TIR-Flow: Active Video Search and Reasoning with Frozen VLMs

Anonymous ACL submission

Abstract

While VideoLLMs have achieved remarkable progress in perception, their reasoning capabilities remain a bottleneck. Existing solutions typically resort to a heavy data engineering paradigm—synthesizing large-scale Chain-of-Thought (CoT) datasets followed by Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL). This pipeline primarily optimizes probability sampling efficiency and aligns output distributions, but fails to activate the intrinsic intelligence required for dynamic visual exploration. In this work, we propose *TIR-Flow*, a novel framework that shifts the paradigm from passive processing to active video searching and reasoning without additional data or parameter updating. Concretely, our framework operates through three synergistic modules: *HDD* decomposes complex queries into a set of verifiable sub-tasks; *HAP* actively directs visual attention to gather high-resolution evidence for hypothesis validation; *EBA* maintains a persistent workspace to accumulate and update the discovered clues for logical reasoning. Extensive experiments on seven benchmarks demonstrate that TIR-Flow significantly outperforms recent strong baselines, delivering an average performance boost of 5.9%, with gains reaching 10.5% on Egoschema. Our analysis confirms that empowering frozen VLMs with System-2-like active perception is a scalable path toward solving long-horizon video reasoning.

1 Introduction

The rapid evolution of Large Vision-Language Models (VLMs) has revolutionized the field of video understanding (Zohar et al., 2025; Zhang et al., 2025a; Li et al., 2024b; Xu et al., 2025b; Jin et al., 2025a). While these models demonstrate impressive capabilities in basic visual perception, they fundamentally struggle with complex reasoning over long horizons. When faced with tasks requiring causal deduction or fine discrimination, standard Video-LLMs often miss critical details.

To bridge this gap, prevailing research relies heavily on data-centric optimization. Approaches such as synthesizing Video Chain-of-Thought (CoT) datasets (Qi et al., 2025; Zhang et al., 2025c) or applying Reinforcement Learning (RL) (Feng et al., 2025; Li et al., 2025b; Wang and Peng, 2025; Jin et al., 2025b; Lin et al., 2026) aim to refine the reasoning distribution.

However, this paradigm faces a critical limitation: it relies heavily on extensive data engineering to align the model with specific reasoning patterns. Consequently, while the model achieves high scores on familiar tasks, it struggles to generalize to out-of-distribution (OOD) scenarios characterized by unseen visual complexities. Drawing inspiration from human cognitive processes (Evans, 2003), we posit that complex video reasoning requires an active search mechanism. Just as a human expert does not passively watch a video but actively rewinds, zooms in, and verifies specific hypotheses, a VLM should be empowered to plan, look, and verify before answering.

To this end, we introduce TIR-Flow (Temporal Intervention Reasoning Flow), a unified, training-free framework designed to unlock the latent reasoning potential of foundation models. TIR-Flow shifts the focus from static parameter optimization to active perception. This framework utilizes a cascaded modular architecture to dynamically bridge the gap between low resolution global contexts and high fidelity local evidence. As illustrated in Figure 2, the framework operates through three cascaded modules: (i) *Hypothesis-Driven Decomposition*: Acting as the Planner, this module utilizes the LLM's semantic priors to decompose complex queries. (ii) *Hierarchical Active Perception*: Acting as the Scout, this module utilizes a Grid-Pyramid mechanism to actively zoom in on evidence from relevant temporal windows. (iii) *Evidence Black-*

board & Arbitration: Acting as the Solver, this module aggregates discrete visual observations into a structured memory buffer to support reasoning.

By decoupling reasoning planning from visual perception, TIR-Flow delivers substantial improvements across diverse benchmarks. Notably, it propels the Qwen2.5-VL-7B to challenge proprietary giants, significantly narrowing the performance gap with Gemini 1.5 Pro on EgoSchema and approaching the level of Gemini 2.5 Pro on MVBench. Our approach demonstrates that a frozen VLM, when equipped with the agency to actively seek information, can outperform models trained with extensive additional data engineering and training costs.

