OASIS: On-Demand Hierarchical Event Memory for Streaming Video Reasoning

Supplementary Material

1. Benchmarks

1.1. OVO-Bench

OVO-Bench [3] is designed to measure the ability of large video understanding models to comprehend online videos under realistic, temporally constrained conditions. The central protocol is simple: at any arbitrary playback time the model receives a question and must answer using only information available up to that time. OVO-Bench contains 644 independent videos spanning seven major domains (e.g., sports, video games, tutorials). Tasks are organized into three complementary categories, including Backward Tracing, Real-Time Visual Perception, and Forward Active Responding. Backward Tracing probes memory and retrospective reasoning, including Episodic Memory(EPM), Action Sequence Identification(ASI), Hallucination Detection(HLD). Real-Time Visual Perception evaluates the model's perceptual acuity at the current time and in the immediate past, which focus on Spatial Understanding(STU), Object Recognition(OJR), Attribute Recognition(ATR), Action Recognition(ACR), Optical Character Recognition(OCR), and Future Prediction(FPD). Forward Active Responding assesses a model's ability to defer response until sufficient evidence accumulates, such as Repetition Event Count(REC), Sequential Steps Recognition(SSR), Clues Reveal Responding(CRR). In this work, we concentrate our experiments on the Backward Tracing and Real-Time Visual Perception categories to study the tradeoffs between episodic recall and instantaneous visual understanding.

1.2. StreamingBench

StreamingBench [2] evaluates multimodal large language models (MLLMs) in realistic streaming video, multi-turn interaction settings. The dataset comprises 900 videos and 4,500 questions across eight video types. To emulate continuous interaction, we sample five timestamps per video and pose questions at each. The benchmark is divided into three evaluation axes. Real-Time Visual Understanding measures immediate visual comprehension such as Object Perception(OP), Causal Reasoning(CR), Clips Summarization(CS), Attribute Perception(ATP), Event Understanding(EU), Text-Rich Understanding(TR), Prospective Reasoning(PR), Spatial Understanding(SU), Action Perception(ACP), Counting(CT). Omni-Source Understanding requires synchronous integration of video and audio sources including Emotion Recognition(ER), Scene Understanding(SCU), Source Discrimination(SD), Multimodal Alignment(MA). Contextual Understanding targets robust reasoning under complex temporal contexts, including Misleading Context Understanding (MCU) where earlier frames may introduce distractors, Anomaly Context Understanding (ACU) which requires detecting subtle anomalies, Sequential Question Answering (SQA) where later questions depend tightly on earlier ones, and Proactive Output (PO) which tests when a model should choose to produce an answer. Our experiments emphasize passive Real-Time Visual Understanding and the ACU/MCU/SQA contextual subsets to evaluate model stability in ambiguous or confounding streams.

Table 1. Results on HourVideo. We report overall accuracy(%). HourVideo contains egocentric videos spanning 20–120 minutes, providing a direct test of hour-scale temporal reasoning.

Model	Accuracy
Qwen3-VL-8B‡	35.11
+ OASIS	**37.35**

1.3. StreamBench

StreamBench [4] focuses on complex, multi-turn interaction in online videos. The collection contains 306 videos and over 1.8k questions, with multiple timestamped queries per video to simulate iterative search and conversational patterns. Questions are open-ended and, for baseline evaluation, we report results using LLaMA3-8B. Tasks fall into six categories: Object Search (OS), Long-term Memory Search (LM), Short-term Memory Search (SM), Conversational Interaction (CI), Knowledge-based QA (KG), and Simple Factual (SF). StreamBench emphasizes the model's ability to maintain retrieval accuracy and dialogue coherence in complex multi-turn interactions with high degrees of freedom.

2. Long-Horizon Evaluation

To further validate OASIS beyond the short-to-medium duration benchmarks in the main paper, we additionally evaluate it in two more challenging long-horizon settings. First, we test on HourVideo [1], whose egocentric videos span 20–120 minutes and therefore directly stress temporal reasoning at hour scale. Second, we construct a long-video slice from OVO-Bench by selecting only videos longer than 15 minutes. These two analyses complement the main paper by isolating the regime where long-term memory maintenance and retrieval quality matter most.

As shown in Table 1, OASIS improves Qwen3-VL-8B from 35.11% to 37.35%, indicating that the proposed hierarchical event memory remains beneficial even when the temporal horizon extends to hour-long videos.

