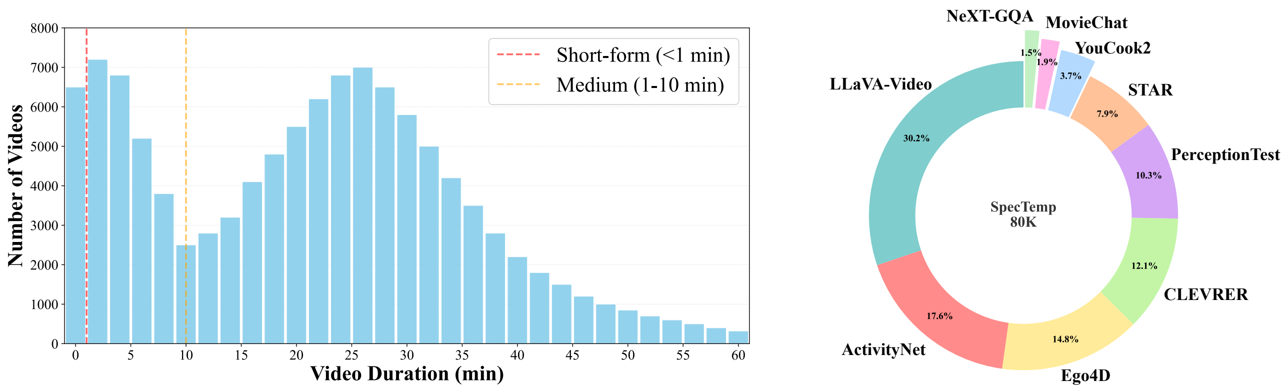
Thinking with Drafts: Speculative Temporal Reasoning for Efficient Long Video Understanding

Supplementary Material



(a) Video duration distribution.

(b) Data source composition.

Figure 1. SpecTemp-80K dataset statistics.

This supplementary material is organized as follows. We begin with a detailed description of the experimental setups, then present additional experimental findings, followed by the visualization results and a discussion of limitations.

Specifically, the **experiment setups** include the following aspects:

- Training configurations.
- Prompt templates for dual models.
- SpecTemp-80K dataset statistics.

More experimental findings include the following aspects:

- Ablation on different iteration numbers.
- Ablation on reward components.
- Frame allocation strategies under fixed budget.
- Needle-in-a-Haystack evaluation.

1. Experiment Setup

Training configurations. We employ a two-stage training strategy. First, we conduct supervised fine-tuning (SFT) for 2 epochs on the SpecTemp-80K dataset with a learning rate of 1×10^{-6}. Subsequently, we perform reinforcement fine-tuning (RFT) using GRPO [2] for 1 epoch with a learning rate of 1×10^{-6}. The KL penalty coefficient β is set to 0.04. All experiments are conducted on 16×L20 GPUs with batch sizes of 32 for SFT and 16 for RFT.

Prompt templates for dual models. The prompt templates for the target model (temporal segment prediction and answer generation) and the draft model (salient frame selection from dense samples) are illustrated in Figure 4 and Figure 5, respectively.

SpecTemp-80K dataset statistics. Figure 1(a) presents the video duration distribution, showing three distinct categories: short-form videos (<1 min, 32.4%), medium-length videos (1-10 min, 51.8%), and long-form videos (>10 min, 15.8%). Figure 1(b) illustrates the data source composition of SpecTemp-80K. The dataset comprises 80,142 video-question-answer triplets collected from 9 diverse sources, with LLaVA-Video (30.2%), ActivityNet (17.6%), and Ego4D (14.8%) constituting the primary sources. This diverse composition ensures comprehensive coverage across varying temporal complexities and video domains, enabling robust training for dual-model collaborative reasoning.

2. More Experimental Findings

Table 1. Ablation study on maximum iteration number $T_{\max}$. We report accuracy (%) on LongVideoBench and Video-Holmes, along with actual iterations executed and inference latency.

$T_{\max}$	Video-Holmes [1]	LongVideoBench [3]	Iterations	Latency (s)
1	42.8	54.7	**0.9**	**1.5**
2	44.6	55.1	1.7	2.0
3 (Ours)	**47.0**	57.5	2.3	2.3
4	46.7	57.3	3.1	2.9
5	46.2	**58.1**	3.7	3.5

Ablation on different iteration numbers. We investigate the impact of maximum iteration number $T_{\max}$ on performance and efficiency. As shown in Table 1, accuracy im-

Table 2. Ablation study on reward components. We evaluate the contribution of IoU reward R_{IoU} for temporal localization in the target model and visual information gain reward R_{visual} for frame selection in the draft model.