In summary, our contributions are three-fold:

- *Unlocking Reasoning Ceiling:* We propose TIR-Flow, a training-free framework shifting from static optimization (SFT/RL) to active perception. Its System-2 "Plan, Look, and Verify" agency enables frozen VLMs to surpass the reasoning capabilities of supervised baselines (see Table 1).
- *Bridging the Gap via Active Perception:* We introduce a closed-loop architecture synergizing decomposition, active scouting, and arbitration. This design aligns semantic planning with visual exploration, enabling the model to dynamically "zoom in" on critical spatiotemporal evidence to resolve information bottlenecks (see Section 3).
- *Superior Performance and Reliability:* TIR-Flow achieves best performance across seven benchmarks with a 5.9% average boost (see Figure 1 and Table 1). Crucially, our ablation analysis confirms these gains stem from genuine evidence grounding , ensuring reliable complex reasoning (see Section 5).

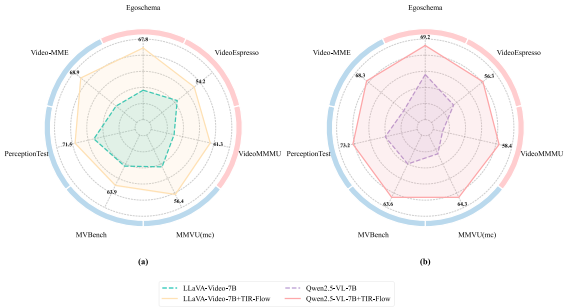

Figure 1: *Unlocking reasoning potential.* TIR-Flow consistently boosts the zero-shot performance of (a) LLaVA-Video-7B and (b) Qwen2.5-VL-7B across seven benchmarks, improving both perception and reasoning capabilities.

2 Related Work

2.1 Video-Language Models.

Recent years have witnessed the surge of Large Vision-Language Models (Video-LLMs) that extend static image encoders to the temporal domain (Lin et al., 2024a; Li et al., 2025b; Zhang et al., 2025a; Wang et al., 2025b; Bai et al., 2025). Standard paradigms typically operate in a single pass, feed forward manner. These models passively consume pre-defined visual tokens, lacking the agency to re-examine or look back at specific regions when initial reasoning proves insufficient. Consequently, fine-grained details are either lost in compressed global embeddings or buried under an overwhelming volume of redundant tokens.

2.2 RFT for Video Reasoning.

To enhance the reasoning capabilities of Video-LLMs, a dominant line of work pursues data-centric alignment via large-scale Video Chain-of-Thought (CoT) synthesis (Wang et al., 2025d; Li et al., 2025a) and post-training RL to enforce step-by-step reasoning (Zhang et al., 2025c; Chen et al., 2025; Feng et al., 2025; Jin et al., 2025b). However, recent studies (Yue et al., 2025; Huang et al., 2024; Kirk et al., 2024; Jin et al., 2026b) suggest these paradigms largely act as alignment, concentrating probability mass on pre-defined patterns rather than expanding the reasoning ceiling: pass@1 may improve, yet pass@k often stagnates or even drops. This limitation is amplified in Video-LLMs, where parameter-intensive optimization is costly and performance remains fundamentally bounded by perceptual bottlenecks in frozen encoders.

2.3 Test Time Scaling and Active Agents

A growing body of work studies inference-time techniques for boosting model performance (Yao et al., 2023a; Xu et al., 2025a), echoing System-2-style deliberation in human cognition (Wang et al., 2025c; Zhang et al., 2024; Yao et al., 2023b; Besta et al., 2023). In vision, agent-based approaches (Surís et al., 2023; Yang et al., 2023; Fan et al., 2024; Jin et al., 2026a) decompose tasks into sub-steps and invoke external tools, but often depend on fragmented toolchains (Wang et al., 2025a; Gao et al., 2025), leading to error prop-