Table 2. Results on long clips from OVO-Bench. We select only videos longer than 15 minutes and report accuracy(%) on the resulting subsets.

Subset	#Examples	Qwen3-VL-8B	+ OASIS
Long-clip perception	37	56.76	**70.27**
Long-clip backward	10	30.00	**50.00**

Table 2 further isolates the long-duration regime within a benchmark that otherwise mixes videos of diverse lengths. On videos longer than 15 minutes, OASIS improves the forward/perception subset from 56.76% to 70.27% and the backward subset from 30.00% to 50.00%. These gains are larger than the average gains reported in the main paper, suggesting that OASIS becomes increasingly advantageous as the amount of historical context grows and naive context accumulation becomes more brittle.

3. Detailed Results on StreamBench

StreamBench [4] provides a complementary evaluation setting to OVO-Bench and StreamingBench: its questions are open-ended, the task space is more diverse, and the multi-turn interaction pattern is less constrained. This makes it particularly useful for assessing whether OASIS generalizes beyond tightly structured benchmark formats. Following the official protocol, we report semantic-similarity-based evaluation with LLaMA-3-8B as the judge model.

As shown in Table 3, OASIS achieves strong gains on StreamBench across diverse open-ended tasks. In particular, the improvements on Long-term Memory Search (LM) and Short-term Memory Search (SM) show that the coarse-to-fine policy benefits both recent grounding and historical retrieval under freer-form interaction. The gain on Conversational Interaction (CI) further suggests that the QA summary and retrieval mechanism help maintain consistency across multi-turn dialogues. Overall, these results complement the main-paper benchmarks by showing that OASIS remains effective when answer spaces are less constrained and questions require more flexible retrieval and reasoning.

4. Details of Retrieval Algorithm in Fine Reasoning

To avoid redundancy where a parent node and its children are retrieved simultaneously, we employ a Greedy Pruning Strategy. The algorithm 1 illustrates this process: we calculate similarity scores for all nodes in the Event Forest against the retrieval query I_i. We iteratively select the highest-scoring node and remove its ancestors and descendants from the candidate pool. This forces the retriever to seek evidence from distinct event branches, ensuring that the final set $\mathcal{E}^\star$ maximizes information diversity while selecting the optimal granularity for each event.

5. Analysis of hyper-parameters

We performed sensitivity analyses on three key hyperparameters λ, k_f, and k_q in the proposed OASIS method to evaluate the robustness of model performance to parameter selection.

5.1. Sensitivity Analysis of λ

λ is a hyperparameter that controls the trade-off between the similarity of the event nodes and the hierarchical level of the nodes in the event forest. As shown in Table 4, we compare the performance of our method on OVO-Bench under different λ values. The performance demonstrates strong robustness to the selection of λ; across the tested range, the overall difference in Perception Avg is less than 1%.

5.2. Sensitivity Analysis of k_f

k_f is a hyperparameter that controls the number of event nodes retrieved from the event forest. Table 5 shows the performance variation with respect to k_f on OVO-Bench. Increasing k_f from 1 to 2 yields a significant performance gain, indicating that retrieving more nodes is crucial. However, further increasing k_f to 3 only results in marginal improvement on the Perception subset while slightly decreasing performance on the Backward subset. Thus, we select $k_f = 2$ to maintain an optimal balance between performance and computational efficiency.

5.3. Sensitivity Analysis of k_q

k_q is a hyperparameter that controls the number of previous questions retrieved from the question memory. As shown in Table 6, we evaluate the performance of our method on the Sequential Question Answering (SQA) subset of StreamingBench. The results clearly indicate that the SQA accuracy scales positively with the value of k_q within the tested range, achieving the best result of 49.20 at $k_q = 3$.

6. Computational Efficiency Analysis

We provide a comprehensive analysis of the computational efficiency of OASIS, focusing on two critical metrics: Peak GPU Memory Footprint and Request Processing Delay [4].

6.1. Peak GPU Memory Footprint

The high memory consumption of long-context MLLMs is the primary bottleneck preventing their deployment on edge devices or standard servers. We measured the peak GPU memory usage during inference on OVO-Bench.