R_{IoU} (Target)	R_{visual} (Draft)	Video-Holmes [1]	LongVideoBench [3]
-	-	42.1	53.1
✓	-	45.5	55.8
-	✓	44.7	54.6
✓	✓	**47.0**	**57.5**

Table 3. Ablation study on frame allocation strategies (Initial+Per-Iter×Max-Iter) under fixed budgets of 16 and 64 frames. We report accuracy (%) and inference latency.

Strategy	Video-Holmes [1]	LongVideoBench [3]	Latency (s)
16-frames budget			
4+4×3	44.3	55.8	**1.5**
10+2×3 (Ours)	**47.0**	**57.5**	1.8
13+1×3	45.6	56.2	2.0
64-frames budget			
32+8×4 (Ours)	**47.8**	**61.4**	**4.7**
48+4×4	46.1	59.7	5.2
56+2×4	46.9	60.2	5.5

proves steadily as $T_{\max}$ increases from 1 to 3, achieving 57.5% on LongVideoBench (+2.8%) and 47.0% on Video-Holmes (+4.2%). However, beyond $T_{\max} = 3$, we observe diminishing returns and even performance degradation on Video-Holmes (46.2% at $T_{\max} = 5$), suggesting that excessive iterations may introduce redundant frames and impair reasoning coherence. The actual average iteration numbers (0.9 to 3.7) are consistently lower than $T_{\max}$, demonstrating effective early termination. Considering that $T_{\max} = 3$ achieves the best Video-Holmes performance and competitive LongVideoBench accuracy with only 2.3s latency, we adopt this configuration as the optimal balance between performance and efficiency.

Ablation on reward components. We conduct ablation studies to analyze the contribution of individual reward components in our reinforcement learning framework. As shown in Table 2, both the IoU reward R_{IoU} for the target model and the visual information gain reward R_{visual} for the draft model provide substantial improvements over the SFT baseline (53.1% on LongVideoBench, 42.1% on Video-Holmes). Individually, R_{IoU} contributes larger gains (+2.7% on LongVideoBench, +3.4% on Video-Holmes) compared to R_{visual} (+1.5% on LongVideoBench, +2.6% on Video-Holmes), indicating that accurate temporal localization is the primary bottleneck for long video understanding. Importantly, combining both rewards yields the best performance (57.5% on LongVideoBench, 47.0% on Video-Holmes), with improvements of +4.4% and +4.9% over baseline respectively, demonstrating that the two reward signals provide complementary supervision for effective dual-model collaboration.

Frame allocation strategies under fixed budget. To investigate the optimal frame allocation strategy under a fixed computational budget, we conduct ablation studies with different distributions of initial frames and per-iteration frames. As shown in Table 3, we evaluate three configurations for both 16-frame and 64-frame budgets, where the strategy is denoted as Initial+Per-Iter×Max-Iter.

The results demonstrate that our default configurations consistently achieve the best performance across both settings. For the 16-frame budget, the 10+2×3 strategy attains 57.5% on LongVideoBench and 47.0% on Video-Holmes, outperforming both alternatives with only 1.8s latency. For the 64-frame budget, our 32+8×4 configuration achieves 61.4% and 47.8% respectively while maintaining the lowest latency (4.7s) among all 64-frame strategies.

These findings reveal an important trade-off in frame allocation strategies. Aggressive early sampling (4+4×3, 48+4×4) prioritizes many frames per iteration but suffers from insufficient global temporal context during initial reasoning, leading to suboptimal segment prediction. Conversely, conservative late sampling (13+1×3, 56+2×4) allocates excessive frames to initial observation but lacks the granularity needed for fine-grained verification in subsequent iterations. Our balanced approach allocates moderate initial frames for coarse temporal understanding while reserving sufficient budget for iterative refinement, effectively bridging global perception and local examination. This validates that our frame allocation strategy achieves an optimal equilibrium between temporal reasoning breadth and verification depth.

Needle-in-a-Haystack evaluation. To evaluate SpecTemp's ability to retrieve visual information from extremely long videos, we conduct a Visual Needle-In-A-Haystack (NIAH) test inspired by [4]. We insert single-frame visual questions at various positions into hour-long video haystacks sampled at 1 FPS. The needle images are designed to be counterfactual or counter-commonsense to ensure the model cannot rely on language priors alone.

