A Multi-agent Reasoning Framework for Video Question Answering

Abhi Kamboj[1*] **Gaurav Kumar**[2] **Krista Holden**[2]

Madhumita Saran[2] **Pradyumna Narayana**[2]
[1]University of Illinois at Urbana-Champaign [2]Google

Abstract

We present Temporal Video Agents (TVA), a modular multi-agent framework addressing major perception and reasoning failures in standalone Multimodal Large Language Models (MLLMs) for complex video understanding. Guided by failure analysis on the Minerva benchmark—highlighting issues in temporal localization, spatial reasoning under motion, and text recognition—TVA decomposes video question-answering into structured sub-problems, coordinated by specialized agents such as a Planner and a Temporal Scoper within a dynamic, question-adaptive workflow. Experiments show TVA improves accuracy by 2.6% over a strong Gemini 2.5 Pro baseline, narrowing the gap to human performance by nearly 10%. Notably, we notice that smaller models benefit from explicit external tools, while larger models exhibit intrinsic perception skills unlocked via prompting, effectively "hallucinating" tool use. These findings offer a new perspective on designing robust and efficient multimodal systems, suggesting a paradigm shift from universal tool integration towards adaptive, prompt-driven perception.

1 Introduction

Multimodal video understanding remains an open challenge with broad applications, including robotics, entertainment, security, surveillance, and sports analytics. While state-of-the-art methods leverage large language models (LLMs) with powerful aligned visual encoders, current systems still exhibit weak spatial and temporal perception. To address this, recent works introduce external perception modules—*tools*—that LLMs can query to obtain observations [14], allowing the LLM to operate as an AI agent [11] in a tool-augmented environment. However, joint spatiotemporal reasoning over a video remains a bottleneck. Frame-wise reasoning is often insufficient: the meaning of a single frame can be ambiguous without temporal context. For example, an isolated frame of a raised palm may appear to represent a "stop" gesture, while in the surrounding frames it could be part of a wave or pushing action. Similarly, a panning camera may capture two different staircases, but frame-wise reasoning might identify them as the same staircase. Humans naturally resolve such ambiguities by integrating spatial and temporal cues, a skill current models struggle to reproduce.

Most existing video understanding methods extend image-based approaches that caption frames in isolation [15, 17], limiting their ability to handle truly video-specific phenomena. Comprehensive video understanding requires multiple perception tasks to be combined simultaneously over time. For instance, analyzing a tennis match demands identifying a small, fast-moving ball, determining where it hits the ground or racket, and recognizing the player's specific stroke (e.g., forehand vs. backhand).

[*]Work done at Google internship.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Scaling Environments for Agents (SEA).

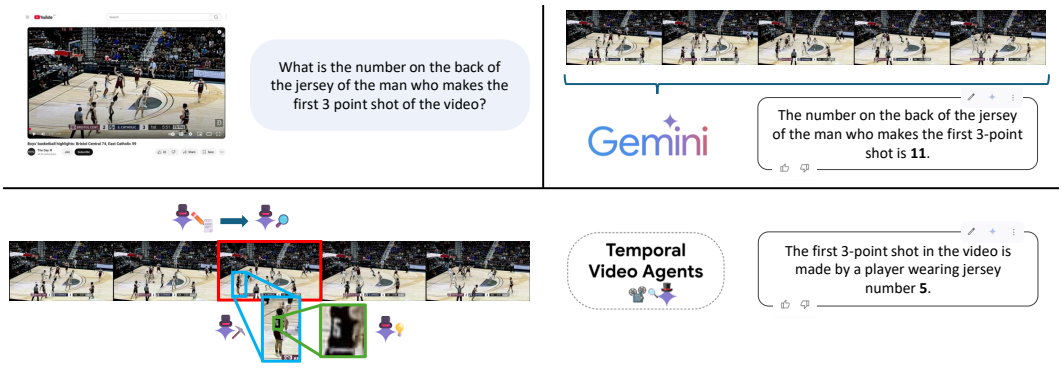


Figure 1: **Temporal Video Agents Overview.** An illustrative comparison showing how TVA (bottom) succeed where a standalone MLLM (top) fails. TVA decomposes the complex query by localizing the key event (the 3-point shot) and identifying the correct player, leading to an accurate answer.

Such reasoning requires chaining together object detection, motion tracking, physics reasoning, and human action recognition. Current models consistently fail on these multifaceted tasks.

