TreeReasoner: Reinforcing Tool-Augmented Tree-of-Videos Reasoning

Supplementary Material

A. Case Study

In the introduction, we posited that our agentic framework enhances reasoning by actively navigating the video timeline—a departure from passive frame processing—and highlighted the interpretability of its search strategies. In this section, we provide qualitative examples (Fig. 1 and 2) to substantiate these claims, specifically demonstrating the model's search behaviors, such as temporal bracketing, causal chain verification, and adaptive temporal resolution adjustment.

Temporal bracketing. As shown in Fig. 1a, consider a query requiring multi-step reasoning, such as "Identify the color of the pants worn by the young woman talking to an elderly man on the subway." Initially, TreeReasoner employs a coarse-grained search with a large temporal stride to rapidly scan the entire timeline, identifying broad temporal regions where key objects (e.g., subway scenes, elderly man) appear. Upon detecting these relevant contexts, the model automatically switches to a fine-grained temporal mode, progressively narrowing down the time window by examining adjacent segments. This allows it to precisely bracket the specific temporal span where the target event—"a young woman conversing with an elderly man"—occurs. Within this bracketed time window, the model then analyzes the fine-grained visual details to extract the answer. By "bracketing" the event temporally in this coarse-to-fine manner, the model avoids processing the redundant background footage, directly contributing to the efficiency gains mentioned in our experiments. Another similar case can be found in Fig. 2a.

Causal chain verification. For complex queries such as "Find the color of the pants worn by the young girl who talks to the old man on the subway", the model exhibits a hierarchical reasoning strategy. TreeReasoner first searches for video segments containing the objects mentioned in the query, including segments with subway scenes, old men, and young girls. After identifying segments with the specified objects, TreeReasoner further examines the consistency between video content and the query-specified events, ultimately pinpointing the segment where "the young girl talks to the old man" occurs. It then comprehends the object information in this segment and provides the answer. This search strategy enables TreeReasoner to retrieve query-relevant video segments at a relatively fine-grained level, forming a complete chain of visual evidence and delivering accurate responses grounded in fine-grained visual details. This behavior demonstrates that TreeReasoner validates relationships by progressively refining its search through multiple verification steps, ensuring the answer is grounded in visual evidence rather than hallucination. Another similar case can be found in Fig. 2b.

Adaptive temporal resolution adjustment. In the final reasoning tree, TreeReasoner exhibits variable temporal resolutions across different video segments through adaptive tool invocation. As shown in Fig. 1c, for static, repetitive, or query-irrelevant portions of the video, TreeReasoner dynamically increases its temporal stride, skipping large chunks of uninformative frames. Conversely, for segments closely related to the query—such as those containing the target individuals and their interaction—the model immediately invokes tools to extract short video clips and applies much smaller step sizes to capture high-frequency details of the girl's appearance and clothing. This adaptive mechanism explains how our method achieves superior accuracy while processing substantially fewer frames than fixed-rate baselines. Another similar case can be found in Fig. 2c.

B. Dataset Details

To endow TreeReasoner with robust tool-use capabilities and ensure generalization across diverse visual domains, we constructed a unified training corpus by aggregating queries from multiple open-source datasets, including CLEVRER [5], LLaVA-Video-178K [6], NExT-QA [4], Perception-Test [3], Video-STaR [7], and LongVideo-Reason [1].

SFT Data Construction. From this broad collection, we first sampled a balanced subset of 20,000 video-question pairs to serve as the source for Supervised Fine-Tuning (SFT). Since these datasets lack explicit annotations for tool-augmented reasoning chains, we employed a knowledge distillation strategy utilizing Gemini-2.5-Pro [2] as the teacher model. Specifically, we prompted the teacher to generate multi-branch Tree-of-Tool reasoning paths based on our defined toolset (zoom, jump, slide). To align with our framework's objective of identifying a *minimal yet sufficient* chain of evidence, we applied an *efficiency-oriented filtering strategy*. For each query, we ranked the successful reasoning paths by length and retained up to the top-k ($k = 2$) shortest trajectories. Additionally, we explicitly preserved a subset of trajectories where the model actively triggered termination to ensure robust exploration capabilities. This process resulted in a final corpus of 33,724 high-quality trajectories for behavioral cloning.