As shown in Figure 3, SpecTemp successfully retrieves needle information within videos containing up to 2000 frames, maintaining accuracy above 80% across different needle positions. The dual-model collaboration (Target 7B + Draft 3B) achieves superior retrieval accuracy compared to using the 7B model alone, demonstrating that our speculative temporal reasoning framework maintains precise long-range visual information retrieval while significantly improving computational efficiency.

3. Visualizations

We provide qualitative examples to illustrate SpecTemp's iterative speculation-verification process across different video understanding scenarios. Figures 6 – Figures 8

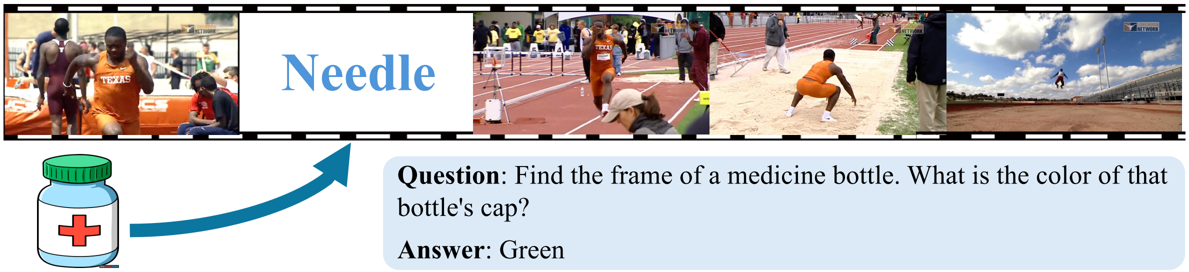


Figure 2. Illustration of the Visual Needle-In-A-Haystack evaluation. A single-frame visual question is inserted at varying positions within long video sequences to assess long-range retrieval capability.

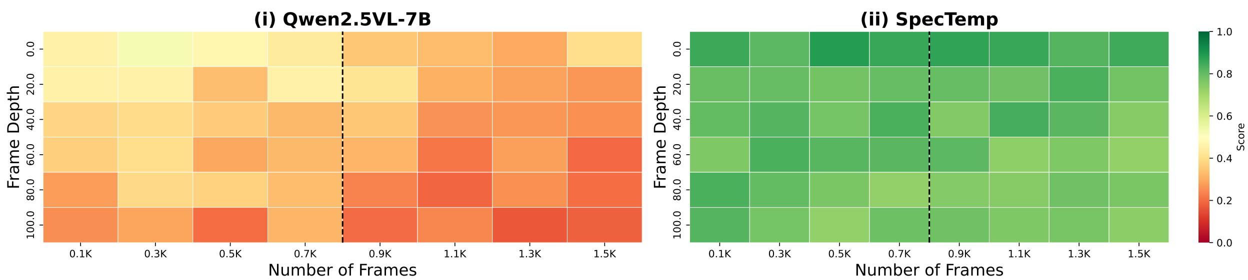


Figure 3. V-NIAH performance across frame counts and needle depths.

demonstrate how the target model progressively refines temporal segments while the draft model proposes salient frames, showcasing effective dual-model collaboration for long video reasoning.

4. Limitations

Despite the effectiveness of SpecTemp, we acknowledge several limitations that present opportunities for future research:

Computational overhead during training. Our dual-model framework requires training two separate MLLMs and performing joint reinforcement learning optimization, which increases computational costs compared to training a single model. The RL stage is particularly resource-intensive, requiring generation of multiple trajectory samples per training example for group reward computation. Future work could explore more efficient training strategies, such as parameter-efficient fine-tuning methods.

Generalization to extremely long videos. While SpecTemp handles hour-long videos effectively, its performance on videos exceeding 2-3 hours has not been extensively evaluated. For such extremely long videos, the number of potential temporal segments grows dramatically, and maintaining coherent reasoning across numerous iterations becomes challenging. Hierarchical reasoning strategies or memory mechanisms may be necessary for handling ultra-long videos.

References

[1] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. 1, 2

[2] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 1

[3] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024. 1, 2

[4] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 2

Question: {question}

Based on the {frames} sampled video frames from a video, analyze whether the provided frames contain sufficient information to answer this question.

Frame indexing: 1 to {frames}

If you need additional visual information from a specific time interval, provide your detailed reasoning of what information is missing between the <think> </think> tags, then specify the interval using <segment>(x,x+1)</segment> which indicates you need frames between frame x and frame x+1.

You will receive new video frames. You may request additional video frames up to a maximum of 3 times. Always choose an interval according to the most recent video frames.

IMPORTANT: Each round of reasoning must be based solely on the currently visible video frames. Do not rely on or repeat reasoning patterns from previous iterations.