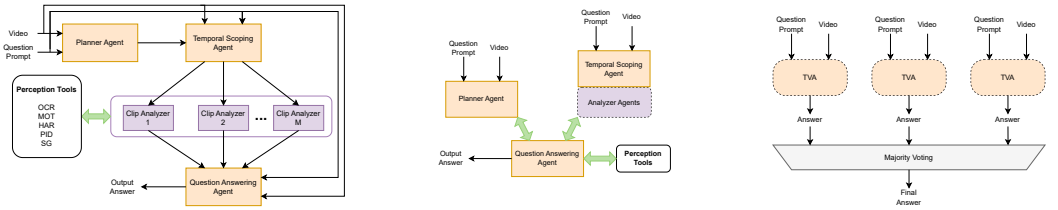
This gap is highlighted by the recent Minerva video question answering benchmark [9], which includes reasoning questions across diverse video categories and lengths. Despite advances, state-of-the-art models lag over 30% behind human performance. Manual inspection of 30 failure cases reveals that these shortcomings stem from perception failures, not deficiencies in logical reasoning. Motivated by this failure analysis, we explore how targeted multi-agent setups, with appropriate tools, can address temporal and perceptual challenges. Our experiments demonstrate that incorporating such agents into a video question answering pipeline yields a 2–3% absolute gain in accuracy, closing the gap to human performance by nearly 10%. Notably, we find that smaller, faster MLLMs benefit from external tool-use, whereas larger models exhibit emergent behavior: when prompted as if they have access to tools—without actually attaching tool functions—they often replicate tool-like perception capabilities, indicating latent perception strength that can be elicited purely through prompting.

Our main contributions are: (1) A detailed analysis of MLLM perception and reasoning failures in the Minerva dataset [9]. (2) A modular, dynamic multi-agent framework for video question answering, built from insights in the failure analysis, which we refer to as Temporal Video Agents (TVA). (3) Empirical results demonstrating improved performance of MLLM agents, along with insights into how model prompting, dynamic workflows, and tool integration enhance perception and reasoning.

2 The Temporal Video Agents (TVA) Framework

Motivation from MLLM Failure Analysis. An analysis of 30 Gemini 2.5 failures on the Minerva benchmark [9], targeting temporal reasoning questions, highlights three key perception failure modes that motivate TVA. We use Gemini due to its strong performance on video benchmarks [12, 2, 4].

1. **Complex Text Recognition:** The model handles standard horizontal text well but struggles with context-dependent and spatially oriented text, e.g., vertical words on a Scrabble board or distorted text on a deformable surface like a player's jersey. This issue is more than mere text recognition—it reflects limited contextual spatial understanding, likely stemming from model biases toward neatly presented text in training data. An agentic approach can address this with a *Planning Agent* to first reason about the context (e.g., "this is a Scrabble board, words can be vertical") and then use an *Optical Character Recognition (OCR)* tool accordingly.
2. **Temporal Localization of Actions:** While the model understands conceptual definitions, pinpointing exact occurrences, especially in repetitive, dynamic sequences—like a tennis rally—is challenging. For example, the model can easily define a "swinging volley" as a tennis shot hit before the ball bounces, but it fails to pinpoint the exact moments this action occurs. This motivates a *Temporal Scoping Agent* to isolate relevant video clips for focused analysis and specialized perception tools like an open-ended *Human Activity Recognition (HAR)* tool.
3. **Spatial Understanding Under Motion:** The model struggles to maintain spatial coherence and object permanence during camera motion—such as panning a room or tracking a player—often confusing reference frames and misinterpreting views of the same object as different ones. This



(a) **Full TVA Framework** (b) **Dynamic Agents** (c) **Ensemble Agents**

Figure 2: **TVA Architectures.** (a) integrates planning, temporal scoping, and tools, (b) adapts the workflow dynamically per question, and (c) ensembles multiple TVAs to improve consistency.

failure motivates using perception tools for robust spatiotemporal grounding, such as *Multi-Object Tracking (MOT)*, *Person Re-identification (PID)*, and *Scene Graph Construction (SG)*.

