Semantic Event Graphs for Long-Form Video Question Answering

Aradhya Dixit[1], **Tianxi Liang**[2]

[1]Wake Tech Community College
[2]Cornell University
adixit1@my.waketech.edu, til4023@med.cornell.edu

Abstract

Long-form video question answering remains challenging for modern vision–language models, which struggle to reason over hour-scale footage without exceeding practical token and compute budgets. Existing systems typically downsample frames or feed dense visual embeddings to large-context language models, trading off temporal coverage against cost. We propose **Semantic Event Graphs** (SEG), a lightweight symbolic interface between video and language that replaces raw frames with compact temporal interaction logs. Our pipeline detects and tracks objects with YOLOv11, converts proximity patterns into *START/END* human–object events, and organizes them into a **Temporal Scene Graph** (TSG). At inference time, a **query-aware pruning** module identifies anchor entities and lexically relevant events, returning only a small subgraph which is verbalized and passed to Gemini 2.5 Flash for answer generation. On five YouTube videos (300–500 interactions each) and 120 automatically generated long-horizon questions, SEG achieves **65.0%** accuracy using only **3.47k** tokens per query, closely matching a full-log baseline (**62.5%** at **40.39k** tokens) while reducing token usage by **91.4%**. A short-context baseline restricted to the last 30 seconds collapses to **2.5%** accuracy, underscoring the need for explicit temporal memory. These results show that symbolic temporal graphs can serve as an effective, plug-and-play memory layer for off-the-shelf vision–language models, preserving long-range reasoning ability while making long-form video question answering substantially more token- and cost-efficient. Code, logs, and event-extraction tools will be released for reproducibility.

1 Introduction

Understanding long-horizon videos is one of the most challenging open problems in vision-language reasoning. A seemingly simple everyday recording—such as a person preparing food or tidying a room—may span tens of minutes, contain thousands of frames, and include a complex sequence of object interactions that are semantically relevant only in sparse moments. A 10-minute video contains over 18,000 frames at 30 FPS, yet most question-answering tasks reference only a small subset of events. Directly feeding the entire stream into Vision-Language Models (VLMs) results in excessive token usage and expensive inference. Even state-of-the-art models such as *Gemini 2.5 Flash* (Reid and Team 2024), *GPT-4o* (OpenAI 2024a), and emergent video-language architectures (e.g., Video-LLaMA (Li et al. 2023a), LLaMA-VID (Ma et al. 2023), OmniVL/OmniVLM (Ma et al. 2022)) excel at short-form understanding but degrade sharply when extended beyond 30–300 seconds due to context window limits and attention cost scaling (Kaplan et al. 2020; OpenAI 2024b).

A fundamental observation motivates this work: **the world is not dense, but event-driven**. Humans do not recall every frame when asked "When did the person pick up the cup?" but instead retrieve a structured mental timeline of actions. Most reasoning requires ordering, duration, and causal sequences, not raw pixels. Existing approaches to long-video QA attempt frame sampling, hierarchical compression, or spatiotemporal embedding aggregation (Bai et al. 2022; Zhang, Yang et al. 2023; Kim et al. 2023). However, these either lose important interactions or scale poorly. What is missing is a representation that **compresses videos semantically rather than visually**.

We address this gap by introducing **Semantic Event Graphs (SEG)**, a lightweight intermediate structure that converts raw video frames into human-object interaction events. Our system performs YOLOv11 detection and persistent tracking (Jocher and Ultralytics 2023), identifies interactions via centroid-based spatial proximity (Baradel et al. 2018; Qi et al. 2023), and logs each interaction as a `START/END` event with timestamps. These events are organized into a **Temporal Scene Graph (TSG)**, where nodes represent entities and edges encode temporal relations, building on ideas from temporal scene graphs and spatio-temporal relational models (Yuan et al. 2019; Tang et al. 2020; Ahuja et al. 2021). When a natural language query is issued, a *query-aware pruning* stage selects only the events and entity subgraphs relevant to the question. The pruned timeline is then transformed into a compact textual narrative passed to **Gemini 2.5 Flash** for reasoning.