As shown in Table 7, the baseline method consumes a massive 76.59 GB of memory, pushing the limit of a single A800 (80GB) GPU. This excessive footprint is largely due

Table 3. Results on StreamBench. We report accuracy(%) on 6 subsets. The best scores are highlighted in **bold**. ‡ denotes results reproduced by us, while others are taken from prior works. 'Avg' denotes the average performance.

Model	frames	OS	LM	SM	CI	KG	SF	Avg
GPT-4o	50	60.5	61.2	64.4	72.3	93.9	74.7	71.0
LLaMA-VID	180	33.9	38.2	44.1	58.4	76.9	57.1	51.2
LLaVA-Hound	8	37.6	43.2	53.4	55.7	76.3	62.0	54.7
LongVA	8	41.1	47.4	57.6	59.8	80.7	66.1	52.4
MiniCMP-v2.6	8	37.6	51.9	43.7	65.7	66.2	64.2	56.6
VILA-1.5	8	36.1	44.4	50.8	68.3	78.6	65.5	57.1
InternVL-V2	8	38.5	46.6	50.9	67.6	81.0	62.2	57.6
InternLM-XCP2.5	8	38.8	43.3	50.8	65.6	88.4	60.5	57.7
MovieChat	32	18.6	20.4	26.5	42.3	67.2	35.8	35.3
FreeVA	4	35.6	37.5	43.7	58.8	84.0	53.7	56.3
Video-online	5 fps	41.4	48.8	52.9	62.7	69.2	64.1	56.4
Flash-VStream	1 fps	37.1	44.5	48.6	58.1	66.4	59.2	52.1
StreamChat‡	15 fps	40.7	43.6	47.1	63.0	**89.1**	**73.8**	59.5
Qwen2.5-VL-7B‡	0.5 fps	36.1	41.3	41.0	50.5	76.5	62.3	51.1
+ OASIS	-	35.9	39.4	44.1	54.5	76.9	64.9	52.4
Qwen3-VL-8B‡	0.5 fps	33.8	50.5	47.8	69.4	73.5	62.6	56.0
+ OASIS	-	**47.0**	**53.6**	**54.3**	**71.7**	86.1	67.6	**62.1**

Algorithm 1 Greedy pruning strategy retrieval for Event Forest

Require: Event forest $\mathcal{F}$ with node set $\{\mathbf{R}_j\}$, retrieval query I_i, embedding encoder $\mathbf{E}$, number of retrieved nodes k_f
Ensure: Selected event nodes $\mathcal{E}^\star$ (up to k_f, hierarchically deduplicated)

$\mathcal{C} \leftarrow \{\mathbf{R}_j \mid \mathbf{R}_j \in \mathcal{F}\}$ ▷ initialize candidate set with all nodes
$s_j \leftarrow \cos\big(\mathbf{E}(I_i),\ \mathbf{e}_j\big)$ for each $\mathbf{R}_j \in \mathcal{C}$ ▷ compute similarity scores
sort $\mathcal{C}$ by s_j in descending order
$\mathcal{E}^\star \leftarrow \varnothing$
while $|\mathcal{E}^\star| < k_f$ **and** $\mathcal{C} \neq \varnothing$ **do**
    $\mathbf{R}^\star \leftarrow$ pop highest-scoring node from $\mathcal{C}$
    $\mathcal{E}^\star \leftarrow \mathcal{E}^\star \cup \{\mathbf{R}^\star\}$
    $\mathcal{C} \leftarrow \mathcal{C} \setminus \big(\mathrm{Ancestors}(\mathbf{R}^\star) \cup \mathrm{Descendants}(\mathbf{R}^\star)\big)$ ▷ prune lineage nodes
end while
return $\mathcal{E}^\star$

to the linearly growing KV-cache required to store the entire video history.

In contrast, the full OASIS framework consumes only 28.48 GB, a significant reduction in memory requirements. Even with all memory components enabled, OASIS fits comfortably within consumer-grade hardware limits or allows for larger batch sizes on enterprise GPUs. The ablation rows further demonstrate that our memory components are lightweight, with the Event Forest adding minimal overhead compared to the base model weights.

Table 4. The ablation of λ on OVO-Bench. Perception Avg and Backward Avg denote the average performance on the Perception and Backward subset, respectively.