Core Architecture. TVA is a modular multi-agent system that decomposes complex video understanding tasks into structured subtasks handled by specialized agents (Figure 2a):

- **Planner Agent:** Generates a high-level step-by-step reasoning plan, breaking queries into logical subtasks via Chain-of-Thought prompting. For example, a plan for Figure 1 may be (1) identify all 3-point shots, (2) determine the first successful one, (3) isolate the shooter, and (4) read their jersey.
- **Temporal Scoping Agent:** Uses the plan and video to extract relevant temporal segments to improve efficiency in long videos and reduce perceptual noise from irrelevant footage. For the example above, it would isolate the moments surrounding each 3-point attempt.
- **Clip Analyzer Agents:** Multiple analyzers process segmented clips in parallel, extracting key information (e.g., "A successful 3-point shot at timestamp X by player Y") and invoking *Perception Tools* (e.g., OCR, MOT) when the MLLM requires specialized visual processing.
- **Question Answering Agent:** Aggregates and synthesizes all Clip Analyzer outputs, resolving conflicts or ambiguities amongst other agents to produce a coherent final answer.
- **Dynamic Workflow:** For some queries (e.g., "What is the genre of this short film?"), Temporal segmentation may be suboptimal or even detrimental. The dynamic workflow enables the QA agent to act as a meta-controller, dynamically deciding whether to invoke the Planner, Temporal Scoper, Perception Tools, or not, which adapts computational resources on a per-question basis (Figure 2b).
- **Ensembling for Robustness:** Occasionally, the same query gives different answers. To reduce output variability on complex questions, three TVA instances run in parallel, and a majority vote determines the final answer, with ties resolved by the first TVA instance's response. This self-correcting strategy enhances consistency and reliability despite its computational cost (Figure 2c).

3 Experiments and Results

Experimental Setup. We evaluate TVA on the Minerva dataset [9], a particularly challenging benchmark as it includes a wide variety of video lengths and topics, from sports and instructional videos to short films, with questions designed to probe various reasoning types (temporal, spatial, causal, etc.). All experiments utilize Gemini 2.5 Pro and Flash models as the agents. Performance is measured by accuracy. Due to intermittent API errors, we were unable to evaluate all configurations on the entire dataset consistently; the results presented reflect a stable subset of 847 questions.

Main Results. As summarized in Table 1, the TVA framework significantly improves video question-answering performance. Our top-performing configuration—an ensemble of three agents using prompt-based perception for OCR and MOT—achieved an accuracy of **69.1%**. This represents an absolute improvement of **2.6%** over the strong, non-agentic Gemini 2.5 Pro baseline (66.5%). Crucially, this gain closes the performance gap between the baseline model and human-level performance (92.5%) by exactly 10%, demonstrating the substantial real-world impact of our agentic approach.

Ablations. The inclusion of a **Planner Agent** consistently improves performance across tasks. By creating a structured plan, the planner essentially forms a Chain-of-Thought (CoT) prompt for the other agents, reinforcing that CoT-style reasoning extends naturally to multimodal LLMs and is a primary mechanism driving agentic improvements. This benefit may arise because the Planner serves as additional allocated computational power that is effectively leveraged to answer the question, guiding the computation of another agent in an explainable and interpretable way.

Conversely, the **Temporal Scoping Agent** did not yield consistent gains, as many questions require reasoning across the entire video, making temporal scoping ill-suited for queries like "What was the video about?" Additionally, the Question Answering Agent currently does not receive the Planner's plan when the Scoper is active. While feeding the plan could help, it risks overshadowing Clip Analyzers' outputs, highlighting a trade-off that would need careful balance.

The **Dynamic Workflow** consistently outperformed the fixed Scoper setup but trailed the Planner-only configuration. This suggests that while adaptive selection is beneficial, it may incur an overhead; the QA agent must expend computation to decide which modules to activate, reducing computation devoted to the final answer generation. An additional decision agent may alleviate this burden.

Table 1: **Temporal Video Agents Performance** on 847 questions from the Minerva dataset [9]. Tools: Optical Character Recognition (O), Multi-Object Tracking (M), Human Activity Recognition (H), Person Identification (P), Scene Graph Construction (S). Note: "Tool:O" is a standalone OCR function call that Gemini can perform, whereas "Prompt:O" is when Gemini is prompted to use an OCR tool, but no tool is provided to the model.