When sufficient information is available, provide your reasoning of answering the question in <think> </think> tags and your final answer in <answer> </answer> tags.

Format: <think>your reasoning process here</think> <segment>(x,x+1)</segment> OR <think>your reasoning process here</think> <answer>your final answer here</answer>

Figure 4. **Prompt template for the target model.** The target model is guided to perform temporal reasoning and either predict evidence segments for further exploration or generate final answers.

You will be shown {num_frames} video frames. Your task is to select the two frames that are most relevant to answering the following question:

Question: {question}

Frame indexing: 1 to {num_frames}

Provide your detailed reasoning between the <think> </think> tags, and then give the two frame numbers using <frame>(x,y)</frame> tags, where x ≠ y.

Format: <think>your reasoning process here</think> <frame>(x,y)</frame>

Figure 5. **Prompt template for the draft model.** The draft model is instructed to select the most informative and diverse frames from densely sampled temporal regions.

CVPR
#***

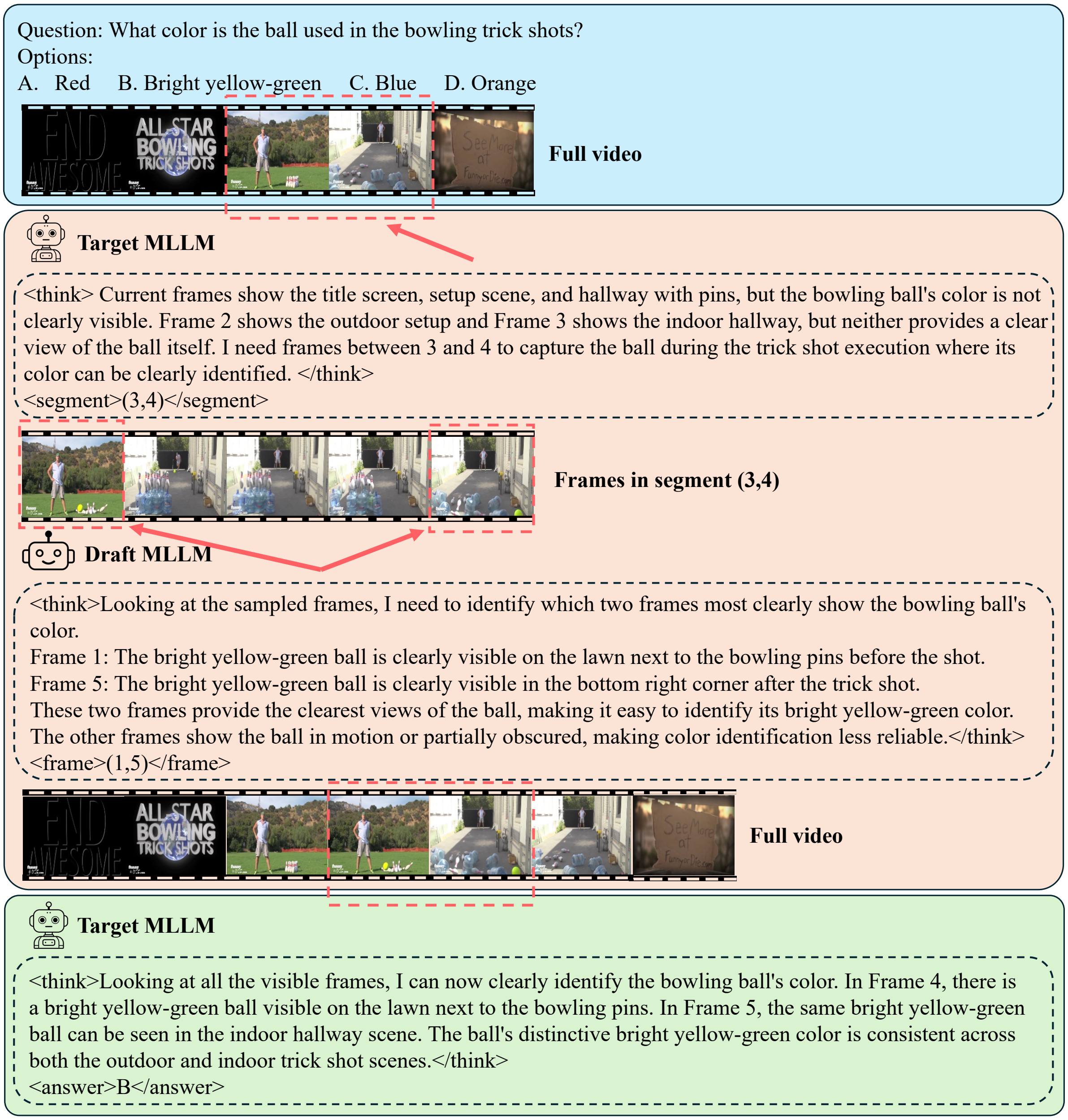


Figure 6. Qualitative case: 1 iteration for bowling ball color identification.