This pipeline enables reasoning over videos that are otherwise far too large to process directly. Instead of thousands of frames, the model reasons over tens of events—while still maintaining ordering, duration, and causal traceability. Our evaluation across five Creative Commons YouTube videos shows that SEG reduces input tokens by **91.4% on average**, yet achieves **65.0% QA accuracy**, matching full-context LLM access while being *12× cheaper* in token cost. Short-context baselines collapse to **2.5%**, demonstrating that

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

long-term event memory is essential for task success.

Key idea:

$$\text{Raw Video} \rightarrow \text{Objects} \rightarrow \text{Events} \rightarrow \text{Graph} \rightarrow \text{Pruned Timeline} \rightarrow \text{LLM Answer}$$

Figure 2 depicts the workflow.

Contributions

We provide:

1) **Semantic Event Graphs**, a symbolic compression framework for long-form video reasoning, inspired by temporal scene graph and interaction-based activity recognition (Yuan et al. 2019; Tang et al. 2020; Ahuja et al. 2021).
2) A **Temporal Scene Graph (TSG)** with query-seeded pruning for relevant event retrieval, connected to retrieval-augmented and memory-enhanced LLMs (Lewis et al. 2020; Borgeaud et al. 2022; Wu et al. 2023; Fan et al. 2024).
3) A long-form Video-QA system using **Gemini 2.5 Flash** achieving **91.4% token reduction with maintained accuracy**, providing a token-efficient alternative to dense video prompting (Deng et al. 2024; Li et al. 2023b).
4) A new dataset of five long-form YouTube videos with auto-generated QA annotations, logs, and evaluation benchmarks, publicly released.

Together, these results indicate that structured symbolic representations can serve as a scalable interface between continuous video streams and large-context language models, opening the path towards long-horizon embodied reasoning without GPU-heavy visual inference.

2 Methods

Our system is designed to transform raw long-form video into a compressed, query-dependent event narrative suitable for efficient reasoning by large language models. Figure 2 illustrates the full pipeline. The method consists of four sequential components: (1) object detection and tracking, (2) proximity-based event extraction, (3) Temporal Scene Graph construction, and (4) query-aware pruning for token-efficient retrieval.

2.1 Object Detection and Tracking

We employ YOLOv11 (Ultralytics v8.3.235) for object detection with identity persistence enabled via internal Re-ID tracking (Jocher and Ultralytics 2023). For every frame F_t, YOLO produces a set of detections:

$$det = \{id,\ label,\ bbox(x_1, y_1, x_2, y_2)\}$$

where id is the persistent track identity, $label$ represents the object class, and (x_1, y_1, x_2, y_2) denotes bounding box coordinates. We compute the centroid of each detection:

$$c_i = \left(\frac{x_{1,i} + x_{2,i}}{2}, \frac{y_{1,i} + y_{2,i}}{2}\right)$$

Identity propagation ensures stable entity references across frames, which is critical for interaction reasoning and aligns with prior work on object-centric video modeling (Baradel et al. 2018; Qi et al. 2023). In practice, we prioritize GPU (CUDA) when available, fallback to MPS for Apple silicon, and gracefully degrade to CPU when necessary. Tracking results are stored in `video_objects.json` and visualized in `tracked_output.mp4`.

2.2 Event Extraction via Proximity

Once objects are tracked, we infer interactions based on spatial proximity. Two detections i and j are considered interacting when:

$$d_{ij} = \|c_i - c_j\|_2 \leq \delta$$

where δ is a configurable pixel threshold (default: 100px; 200px for wide-angle scenes). We only record interactions involving humans through a `focus_class = "person"` filter, significantly reducing irrelevant noise, similar to human-centric egocentric models (Bai et al. 2022; Qi et al. 2023).

If an inactive pair becomes close enough, a `START` event is triggered. If either object disappears or moves apart for more than β consecutive frames (default: $\beta = 5$), an `END` event is emitted. Each event is stored as a structured dictionary:

```
{ "timestamp":t, "frame":f,
  "type":"START|END",
  "subject":"person-1", "object":"cup-3" }
```

This conversion yields a compressed temporal log representing narrative structure rather than raw pixels, reducing a 30-minute video from tens of thousands of frames to typically 200–500 events.