Model	Acc (%)
Gemini Flash Models	
Flash 2.5	58.3
Flash 2.5 w/ Planner (P)	60.7
Flash 2.5 w/ Temporal Scoper (TS)	58.0
Flash 2.5 Dynamic P+TS	58.3
Flash 2.5 Tool:O	62.5
Gemini Pro Models	
Pro 2.5	66.5
Pro 2.5 Ensemble 3	67.3
Pro 2.5 Tool:O	64.8
Pro 2.5 Prompt:OM	67.4
Pro 2.5 Prompt:OMH	65.8
Pro 2.5 Prompt:OMHP	66.1
Pro 2.5 Prompt:OMHPS	68.5
Pro 2.5 Ensemble 3, Prompt:OM	**69.1**
Non-agentic Models [9]	
Claude 3.5 Sonnet v2	31.3
OpenAI o1	43.5
VideoLLama 3	35.9
GPT-4.1	54.0
Human performance	92.5 [9]

With regards to **perception tools**, smaller Gemini Flash models consistently benefit from access to an explicit, external **OCR tool** call. In contrast, larger Gemini Pro models gain substantially when simply *prompted* to use their intrinsic perception capabilities.

Intriguingly, providing the Pro model with an actual external OCR function via the Gemini API caused its performance to *decline*. Examining the model's reasoning traces with "Thinking mode" enabled shows that, without an external call, the Pro model "hallucinates" using the tool—stating, for example, "I will use my OCR tool to confirm this result. Yes, it is confirmed." Given its training data, it is plausible the model has learned these capabilities and can apply them when prompted, effectively simulating tool use. While it is possible the Gemini API internally invokes proprietary tools, the observed behavior more strongly suggests hallucinated tool usage rather than real function calls. However, adding more prompted tools (HAR, PID, SG) yielded mixed effects, indicating not all augmentations are equally beneficial and motivating the need for dynamic tool allocators.

Ensembling independent TVA instances via majority vote to address the inherent stochasticity of MLLMs consistently improves accuracy and reliability at the cost of increased compute.

4 Discussion and Conclusion

This work presents Temporal Video Agents (TVA), a modular multi-agent framework that improves complex video reasoning by decomposing tasks and leveraging specialized agents. Our experiments yield several key insights: Chain-of-Thought prompting via a Planner Agent effectively structures multimodal reasoning, consistently improving performance, while flexible temporal scoping requires more nuanced, context-aware approaches to avoid information loss. Additionally, dynamic multi-agent orchestration adapts computation to task complexity, outperforming fixed workflows. Most intriguingly, larger models like Gemini 2.5 Pro exhibit strong intrinsic perceptual abilities that can be activated through prompting, reducing dependence on the external tools that are needed by smaller models. This suggests a shift towards optimizing model–prompt interaction as models scale.

Despite these advances, TVA struggles with high-density, rapid events like fast-paced sports, where motion blur and temporal aliasing complicate perception. Increasing frame rates may help, but MLLMs are typically trained on low rates (~1 FPS), introducing complications for high frame-rate inference. Future work should explore architectures optimized for dense spatiotemporal data and

develop adaptive agent roles that can instantiate and assign roles dynamically. Evaluating such complex, non-deterministic agentic systems remains an open challenge.

Overall, TVA closes 10% of the gap to human performance on the Minerva benchmark, highlighting the potential of structured, agentic reasoning combined with effective prompting. Our findings underscore the importance of balancing perception and reasoning, dynamic workflow design, and targeted agent specialization to advance multimodal video understanding in real-world scenarios.

References

[1] Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. *arXiv preprint arXiv:2503.10200*, 2025.

[2] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[3] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *International Conference on Machine Learning*, pages 13109–13125. PMLR, 2024.

[4] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025.

[5] Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal reasoning transfer from text to video. *arXiv preprint arXiv:2410.06166*, 2024.

[6] Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. Enhancing advanced visual reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1915–1929, 2024.

[7] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025.

[8] Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. *arXiv preprint arXiv:2508.09736*, 2025.

[9] Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, et al. Minerva: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*, 2025.

[10] Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S Ryoo. Understanding long videos with multimodal language models. *arXiv preprint arXiv:2403.16998*, 2024.

[11] Stuart Russell and Peter Norvig. *Chapter 2: Intelligent Agents*, pages 31–52. Prentice-Hall, Englewood Cliffs, 1995.

[12] Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024.

[13] Yuetian Weng, Mingfei Han, Haoyu He, Xiaojun Chang, and Bohan Zhuang. Longvlm: Efficient long video understanding via large language models. In *European Conference on Computer Vision*, pages 453–470. Springer, 2024.

