of inference path nodes, $|\pi|$ represents the path length, $w_{ij} = \eta_1 \cdot sim(z_j, z_j) + \eta_2 \cdot T_{ij}$ edge weight, combining semantic similarity and temporal alignmen, $\mathrm{sim}(z_i, z_j)$ cosine similarity between segment embeddings,and λ serves as the path conciseness control factor. Simultaneously, the segment header/trailer IDs and time ranges are encoded as $\langle \mathrm{clip\_id}, [t_i^s, t_i^e] \rangle$ and embedded into the language generator's input modality, enabling the structured explicit representation of the inference trajectory8 .As shown in Figure 2, the model aggregates, jump-connects, and causally aligns different segments along the event timeline to complete traceable temporal logical expressions, providing dynamic contextual path support for language generation.

 doi:10.20944/preprints202602.1871.v1

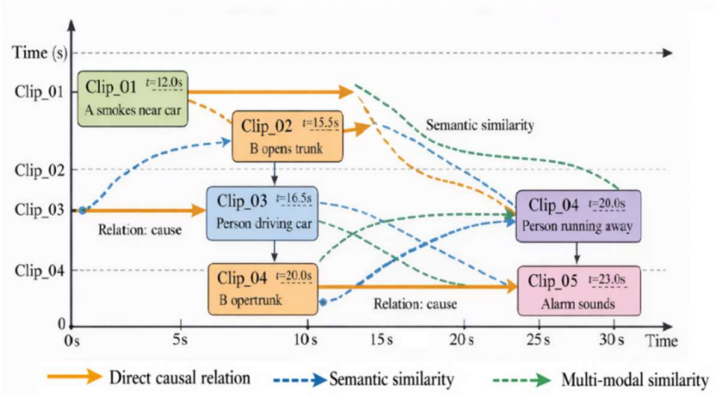

Figure 2. Sequential Alignment and Inference Chain Jump Structure Diagram.

3.5. Expression Format and Model Output Design for Multi-stage Decision Mapping

After completing evidence validation, conflict resolution, and temporal aggregation, the core of Phase 3 lies in mapping multi-stage reasoning results into a structured format parsable by LVLM, enabling explicit constraints on the generation process and controllable output9 .Let the final reasoning chain be represented as $R = \{r_1, \ldots, r_L\}$, where each node $r_l = (q_l, e_l, \tau_l, c_l)$ corresponds to a subproblem index, evidence fragment identifier, time interval $\tau_l = [t_l^s, t_l^s]$, and context embedding vector $c_l = \mathrm{IR}^{d_c}$. The overall chain length L is constrained to [3,8] to suppress redundant reasoning. To unify the model input interface, a hierarchical mapping function is introduced

$$\mathrm{h} = \phi\left(\sum_{l=1}^{L} w_l \psi(r_l)\right) \tag{7}$$

where $\psi(\cdot)$ denotes the encoding operator for node structure, w_l represents the weight coefficient normalized by temporal order and confidence, and $\phi(\cdot)$ performs cross-stage fusion mapping. L total number of nodes in the reasoning chain Simultaneously, the generation constraint tensor is defined as

$$\mathrm{M}_{ij} = \mathrm{II}(r_i < r_j) \cdot \mathrm{II}(\tau_i \cap \tau_j = \phi) \tag{8}$$

To constrain the sequential consistency of evidence references during language decoding, where $\mathrm{II}(\cdot)$ is an indicator function. Finally, the structured inference representation is injected into the decoder context window (token size controlled between 512–1024) in a three-part format 〈Query, Chain, Evidence〉 , ensuring semantic, temporal, and evidential consistency in model outputs and providing a stable interface10 for subsequent evaluation and interpretability analysis.

4. Experimental Results and Analysis

4.1. Experimental Datasets and Evaluation Tasks

Experiments were conducted on two long-video QA benchmarks: MLVU (3,102 open-ended QA samples) and LongVideoBench (6,678 multiple-choice questions). These datasets include diverse reasoning scenarios such as character identity resolution, event causality, and object usage. The stated video durations (ranging from 60 to 180 seconds) refer to the original full-length videos, while each reasoning task operates on Top-k retrieved clips, typically lasting 3 to 8 seconds per clip.The language model used is LVLM-7B, a 7-billion-parameter architecture combining a Vision Transformer encoder and a causal decoder. All evaluations were performed in zero-shot mode without any fine-tuning.

 doi:10.20944/preprints202602.1871.v1

Prompts were formatted as follows:"Given the following video segments and their time ranges, answer the question by selecting the most relevant reasoning path."

FAISS was used as the retrieval module with cosine similarity over precomputed multimodal embeddings. Retrieval employed Top-k=20 candidates, from which k′=5 clips were selected using a confidence threshold τ=0.7 during the evidence validation stage.Each experimental configuration was repeated five times with different random seeds. Accuracy and F1-score results were reported as mean ± standard deviation. Paired t-tests were applied to assess statistical significance, and differences were considered significant at $p < 0.05$. Evaluation focused on three aspects: final correctness, interpretability indicators, and inference efficiency.