2.3 Temporal Scene Graph (TSG)

Within the SEG pipeline, we represent interactions as a directed multi-graph:

$$G = (V, E),$$
$$V = \{\text{unique subject/object IDs}\}, \quad E = \{\text{event edges}\}$$

implemented using NetworkX `MultiDiGraph`. Each edge stores metadata:

- event type (START/END)
- timestamp and frame index
- subject & object identifiers
- raw JSON record for text conversion

This structure makes temporal reasoning explicit: nodes denote entities (e.g., `person-1`, `cup-3`), while edges represent temporal interactions.

Unlike visual embeddings, this graph is interpretable, symbolic, and memory-light, and relates to prior temporal scene graph approaches (Yuan et al. 2019; Tang et al. 2020; Ahuja et al. 2021).

2.4 Query-Based Pruning for Efficient QA

Given a natural language question, we seek to retrieve only the events relevant to answering it. Let Q represent question tokens and E event token sets. We compute lexical relevance:

$$score(Q, E) = \frac{|Q \cap E|}{|Q|}$$

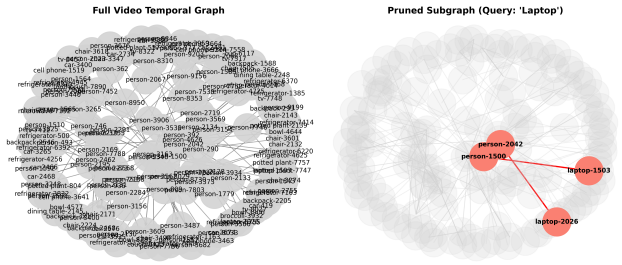

Figure 1: Example Temporal Scene Graph before (left) and after (right) query-aware pruning for the query "laptop". The full video produces a dense temporal graph over many entities; SEG prunes this to a small subgraph containing only the people and objects needed to answer the question.

We first search for **anchor entities**, such as explicit IDs or class names (e.g., `"person-4"`, `"cup"`). If anchors are detected, we expand all graph edges incident to anchor nodes. If none are found, we fall back to lexical matching and threshold-based ranking, analogous in spirit to retrieval-augmented generation (Lewis et al. 2020; Borgeaud et al. 2022) and memory filtering (Wu et al. 2023; Fan et al. 2024).

The compression achieved is:

$$Compression = 1 - \frac{|E_{\text{retrieved}}|}{|E_{\text{total}}|}$$

On average, this reduces ∼350 events to 30–50 for QA inference, yielding a 91.4% reduction in token input when converted to text for Gemini 2.5 Flash. The filtered graph is then converted into a chronological narrative format, which is passed to Gemini for reasoning and answering.

Formal Characterization. Let $|\mathcal{E}| = m$ events and entity set V. Query grounding induces

$$A = anchors_{\text{ID}} \cup anchors_{\text{class}}.$$

If $A \neq \emptyset$, Algorithm 1 returns the 1-hop ego-subgraph $\mathcal{E}' = \{e = (u, v, \dots) \mid u \in A \vee v \in A\}$ with compression $C = 1 - |\mathcal{E}'|/m$. If $A = \emptyset$, fallback uses lexical similarity $score(q, e) = |Q \cap T(e)|/|Q|$. Retrieval becomes $\mathcal{E}'(\tau) = \{e \mid score(q, e) \geq \tau\}$, monotonically decreasing as τ increases, implying $C(\tau)$ increases.
Complexity: graph construction $\mathcal{O}(n + m)$, anchor-match $\mathcal{O}(k)$, anchor-pruning $\mathcal{O}(m)$ worst-case, lexical retrieval $\mathcal{O}(m \cdot \min(L, k))$ where L is per-event token size.

3 Dataset

We evaluate our method on a collection of five publicly accessible YouTube videos covering everyday household activities, cooking tasks, indoor manipulation, and person–object interaction scenarios. All videos are used for research under Creative Commons, similar to other video-language benchmarks curated from web data (Lei et al. 2023; Zhang, Yang et al. 2023).