RL Data Construction. Distinct from the SFT dataset, we sampled an *additional, non-overlapping subset* of 40,000

video-question pairs from the aggregated corpus to drive the RL stage. By using raw queries instead of teacher trajectories, the model generates its own reasoning paths optimized via ToT-RPO, balancing accuracy with efficiency and enhancing generalization.

C. Limitations

Despite the demonstrated effectiveness of TreeReasoner in handling complex video reasoning tasks, we acknowledge several limitations. First, the reliance on a tree search mechanism introduces a computational overhead compared to standard end-to-end IO of MLLMs. While our Reinforcement Learning strategy explicitly optimizes for shorter reasoning paths, the inference process still requires multiple forward passes to explore and prune branches, which may hinder deployment in strictly real-time scenarios. Second, our current toolset is limited to visual navigation primitives (zoom, jump, slide). While these are sufficient for spatiotemporal grounding, the model lacks tools for processing other modalities (e.g., audio analysis). Finally, our training paradigm depends on knowledge distillation from a teacher model (Gemini-2.5-Pro). Although we apply rigorous filtering to the distilled data, the student model's upper bound is inevitably influenced by the teacher's capabilities.

References

[1] Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, Sifei Liu, Hongxu Yin, Yao Lu, and Song Han. Scaling rl to long videos. 2025. 1

[2] Google. Gemini 2.5 pro: A multimodal reasoning model for video, audio, and text. .https://deepmind.google/models/gemini/pro/, 2025. 1

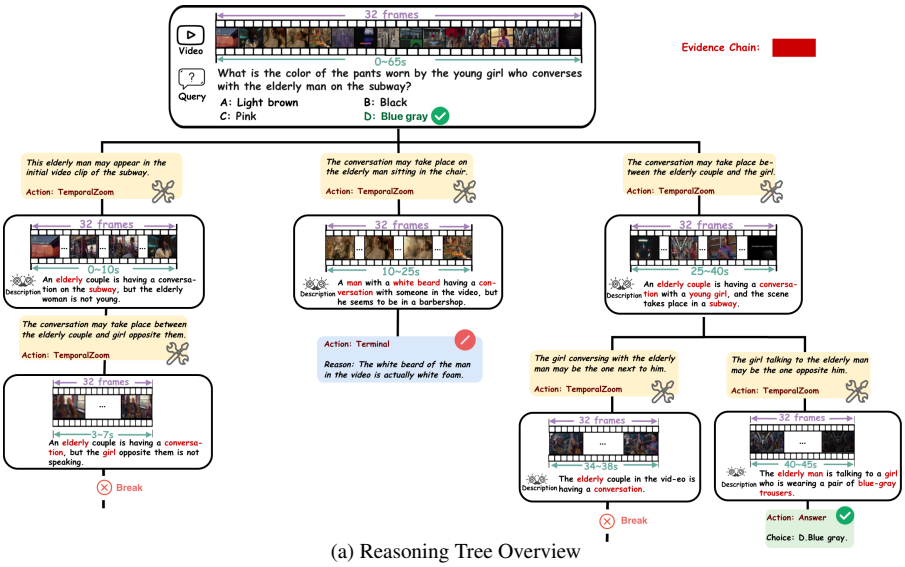
[3] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems*, 2023. 1

[4] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 1

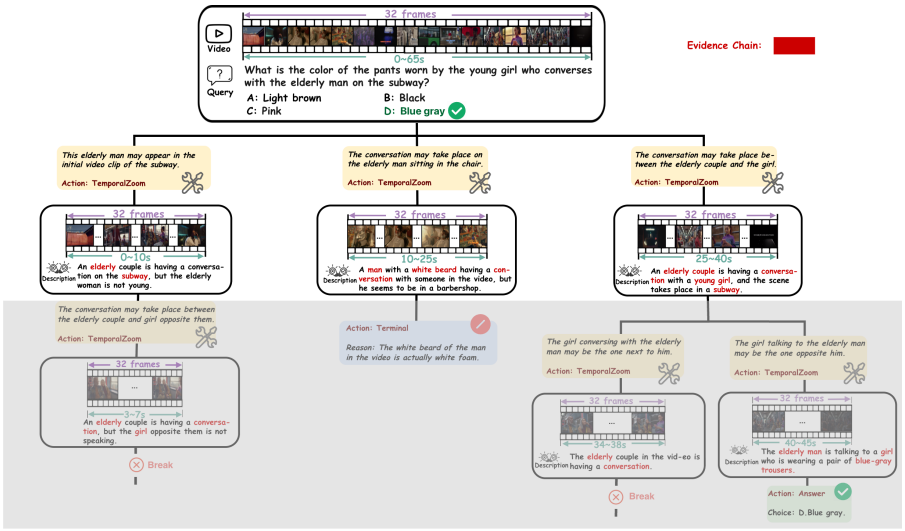
[5] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020. 1