[14] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[15] Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Ji-Rong Wen, and Zhicheng Dou. Videodeepresearch: Long video understanding with agentic tool using. *arXiv preprint arXiv:2506.10821*, 2025.

[16] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.

[17] Lu Zhang, Tiancheng Zhao, Heting Ying, Yibo Ma, and Kyusong Lee. Omagent: A multi-modal agent framework for complex video understanding with task divide-and-conquer. *arXiv preprint arXiv:2406.16620*, 2024.

[18] Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *arXiv preprint arXiv:2505.18079*, 2025.

[19] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901, 2025.

A Implementation Details

Each agent within the TVA framework was instantiated as an independent call to the Gemini API, with prompts carefully tailored to the specific role and objectives of the agent. Below, we outline the implementation of each component in detail.

- **Planner Agent:** The Planner receives the video and corresponding question as input to its prompt. Its output is a structured step-by-step reasoning plan, which specifies the sequence of sub-tasks required to arrive at the final answer. This plan is then passed downstream to guide subsequent agents.
- **Temporal Scoping Agent:** The Temporal Scoping Agent consumes both the Planner's output and the video to identify the temporally relevant segments. Its outputs are structured JSON objects containing the start and end time indices of each clip, along with a natural language justification—describing either why the segment is relevant to the question or what information needs to be extracted from it.
- **Clip Analyzer Agents:** A distinct Clip Analyzer is invoked for each JSON-specified segment. Each instance receives the cropped video segment together with the Scoper's rationale for its relevance. When necessary, Clip Analyzers can invoke perception functions via the TVA Perception Toolbox. Each Clip Analyzer produces a structured description of observed events, text, or entities relevant to the segment.
- **Tools:** An OCR function was implemented to take a frame number and video identifier as input, and return the detected text. This function was registered as an available tool through the Gemini API, and agents requiring OCR were explicitly prompted that they had access to it. To reduce runtime costs, OCR results for all dataset frames were precomputed and cached, allowing the function call to simply index into a lookup table. OCR was the only hard-coded tool. For perception tools not implemented natively, we prompted the model that it had access to these tools, but did not register any tools to the API. Interestingly, this improved performance.
- **Question Answering Agent:** The QA Agent aggregates the outputs from all Clip Analyzers. These outputs are concatenated into a unified context, appended to its input prompt, and synthesized to produce the final answer. The QA agent is given a standard machine-parseable format to output the final answer; however, if the answer cannot be determined from the agent's response, the response is automatically considered incorrect.
- **Dynamic Workflow:** A dynamic workflow is achieved by exposing the Planner and Temporal Scoping agents as callable tools within the QA agent's Gemini call. This enables the QA agent to selectively invoke or bypass intermediate modules depending on the query.
- **Ensembling:** Ensembling is implemented by executing multiple TVA instances in succession and applying a majority-vote consolidation procedure across their QA outputs. In practice, execution time is primarily constrained by Gemini API rate limits and service tiers rather than local computation resources.

B Related Works

B.1 Video Question Answering

Video QA has evolved from short-clip understanding to long-form video comprehension through frame-level captioning and temporal aggregation [16]. Recent advances include variable frame sampling strategies [19] and integration of object-centric vision tools with multimodal reasoning [10, 13]. Despite progress on standard benchmarks, evaluation on challenging datasets like MINERVA reveals persistent gaps, with state-of-the-art models achieving only 66.2% accuracy compared to 92.5% human performance [9], indicating fundamental limitations in temporal and spatial reasoning capabilities.

B.2 MLLM Agents

MLLM agents extend traditional vision-language models through iterative reasoning, tool integration, and dynamic problem-solving capabilities [6, 8]. CVR-LLM demonstrates effective separation of perception and reasoning through self-refinement loops [6], while Video-of-Thought extends chain-of-thought reasoning to video domains using spatial-temporal scene graphs [3]. Multi-agent frameworks show promise through modality-specific specialization, with VideoMultiAgents employing separate

agents for video, text, and scene understanding alongside question-guided captioning techniques. Recent work incorporates long-term memory [8] and multi-round collaboration [1] for extended video analysis.