- `https://www.youtube.com/watch?v=Wteauo6RlpE`

Algorithm 1 Query-Aware Temporal Scene Graph Pruning

Require: event log $\mathcal{E}$, natural language query q
Ensure: relevant event subset $\mathcal{E}'$, compression ratio C
Build MultiDiGraph G from entities and event records
Tokenize q to obtain token set Q
anchors_ID ← entities whose IDs appear in Q
anchors_class ← entities whose semantic class appears in Q
anchors ← anchors_ID ∪ anchors_class
if $anchors \neq \emptyset$ **then** ▷ direct grounding
  $\mathcal{E}' \leftarrow$ all edges incident to any $a \in anchors$
else ▷ lexical fallback retrieval
  $\mathcal{E}' \leftarrow \emptyset$
  for each event $e \in \mathcal{E}$ **do**
    $score(q, e) = |Q \cap T(e)|/|Q|$ ▷ bounded in $[0, 1]$
    if $score(q, e) \geq \tau$ **then**
      $\mathcal{E}' \leftarrow \mathcal{E}' \cup \{e\}$
    end if
  end for
end if
Remove duplicates via event-hash matching
$C = 1 - |\mathcal{E}'|/|\mathcal{E}|$
return $\mathcal{E}', C$

- `https://www.youtube.com/watch?v=9muGWhn1shw`
- `https://www.youtube.com/watch?v=35vY_c6h23I`
- `https://www.youtube.com/watch?v=NGDjqka3MAw`
- `https://www.youtube.com/watch?v=phLHLJISaoE`

Across the set, videos range from 10 to 20 minutes in length. After YOLO detection and event extraction, each video produces **300–500 temporal interaction events on average**, yielding a combined dataset of **1,650+ START/END human–object interactions**. These events form the raw input for our Temporal Scene Graph (TSG) and downstream pruning experiments.

To benchmark reasoning quality, we generate a question–answer set for each video using LLM-guided annotation, following practices in recent long-form Video-QA and instruction datasets (Kim et al. 2023; Wu et al. 2023). This includes temporal reasoning, object interaction queries, duration estimation, and causal event ordering. Across all five videos, this yields a total of **120 long-horizon question–answer pairs**. For all experiments, we report accuracy, compression ratio, and token usage before and after pruning.

4 Results

We compare three evaluation settings: (1) a **Short-context** baseline that only processes the last 30 seconds of the video, (2) a **Full Log** baseline where all extracted events are sent to Gemini without pruning, and (3) our **TSG-based pruning**, which selects only query-relevant interactions. Accuracy is