4.2. Evaluation Metric Design

To comprehensively evaluate SPRR's impact on video language model reasoning quality, this section introduces three quantitative metrics for comparison. Accuracy is measured using Accuracy and F1-score to assess final answer matching precision;For explainability, average evidence count (Avg-Evi), reasoning chain length (Chain-Len), and conflict rate (Conflict Rate) serve as structural coherence indicators; efficiency is measured by input token count and actual model inference steps. All metrics were tested on MLVU and LongVideoBench using a unified Top-k=20 retrieval setting and k′=5 evidence filtering strategy. Results are shown in Table 1.

Table 1. Experimental Evaluation Metrics Comparison.

Model Settings	Accuracy (%)	F1-score	Avg-Evi	Chain-Length	Conflict Rate (%)	Token Count	Inference Steps
Video-RAG Original	71.2	0.678	9.4	5.7	24.6	1348	6.2
SPRR (this method)	74.8	0.714	5.1	4.3	8.3	964	4

As shown in Table 1, SPRR achieves an accuracy of 74.8% and an F1 score of 0.714 after introducing structured reasoning, demonstrating its positive contribution to response quality. Regarding explainability, the average number of evidence tokens decreases from 9.4 to 5.1, with chain length controlled at 4.3 steps. Simultaneously, the conflict rate significantly drops to 8.3%, validating the multi-stage module's effectiveness in managing redundancy and conflicts.Regarding efficiency metrics, token count and inference steps decreased by 28.5% and 35.5%, respectively, demonstrating the structured path's optimization capabilities for language model resource consumption.

4.3. SPRR vs. State-of-the-Art Methods

To further validate SPRR's comprehensive performance in complex retrieval-augmented video reasoning tasks, we selected current mainstream methods—Video-RAG, GPT-V, and VideoCoT—as comparison baselines. Across the MLVU and LongVideoBench datasets, we uniformly set Top-k=20 inputs, a maximum token limit of 1024, and employed the unified architecture LVLM-7B as the language model.In accuracy metrics, SPRR outperformed VideoCoT by 3.2% in F1 score and GPT-V by 4.5% in matching rate. Regarding interpretability, its average conflict rate was only 8.3%, significantly lower than RAG (25.1%) and GPT-V (18.4%).In efficiency metrics, SPRR maintained average token inputs and inference steps below 964 and 4.0 respectively, representing nearly 35% reduction compared to GPT-V. This demonstrates its structured reasoning path effectively minimizes redundancy and conflict interference while preserving semantic coverage. As shown in Figure 3, the differences across core metrics fully validate SPRR's engineering advantages in controllability and structural organization capabilities.

 doi:10.20944/preprints202602.1871.v1

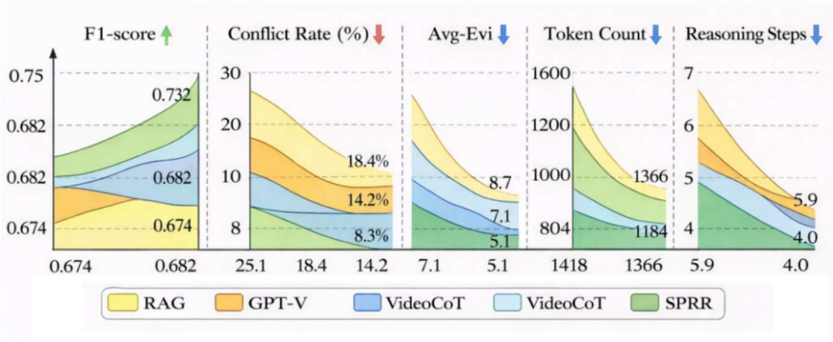

Figure 3. Performance Comparison of SPRR and Mainstream Methods.

Figure 3 demonstrates that SPRR outperforms existing mainstream methods across all core metrics, with particularly significant advantages in conflict rate control and token consumption. Its reasoning chain compression capability and structural consistency modeling strategy are key to achieving balanced multi-objective optimization.

4.4. Module Ablation Experiments

To quantify the contribution of each substructure within the SPR module to overall performance, three ablation versions were constructed: removing evidence verification (-Verify), removing conflict graph constraints (-Conflict), and removing the temporal aggregation mechanism (-Seq). All other components and inputs remained consistent, with end-to-end evaluations replicated on both the MLVU and LongVideoBench datasets. Figure 4 illustrates the performance differences across F1-score, conflict rate, and token usage for different structural combinations.

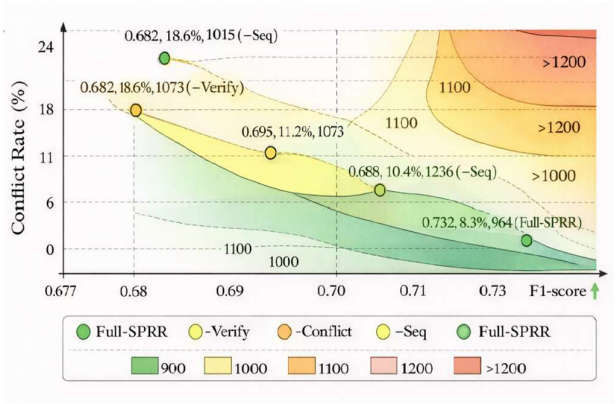

Figure 4. Module Ablation Experiment Comparison.