[6] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024. 1

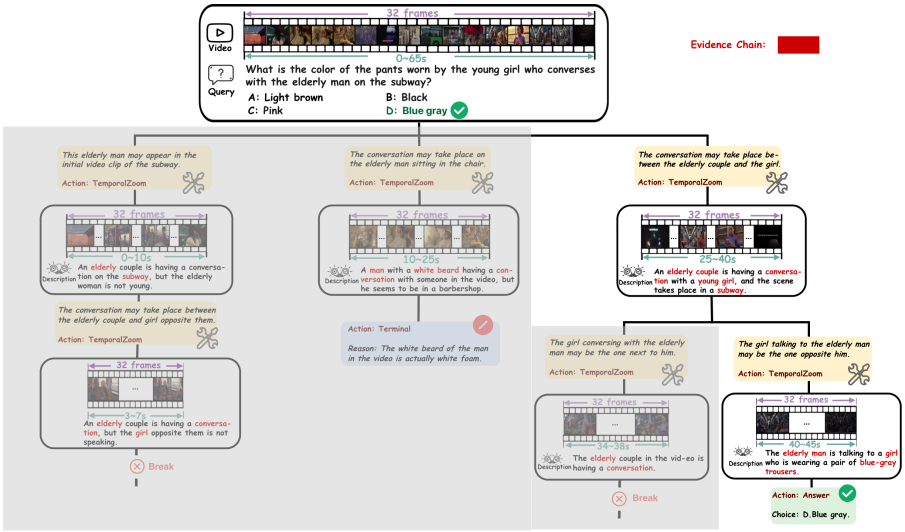
[7] Orr Zohar, Xiaohan Wang, Yonatan Bitton, Idan Szpektor, and Serena Yeung-levy. Video-star: Self-training enables video instruction tuning with any supervision. In *arXiv preprint arXiv:2407.06189*, 2024. 1