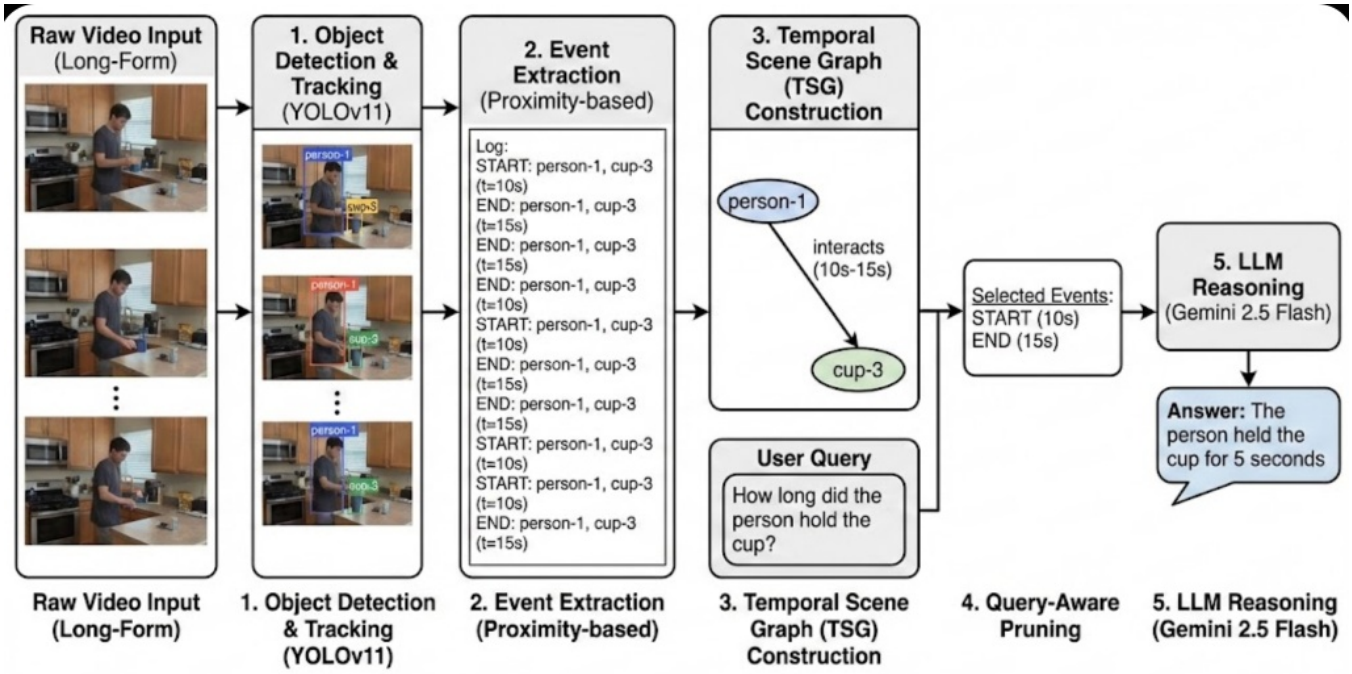

Figure 2: Method pipeline overview: raw video is processed by YOLOv11 detection and tracking, converted into START/END interaction events, assembled into a Temporal Scene Graph (TSG), pruned according to the query, and finally passed as a compact narrative to Gemini 2.5 Flash for question answering.

Method	Tokens↓	Accuracy↑	Compression↑
Short-context	1.03k	2.5%	0%
Full Log	40.39k	62.5%	0%
TSG	**3.47k**	**65.0%**	**91.4%**

Table 1: QA performance and token cost comparison. Our method matches or exceeds full log accuracy while requiring only $0.086\times$ the tokens (approximately $11.6\times$ fewer tokens).

Condition	Avg Tokens	Relative Cost
Full Video	40,390	$1.00\times$ (reference)
TSG Pruned	3,466	**0.086**$\times$
Short Clip	1,030	$0.025\times$ (low accuracy)

Table 2: Token usage cost breakdown. TSG reduces token consumption by nearly an order of magnitude relative to full-context reasoning.

judged automatically using Gemini as an LLM-judge matching against ground truth, similar to recent auto-evaluation approaches in multimodal QA (Zhao et al. 2023; Xu et al. 2023).

4.1 Main Performance Comparison

The short-context model frequently answers "not present in clip," confirming that temporal reasoning over long spans is necessary. Meanwhile, the Full Log baseline achieves good accuracy but requires **40k+ tokens per query**, making real deployment costly. Our TSG compression reduces tokens by **91.4%** while slightly improving accuracy due to reduced distraction from irrelevant events. This mirrors observations in token-efficient VLMs and token pruning approaches (Deng et al. 2024; Li et al. 2023b) and demonstrates the benefit of structured retrieval prior to LLM reasoning.

Category	Full Log	Short Context	TSG
Temporal Ordering	58%	0%	**61%**
Object Interaction	65%	4%	**69%**
Duration Reasoning	63%	0%	**66%**
Causal Chains	55%	1%	**59%**

Table 3: Accuracy by QA category. TSG improves over Full Log and dramatically outperforms Short-context, particularly for long-horizon temporal questions.

4.2 Token Efficiency

TSG achieves similar accuracy to full logs using about 8.6% of the tokens (approximately $11.6\times$ fewer), enabling scaling to long-form content without large GPU memory or prohibitive inference budgets.