λ	Perception Avg	Backward Avg
0.0	78.26	57.06
0.1	78.14	57.21
0.2	78.74	57.37
0.5	78.62	57.84
1.0	78.65	57.82

Table 5. The ablation of k_f on OVO-Bench. Perception Avg and Backward Avg denote the average performance on the Perception and Backward subset, respectively.

k_f	Perception Avg	Backward Avg
1	77.12	56.74
2	78.14	57.21
3	78.39	56.90

Table 6. The ablation of k_q on Sequential Question Answering of StreamingBench.

k_q	SQA Acc
1	48.40
2	48.80
3	49.20

Table 7. **Component-wise breakdown of Peak GPU Memory usage on OVO-Bench.** We compare the full-context baseline against OASIS with various memory components enabled. Event Forest, Medium Buffer, and Short Window denote the three tiers of our hierarchical memory. The baseline nearly exhausts the A800 memory, while OASIS significantly reduces the footprint.

Method	Medium Buffer	Event Forest	Short Window	GPU Memory(GB)
OASIS	✓	✗	✗	20.91
OASIS	✗	✗	✓	20.09
OASIS	✗	✓	✓	24.40
OASIS	✓	✗	✓	22.28
OASIS	✓	✓	✗	26.15
OASIS	✓	✓	✓	28.48
Qwen3-VL-8B	-	-	-	76.59

7. Qualitative Analysis

We provide concrete visualization examples to demonstrate how OASIS adaptively handles different types of queries in streaming scenarios. We select two representative cases: one requiring Real-time Perception and one requiring Long-term Retrieval.

Figure 1 visualizes OASIS handling a real-time perception query where full-context modeling fails due to attention collapse. Despite the explicit temporal cue "now," the baseline misattends to historically-presented cards, yielding an incorrect answer. This failure case underscores the efficacy of our Coarse Reasoning stage: OASIS grounds its reasoning in the Short Window, accurately perceiving the five cards currently on the table. Crucially, the model autonomously assesses this local context as sufficient, thereby avoiding noisy historical information and producing a precise, hallucination-free response.

Figure 2 compares the full OASIS framework against its Coarse-reasoning-only variant (left) to demonstrate the necessity of the fine reasoning stage. Without the retrieval capability, the model relies solely on the ShortWindow and summaries, failing due to Evidence Missing and incorrectly concluding the white plate never appeared. In contrast, the full OASIS framework (right) successfully bridges this gap. Upon identifying the information deficiency in the current context, it actively plans a semantic hypothesis, inferring the object likely appeared in a kitchen or dining area, and executes a precise retrieval. This action retrieves the specific dining table event from the long-term history, enabling the model to accurately ground the answer.

8. Details of Prompts

We provide the exact prompts used in our implementation. Figure 3 illustrates the system instructions for the main inference model, specifically detailing the guidelines for the Two-Stage Reasoning policy. Regarding memory maintenance, Figure 4 presents the prompt responsible for generating event nodes from the medium buffer, while Figure 5 depicts the instruction for dynamically merging two adjacent event nodes into a unified summary. Figure 6 outlines the prompt used for iteratively updating the QA summary.

9. Future Work

OASIS currently operates in a training-free regime, leveraging MLLM zero-shot reasoning. While effective for standard scenarios, this paradigm exhibits limitations on tasks requiring multi-hop retrieval or fine-grained visual grounding. We will address this by curating instruction-tuning datasets for temporal planning, enabling the model to learn optimal search strategies rather than relying exclusively on prompting. This supervised refinement will enhance the robustness and precision of agentic retrieval in long-form video streams.