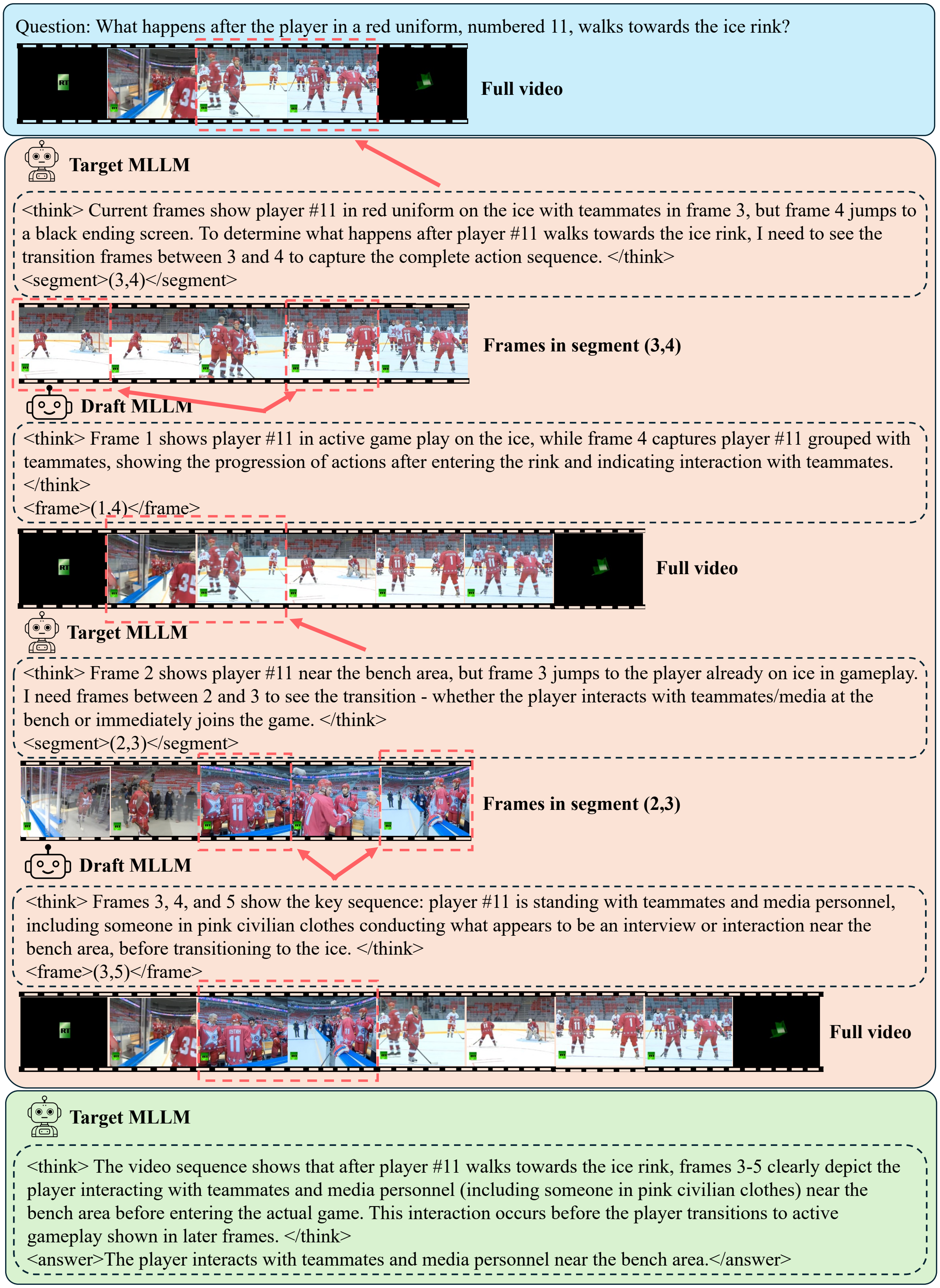


Figure 7. Qualitative case: 2 iterations for sports action sequence tracking.

Figure 8. Qualitative case: 3 iterations for temporal ordering reasoning.