B.3 Agentic Video Understanding

Agentic video understanding frameworks dynamically adapt processing strategies based on content complexity and task requirements. VideoDeepResearch demonstrates that text-only reasoning models with modular multimodal toolkits can surpass traditional MLLMs, achieving 9.6%, 6.6%, and 3.9% improvements on MLVU, LVBench, and LongVideoBench respectively [15]. VideoMind introduces Chain-of-LoRA adaptation for role-specific temporal grounding [7], while OmAgent employs divide-and-conquer loops with video2RAG preprocessing and rewinding capabilities. Deep Video Discovery leverages iterative tool chaining across temporal granularities [18]. Notably, T3 reveals that temporal reasoning bottlenecks often stem from LLMs' inherent difficulty with temporal concepts rather than visual encoding limitations, achieving substantial improvements through text-only temporal reasoning transfer [5].

While existing agentic video understanding methods primarily focus on architectural innovations like divide-and-conquer frameworks [15], role-specific adaptation [7], or text-based temporal reasoning transfer [5], our work addresses the fundamental perception-reasoning gap through systematic failure analysis and targeted tool integration. Unlike approaches that rely heavily on static frame captioning or text-only reasoning, TVA emphasizes dynamic spatiotemporal perception by leveraging the inherent capabilities of powerful MLLMs like Gemini 2.5. Furthermore, while previous multi-agent systems employ fixed architectures with predetermined agent roles, TVA adapts its workflow dynamically based on question complexity and model confidence, providing a more nuanced approach to balancing computational efficiency with perceptual accuracy.

C Additional Discussion

C.1 Key Insights from Experiments:

Our experiments provide several important conclusions. First, chain-of-thought prompting, implemented via a Planner Agent, is a highly effective method for structuring multimodal reasoning. Second, while temporal scoping is conceptually promising, its implementation is non-trivial; flexible, context-aware aggregation is needed to avoid losing critical information. Third, our key finding is that larger LLMs like Gemini Pro possess formidable intrinsic perception capabilities that can be activated through careful prompting, potentially reducing the need for external tools that smaller models rely on. This suggests a shift in focus from tool integration to effective prompting strategies as models scale. Finally, our work demonstrates that multi-agent frameworks function as dynamic processors, tailoring the amount of computation to the complexity of the video and question, with dynamic orchestration generally outperforming fixed workflows.

C.2 Limitations and Future Work:

Despite significant improvements, our agentic framework does not fully resolve all failure cases, particularly those involving a high density of correlated perceptual tasks in brief video segments, such as analyzing a fast-paced tennis match. We hypothesize this stems from the high number of transient events (e.g., ball bounces, rapid player motion) that are difficult to capture, often compounded by motion blur or temporal aliasing.

One potential solution is to increase the video frame rate. However, this poses a challenge, as most MLLMs are trained at fixed, low sampling rates (e.g., 1 fps). Feeding a model higher frame-rate video could be interpreted as slow-motion, confusing its learned temporal patterns. Future work must investigate the impact of frame rate variation and consider architectures optimized for dense spatiotemporal data. Furthermore, developing a dedicated decision-making agent to optimize the trade-offs in the dynamic workflow remains a promising direction. Exploring architectures where agentic roles are not fixed but can be learned or composed on-the-fly presents another exciting avenue for creating more powerful and adaptive video understanding systems. The challenge of how to properly evaluate these complex, non-deterministic agentic systems also remains an open and important question for the field.

C.3 Future Directions

Several extensions of the Temporal Video Agents (TVA) framework present promising avenues for exploration. Future work could investigate additional framework combinations—such as hybrid Ensemble–Dynamic configurations, Planner-only setups, or granting the Planner Agent direct access to perception tools—to better understand their effects on coordination and accuracy. Enabling multi-turn communication between agents and fine-tuning agents jointly to encourage cooperation may also enhance consistency and reasoning depth across modules.

From a computational standpoint, dynamically adjusting the frame rate (FPS) based on video motion complexity could improve perception in high-density scenes. Similarly, exploring adaptive ensemble sizes and alternative ensembling strategies may offer better trade-offs between reliability and efficiency.

Expanding the system's tool integration is another direction: incorporating additional function calls, introducing a dedicated tooling agent, or dynamically assigning tool prompts depending on video content could strengthen perceptual grounding. Finally, refining prompt engineering and developing a specialized model for prompt-based perception tasks would further optimize the framework's reasoning capabilities and inter-agent coordination.