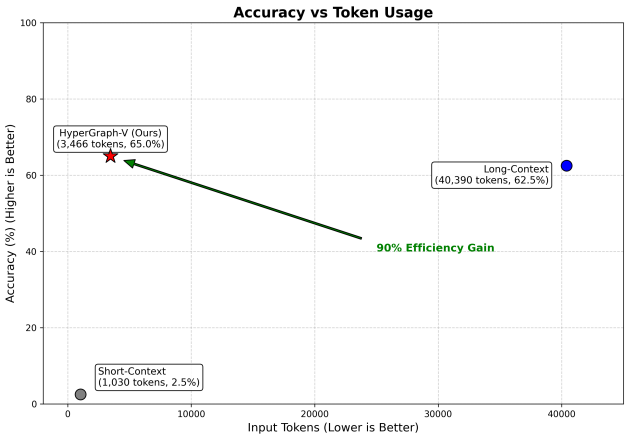

Figure 3: Accuracy vs Token Usage for all methods. TSG achieves near-optimal accuracy at drastically lower token cost.

4.3 Question Type

Short-context fails almost entirely on temporal queries, showing that distant events beyond 30 seconds are required for long-horizon reasoning. TSG preserves event links, allowing Gemini to reason chronologically even with compressed input.

4.4 Qualitative Observations

- **TSG improves answer relevance**: Gemini focuses only on the narrative events instead of irrelevant scene changes, consistent with prior findings that structured context improves LLM reasoning (Wu et al. 2023; Fan et al. 2024).
- **Lexical fallback retrieval works**, but occasionally misses synonyms (*cup* vs *mug*), suggesting embedding-based retrieval with CLIP-like models (Radford et al. 2021; Gao et al. 2021) may help.
- **TSG answers duration questions accurately**, since timestamps were preserved through START/END event structure.
- **Complex multi-hop reasoning works**: e.g., "What did the person do after putting down the bowl?" — TSG context keeps both events proximal, enabling multi-step inference.

4.5 Visualization

As shown from figure 3, short context is cheap but inaccurate; full logs are accurate but expensive; TSG lies closest to the optimal region, achieving both accuracy and efficiency.

5 Discussion

Our findings indicate that symbolic temporal abstraction offers an effective intermediate representation between raw long-horizon video and LLM inference. By converting pixel streams into discrete interaction events, and subsequently operating on the Temporal Scene Graph (TSG), the system constrains context to a semantically meaningful subset of the original video trace. This allows downstream VLMs to reason over long activities without incurring quadratic token overhead or being flooded by visually redundant information. In contrast, the Full Log baseline frequently saturates the model's context window with low-relevance frames, forcing the LLM to perform its own saliency filtering—a failure mode also noted in memory-augmented architectures (Wu et al. 2023; Fan et al. 2024).

5.1 TSG Reasoning Improvement

- **Lower reasoning burden.** Direct event abstraction obviates the need for Gemini to infer importance across hundreds of near-duplicate frames. Instead, the model is presented with only high-utility temporal facts.
- **Temporal relationships remain intact.** The START→END formulation yields explicit segment boundaries, encoding order and duration without additional modeling. This mirrors temporal-relational approaches in video grounding (Yuan et al. 2019; Tang et al. 2020).
- **Noise suppression improves attention allocation.** Eliminating off-task interactions reduces attention diffusion, enabling more stable reasoning compared to full-context prompting—aligning with structured token pruning observations in video compression literature (Deng et al. 2024; Li et al. 2023b).

Together, these results suggest that VLM performance is constrained not only by token count but by *representation granularity*. A TSG serves as an information bottleneck that filters surface detail while preserving causal structure, enabling the LLM to allocate its attention budget on reasoning rather than perception. As such, the quality of long-horizon video reasoning may depend less on scaling model size and more on providing the right symbolic substrate.

5.2 Failure Modes and Challenges

Despite strong performance, several consistent weaknesses emerged:

- **Lexical brittleness**: anchor detection is string-matching based. Queries containing synonyms ("*mug*" vs "*cup*"), plurals, pronouns, or paraphrases may retrieve incomplete events.
- **Off-camera actions**: if the person temporarily leaves view, no event is recorded, leading to ambiguous or partial answers, similar to limitations reported in egocentric benchmarks (Bai et al. 2022).
- **Complex multi-object reasoning**: questions requiring co-reference across more than two interacting entities are harder for the model to reconstruct.
- **No visual grounding**: we rely purely on symbolic events; cases requiring appearance recognition ("red cup" vs "blue cup") cannot be resolved without reintroducing image features or visual embeddings.