References

[1] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024. 1

[2] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo,

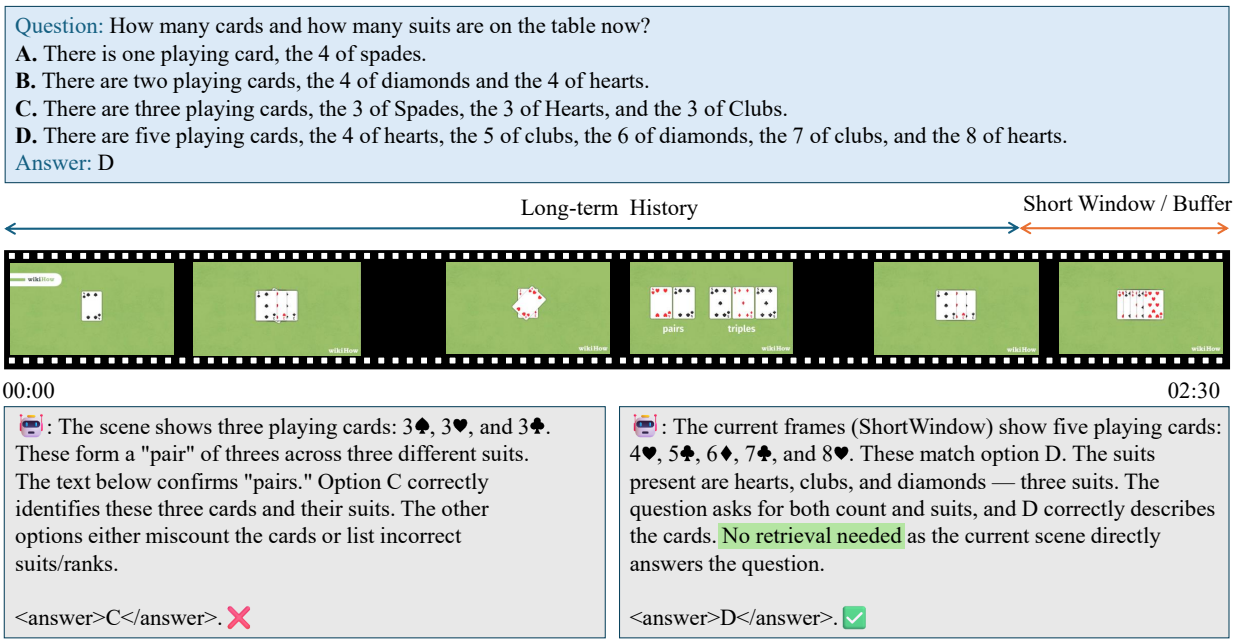

Figure 1. Visualization of OASIS handling a Real-time Perception query.

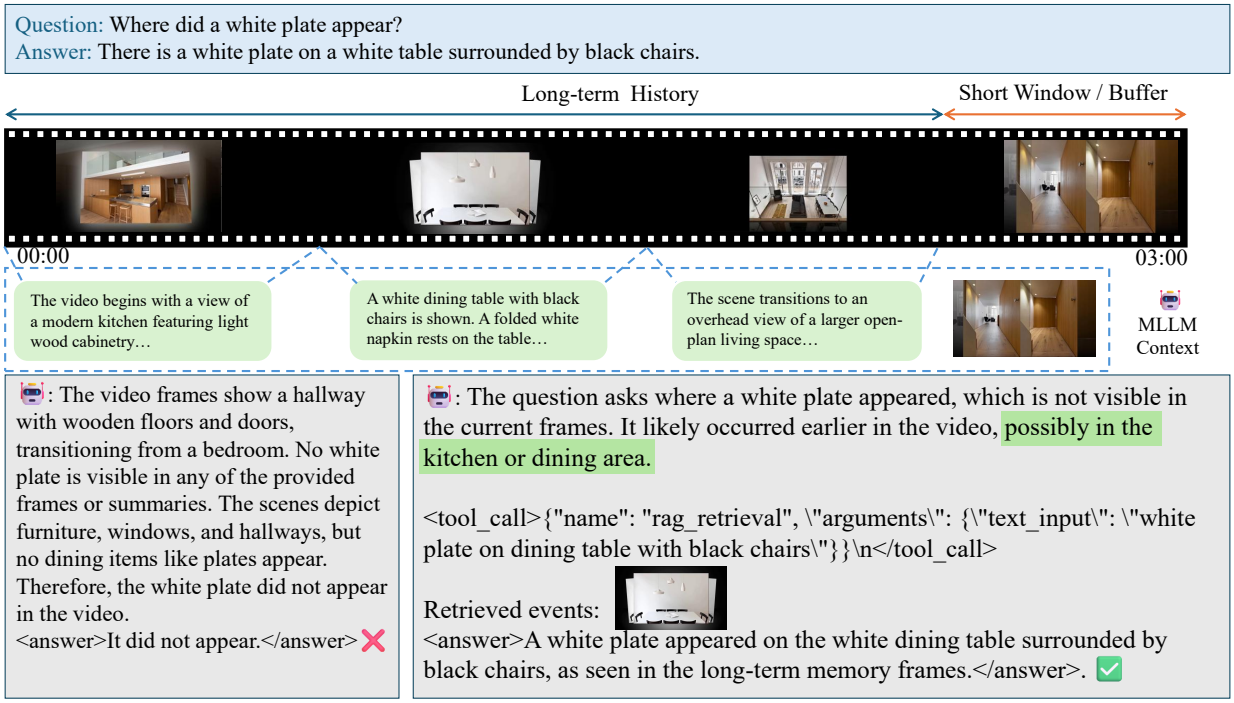

Figure 2. Visualization of OASIS handling a Long-term Memory query.

Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024. 1

[3] Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*,

You are an expert multimodal assistant on virtual reality headset for streaming video QA.
You are looking at a video of a real-world scene, and you are answering real-time questions for user

Inputs (separate fields)
- now_window_frames: fine-grained frames from NowWindow, very recent video frames that the headset is looking at.
- short_term_frames: fine-grained frames from ShortTermWindow.
- long_term_events: textual summaries (and/or coarse frames) of longer segments.
- qa_history_summary: summary of prior Q/A.

Decision Policy
- Read the user question + provided frames/summaries, briefly think step-by-step about the question
- After thinking, you can call `rag_retrieval` with a precise interest description, the tool will return the specific video clips and question-and-answer history that are most relevant to your description.
- Once you confirm your final answer, place the final answer inside <answer> and </answer>.

Forming the Retrieval Query
- Be specific and brief (≤ 20 words), prefer noun phrases.
- Include disambiguators when available: who/what, action, location/region, salient attributes (color/count).
- Avoid vague queries ("more info", "look again").

Tool
You may fetch missing details by issuing a concise interest description (entity, action, time anchor, location, attributes).

```
<tools>
{
  "type": "function",
  "function": {
    "name_for_human": "rag_retrieval",
    "name": "rag_retrieval",
    "description": "Retrieve details based on a concise interest description.",
    "parameters": {
      "type": "object",
      "properties": {
        "text_input": {
          "type": "string",
          "description": "Short, specific description to retrieve."
        }
      },
      "required": ["text_input"]
    }
  }
}
</tools>
```

Tool Call Format (example)

```
<tool_call>
{"name": "rag_retrieval", "arguments": {"text_input": "man opens car trunk, parking lot"}}
</tool_call>
```

Figure 3. System Prompt Template used in OASIS.

pages 18902–18913, 2025. 1

[4] Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced

You are an event summarizer for a Short-Term Memory (STM) video window.

Goal
- Produce ONE self-contained summary describing what happens inside this STM window only.

Inputs
- STM frames (authoritative evidence).

Hard Rules
1) Chronology: Narrate in temporal order within the STM window. No reordering across time.
2) No guessing: Do not infer intentions/causes not shown. If something is unclear, state "unidentified/unclear" rather than guessing.

Content Focus
- Who did what to whom/what, where, with what tool/object, and the immediate result.
- Include objects visible in STM

Style
- Active voice; present or simple past; concrete, observable facts.
- No titles, lists, timestamps, metadata, or markup.
- Length ≤ 300 words.

Output
- Output ONLY the summary text.

Figure 4. Event Summary Prompt Template used in OASIS.

You are a summary merger.
I'm giving you two summaries and their timestamps, each describing the content of two adjacent clips from a video.
You need to merge them into one.

Rules:
1. Do not add any new information that is not already present in either summary.
2. Maintain chronological order and keep the total word count under 300.

Here are the two summaries:
{summary_a}
{summary_b}

Now output your final summary directly, without any additional explanation or title, and you do not need to output the timestamps at the beginning:

Figure 5. Event Merge Prompt Template used in OASIS.

knowledge. *arXiv preprint arXiv:2501.13468*, 2025. 1, 2

You are a QA aggregator. You receive the current QA history summary S and a new QA. Your task is to generate an updated S' for subsequent retrieval and low-cost reasoning.

Hard Rules:
1) Only use information from S and the new QA; no external knowledge or assumptions should be introduced.
2) Preserve the "who/what/key changes"; resolve pronouns and unify entity names.
3) De-duplicate and merge duplicate or synonymous statements; and remove redundant and irrelevant content.
4) Keep the total length to under 300 words.
5) Output only the updated summary text, without any explanations, titles, or additional notes.

Given the QA history summary S:
{QA_summary_all}

Given the new QA (including questions and answers):
{QA}

Now output the updated summary:

Figure 6. QA Summary Prompt Template used in OASIS.