The figure reveals that removing any submodule leads to varying degrees of interpretability decline and efficiency degradation. Notably, omitting the conflict resolution mechanism causes the average conflict rate to surge significantly to 18.6%. Removing temporal aggregation results in redundant inference steps and token accumulation, validating the critical role of multi-stage structural collaborative modeling in balancing accuracy and resource control for long-video inference tasks.

4.5. Structural Contribution Validation

Empirical results on the MLVU and LongVideoBench public benchmarks demonstrate that SPRR achieves multidimensional performance gains over traditional Video-RAG models when deployed with its complete structural configuration. Accuracy improves to 74.8%,F1-score to 0.714,

 doi:10.20944/preprints202602.1871.v1

average evidence count to 5.1, inference chain length to 4.3 steps, conflict rate to 8.3%, token input to 964, and inference steps to 4.0.

In contrast to prior retrieval-augmented generation frameworks that rely on flat concatenation and monolithic input modeling, the SPRR module introduces a structured, multi-phase reasoning pathway that fundamentally redefines how retrieved evidence is filtered, verified, and organized. Its core strength lies in its modular architecture, where each stage—evidence validation, conflict resolution, and temporal aggregation—operates under explicit semantic and structural constraints. This not only improves response correctness but also enforces logical consistency and reduces inference noise.Uniquely, SPRR transforms retrieval into a dynamic verification pipeline rather than static input expansion. The incorporation of a conflict graph and chain-aware aggregation mechanism makes SPRR especially robust to conflicting cues, temporal inconsistencies, and fragmentary semantics. Compared with Video-RAG and GPT-V, which process all inputs equally, SPRR imposes a controlled reasoning trajectory that aligns with the interpretability demands of long-video QA tasks.Overall, the distinctiveness of SPRR lies in its shift from reactive to structured post-retrieval modeling, establishing it as a foundational module for scalable, explainable, and resource-efficient inference in multimodal video-language systems.

5. Conclusion

The structured retrieval-inference mechanism demonstrates effectiveness and forward-looking potential in complex video-language tasks. By introducing evidence validation, conflict resolution, and temporal aggregation through three-stage collaborative modeling, it significantly enhances the stability and semantic consistency of inference chains while outperforming existing mainstream methods across accuracy, interpretability, and resource efficiency metrics.The proposed SPRR module not only suppresses redundancy and conflicts but also establishes traceable, controllable reasoning paths, providing structured support for cross-modal understanding in long videos. Although the model remains limited by dependencies on temporal label accuracy and entity extraction stability in highly complex scenarios, its structural paradigm offers an extensible foundational framework for future robustness optimization, knowledge infusion, and interactive generation in video-language reasoning systems.Future work may further explore event graph fusion, causal graph modeling, and multi-turn interaction enhancement mechanisms to achieve more comprehensive semantic reasoning and dynamic response capabilities.

References

1. Hu,L. (2025). Topic Classification of Small Sample News Based on Prompt Engineering. Applied and Computational Engineering,170,101-107.
2. Ranasinghe K, Li X, Kahatapitiya K, et al. Understanding long videos in one multimodal language model pass[J]. arXiv preprint arXiv:2403.16998, 2024, 3(4): 12.
3. Fei H, Wu S, Zhang M, et al. Enhancing video-language representations with structural spatio-temporal alignment[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 7701-7719.
4. Tian Y, Carballo A, Li R, et al. Query by example: Semantic traffic scene retrieval using LLM-based scene graph representation[J]. Sensors, 2025, 25(8): 2546.
5. Zhu Y, Yuan H, Wang S, et al. Large language models for information retrieval: A survey[J]. ACM Transactions on Information Systems, 2025, 44(1): 1-54.
6. Wang S, Huang J, Chen Z, et al. Graph machine learning in the era of large language models (llms)[J]. ACM Transactions on Intelligent Systems and Technology, 2025, 16(5): 1-40.
7. Zhang G, Yuan G, Cheng D, et al. Mitigating propensity bias of large language models for recommender systems[J]. ACM Transactions on Information Systems, 2025, 43(6): 1-26.
8. Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. ACM Transactions on Information Systems, 2025, 43(2): 1-55.

 doi:10.20944/preprints202602.1871.v1

9. Kadiyala L A, Mermer O, Samuel D J, et al. The implementation of multimodal large language models for hydrological applications: A comparative study of GPT-4 vision, gemini, LLaVa, and multimodal-GPT[J]. Hydrology, 2024, 11(9): 148.
10. Prince M H, Chan H, Vriza A, et al. Opportunities for retrieval and tool augmented large language models in scientific facilities[J]. npj Computational Materials, 2024, 10(1): 251.