These limitations provide several directions for improvement.

5.3 Limitations and Evaluation Scope

Our evaluation is intentionally small-scale and should be interpreted as a proof-of-concept rather than a full benchmark study. First, the dataset comprises only five long-form YouTube videos and 120 question–answer pairs. While this is sufficient to expose clear trends between short-context, full-log, and TSG-based prompting, it does not cover the diversity of real-world video distributions. Second, both question generation and answer evaluation rely on large language models: questions are produced with an LLM, and accuracy is computed using Gemini as an automatic judge. Although we follow recent practice in multimodal QA (Zhao et al. 2023; Xu et al. 2023), this introduces potential biases—especially since the answering model (Gemini 2.5 Flash) and the judge belong to the same model family. A full human evaluation and cross-model judging (e.g., evaluating Gemini answers with a different LLM) are left to future work. Finally, we focus exclusively on symbol-level inputs; tasks that require fine-grained visual appearance (color, identity, pose) are out of scope for the current system and would require a hybrid symbolic–visual design.

5.4 Implications

TSG demonstrates that **semantic compression is a viable alternative to dense visual prompting**. Instead of scaling LLM context windows indefinitely, long-form understanding may be achieved through multi-stage processing:

$$\text{Video} \Rightarrow \text{Symbols} \Rightarrow \text{Graphs} \Rightarrow \text{Language}$$

This suggests three promising research directions:

1. **Embedding-aware retrieval** to resolve synonym mismatches and improve anchor recall, using CLIP or sentence-transformer embeddings (Radford et al. 2021; Gao et al. 2021).
2. **Multi-hop reasoning over graph neighborhoods** to recover more complex chains of events, leveraging advances in graph-LLM integration (Huang et al. 2022; Chen et al. 2024).
3. **Hybrid symbolic + visual querying**, retrieving raw frames only when needed, aligned with emerging hybrid video reasoning models (Ma et al. 2022; Yang et al. 2024).

6 Future Work

While TSG-based pruning proves highly effective for long-form reasoning, several extensions could further enhance performance, robustness, and generalization. We outline key directions for future exploration below:

- **Embedding-based retrieval replacing lexical anchors.** Current query grounding relies on exact token overlap, which is brittle to synonym or paraphrase variation. Incorporating sentence-transformer or CLIP embeddings (Radford et al. 2021; Gao et al. 2021) could enable fuzzy matching between query terms and graph nodes (e.g., "cup" $\leftrightarrow$ "mug"), improving recall in ambiguous cases.
- **Multi-hop inference and causal reasoning chains.** Some queries reference multi-event relationships (*after placing the bowl, what did they do next?*) that require reasoning across sequential nodes. Graph search or causal message passing over event timelines could enable multi-step inference beyond direct anchor neighborhoods, drawing on recent graph-LLM methods (Huang et al. 2022; Chen et al. 2024).
- **Real-time streaming Event Graph QA.** Our pipeline currently processes video offline. Extending to streaming input with incremental TSG updates would unlock applications in at-home monitoring, assistive robotics, or live surveillance, where low-latency answers are needed.
- **Hybrid symbolic + visual retrieval.** Certain tasks require visual grounding (color, identity, gesture). A hybrid pipeline could retrieve raw frames associated with selected events only when necessary, preserving efficiency while adding semantic resolution, in line with hybrid video-LLM designs (Ma et al. 2022; Yang et al. 2024).
- **Benchmark comparison with frame-sampling VLM baselines.** Future work should include direct comparison against uniform/semantic frame sampling, token compression models, and visual-memory LLMs (Deng et al. 2024; Li et al. 2023b; Wu et al. 2023) to quantify tradeoffs across cost vs. accuracy.
- **Scalability to hour-long and multi-camera videos.** Our dataset demonstrates feasibility, but scaling to tens of hours or multi-view environments could reveal new challenges in identity maintenance, graph growth, and context window management.
- **Graph reasoning using LLMs as differentiable planners.** Converting TSG sequences to structured natural language allows LLMs to perform temporal deduction. A future system could allow the model to iteratively query the graph, forming a closed-loop reasoning controller over symbolic events.