(a) Reasoning Tree Overview

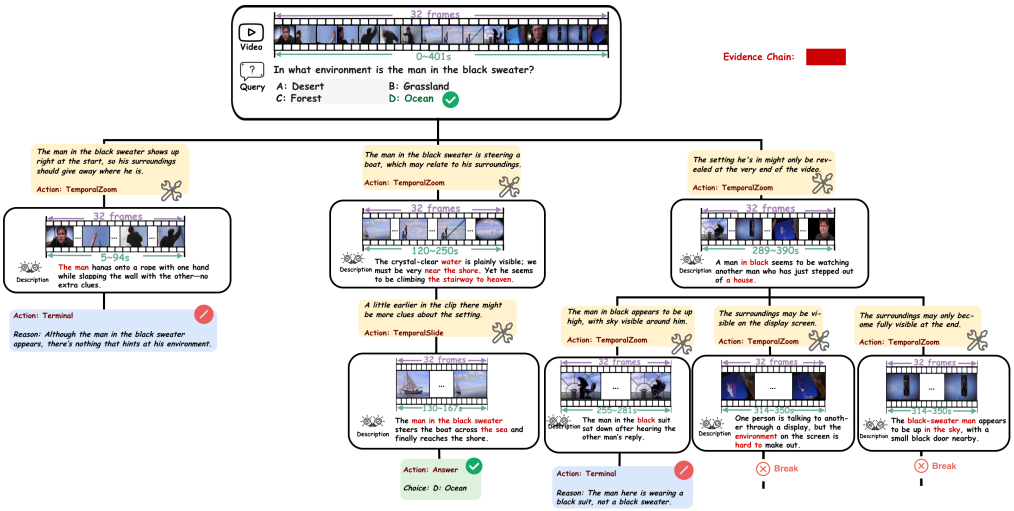


(b) Temporal Bracketing

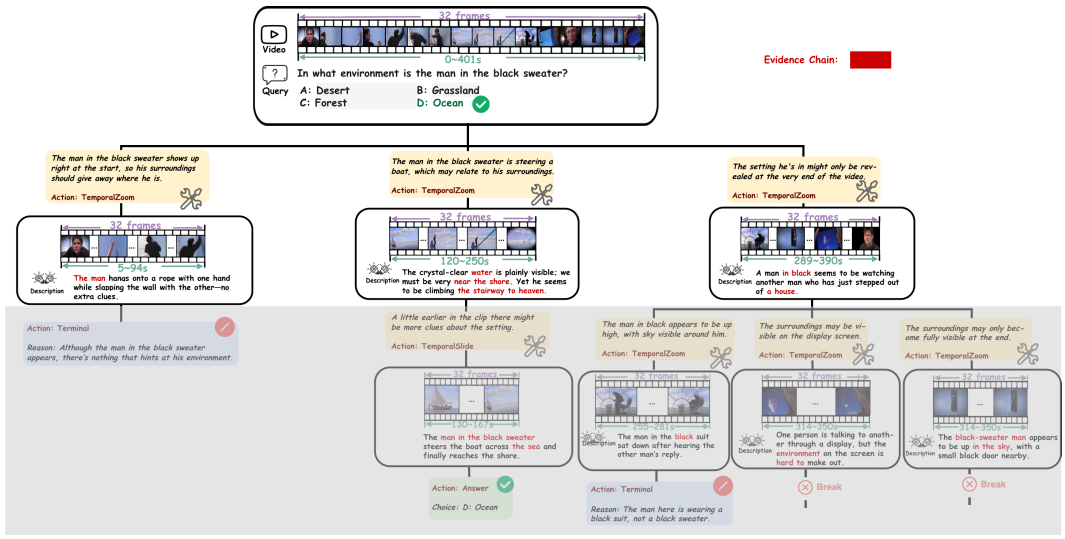


(c) Causal Chain Verification

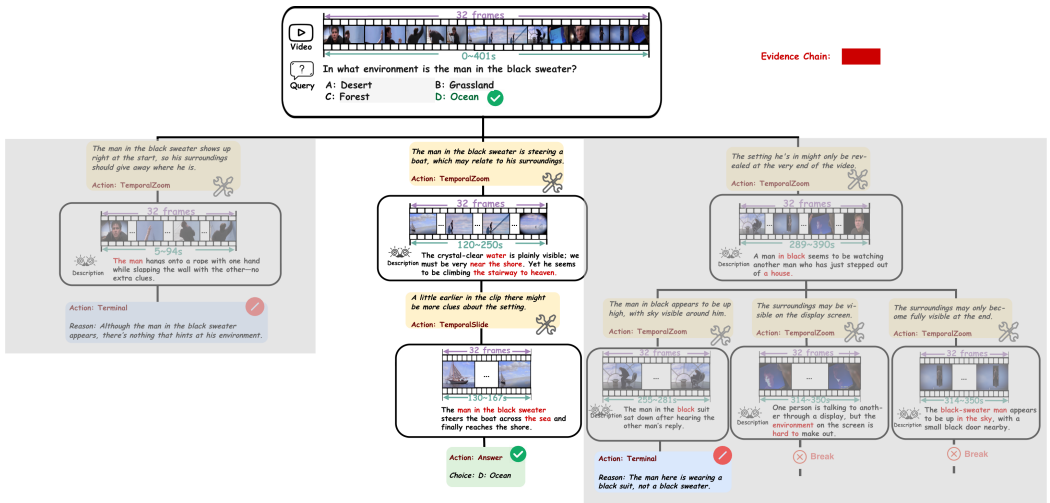
Figure 1. A case study on "identifying a girl's pants color query in subway scene": (a) shows an overview of the complete reasoning tree; (b) illustrates the temporal bracketing mechanism; (c) presents a branch trajectory along with its corresponding visual evidence chain. Irrelevant regions are masked in (b) and (c) to highlight key content.



(a) Reasoning Tree Overview



(b) Temporal Bracketing



(c) Causal Chain Verification

Figure 2. A case study on "the environment query of a man in black sweater": (a) shows an overview of the complete reasoning tree; (b) illustrates the temporal bracketing mechanism; (c) presents a branch trajectory along with its corresponding visual evidence chain. Irrelevant regions are masked in (b) and (c) to highlight key content.