Acknowledgments

References

Ahuja, N.; et al. 2021. Noisy-Labeled Video Scene Graph Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Bai, J.; et al. 2022. EgoVLP: Egocentric Video-Language Pretraining. In *Advances in Neural Information Processing Systems*.

Baradel, F.; Wolf, C.; Mille, J.; and Taylor, G. W. 2018. Object Level Visual Reasoning in Videos. In *Proceedings of the European Conference on Computer Vision*.

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Dohan, D.; Damoc, B.; Hernández, D. P.; et al. 2022. Improving Language Models by Retrieving from Trillions of Tokens. *Proceedings of the International Conference on Machine Learning*.

Chen, Y.; et al. 2024. GraphMamba: Towards Long-Range Graph Sequence Modeling with Selective State Spaces. *arXiv preprint arXiv:2402.00789*.

Deng, Y.; et al. 2024. Video Token Pruning for Efficient Video-Language Models. *arXiv preprint*.

Fan, A.; et al. 2024. LongMem: Long-Context Memory for Large Language Models. *arXiv preprint*.
Gao, Y.; et al. 2021. CLIP-Score: A Reference-Free Evaluation Metric for Image Captioning. In *Proceedings of the 59th Annual Meeting of the ACL*.
Huang, W.; et al. 2022. Graph-LLM: Can Large Language Models Solve Graph Problems? *arXiv preprint*.
Jocher, G.; and Ultralytics. 2023. Ultralytics YOLOv8 and YOLOv11. https://docs.ultralytics.com/.
Kaplan, J.; et al. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*.
Kim, M.; et al. 2023. Video-X: A Large-Scale Video Instruction Benchmark for Video LLMs. *arXiv preprint*.
Lei, J.; et al. 2023. LvBench: A Benchmark for Long-form Video Understanding. *arXiv preprint arXiv:2312.04817*.
Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*.
Li, D.; et al. 2023a. Video-LLaMA: An Instruction-Tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858*.
Li, D. B.; et al. 2023b. Token Merging: Your ViT but Faster. *International Conference on Learning Representations*.
Ma, X.; et al. 2022. OmniVL: One Vision-Language Model for All Tasks. *arXiv preprint arXiv:2209.07526*.
Ma, Z.; et al. 2023. LLaMA-VID: An Image-to-Video Language Model for Long Video Understanding. *arXiv preprint arXiv:2311.17043*.
OpenAI. 2024a. GPT-4o System Card. https://openai.com/index/gpt-4o-system-card/.
OpenAI. 2024b. OpenAI o1 Reasoning Models System Card. https://cdn.openai.com/o1-system-card.pdf.
Qi, Y.; et al. 2023. EgoSchema: A Large-Scale Egocentric Video-Language Benchmark. *Advances in Neural Information Processing Systems*.
Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models from Natural Language Supervision. *Proceedings of the International Conference on Machine Learning*.
Reid, M.; and Team, G. 2024. Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context. *arXiv preprint arXiv:2403.05530*.
Tang, K.; et al. 2020. STTran: Spatial-Temporal Transformer for Dynamic Scene Graph Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
Wu, Z.; et al. 2023. MemoryBank: Enhancing Large Language Models with Long-Term Memory. *arXiv preprint arXiv:2305.10250*.
Xu, H.; et al. 2023. LLaVA-NeXT: Large Language and Vision Assistant for Video Understanding. *arXiv preprint arXiv:2311.00571*.
Yang, Z.; et al. 2024. V2VFormer: A Hybrid Visual-Token to Video-Token Transformer for Long Video Understanding. *arXiv preprint*.
Yuan, L.; et al. 2019. Temporal Relational Reasoning in Videos. In *Proceedings of the European Conference on Computer Vision*.
Zhang, Y.; Yang, A.; et al. 2023. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
Zhao, W. X.; et al. 2023. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355*.