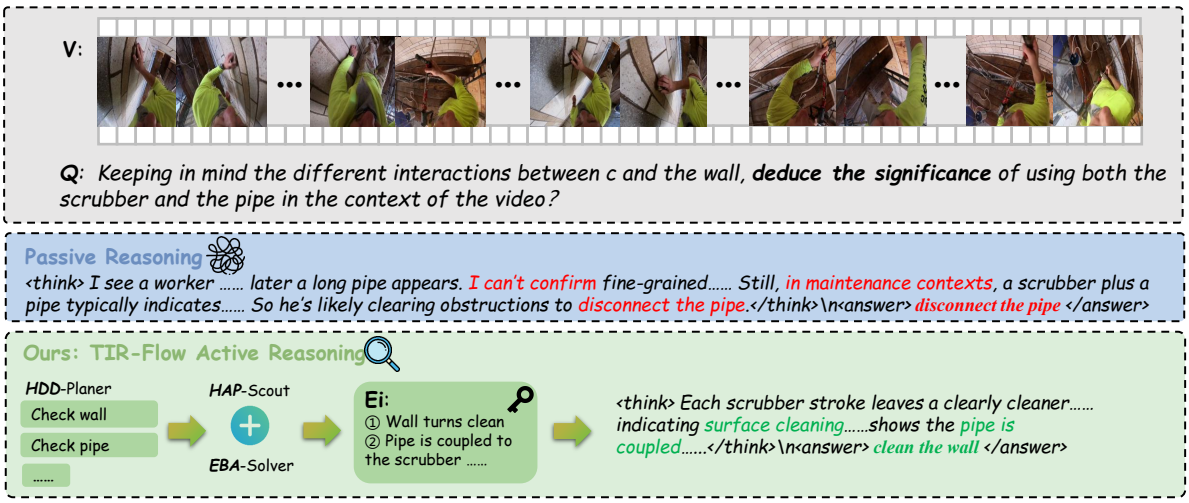

Figure 2: **Comparison of Reasoning Trajectories between Passive CoT and TIR-Flow Paradigm.** Without an explicit framework to guide the model in extracting and verifying visual information, it is challenging for the system to perform logical deductions grounded in the intrinsic content of the video.

agation and maintenance overhead. Meanwhile, training-free methods such as SmartSight (Sun et al., 2025), VideoTree (Wang et al., 2025d), InfiniPot-V (Kim et al., 2025), and STTM (Hyun et al., 2025) mitigate hardware constraints, but typically trade off memory, and reasoning fidelity. In contrast, TIR-Flow champions endogenous perception and leverages the VLM's native visual encoder for lightweight yet high-resolution active search, thereby avoiding error propagation and enabling effective inference-time scaling.

3 Method

3.1 Overview and Problem Formulation

We address the problems of VideoQA under a training-free setting. Given a long video $\mathcal{V} = \{v_1, v_2, ..., v_T\}$ and a complex natural language query Q, our goal is to generate an accurate answer A without fine-tuning the underlying Vision-Language Model (VLM).

Traditional approaches maximize $P(A|\mathcal{V}, Q)$ on fixed token sequences but suffer from redundancy and attention dilution. Instead, we formulate reasoning as a Modularized Flow, decomposing it into discrete steps where the model actively seeks visual evidence to verify sub queries. The inference process is modeled as:

$$A^* = \arg\max_{A} \sum_{\mathcal{P}, \mathcal{E}} P(A \mid \mathcal{V}, \mathcal{E}, \mathcal{P}) P(\mathcal{E} \mid \mathcal{V}, \mathcal{P}) P(\mathcal{P} \mid Q). \quad (1)$$

where $\mathcal{P}$ represents the reasoning plan, $\mathcal{V}$ represents the visual frame information extracted from the video and $\mathcal{E}$ represents the retrieved visual evidence. To implement this, we propose Temporal-Intervention-Reasoning Flow.

In Figure 3, we illustrate the framework of TIR-Flow, which enhances VLMs by following rules to fetch the right visual evidence during inference.

3.2 Hypothesis-Driven Decomposition

Complex queries (e.g., "Why did the driver get pulled over?") often lack direct visual mappings. The HDD module serves as a cognitive Planner. It utilizes the semantic reasoning capability of the LLM backbone to decompose the user query Q into a structured reasoning tree $\mathcal{T} = \{q_1, q_2, ..., q_N\}$.

At step i, the planner generates (q_i^{text}, q_i^{vis}), representing the logical query and a visual matching description, respectively.

$$\mathcal{P} = \text{LLM}_{plan}(Q) \rightarrow \{(q_1^{text}, q_1^{vis}), ...\} \quad (2)$$

This decomposition transforms abstract causal inquiries into verifiable visual existence checks.

3.3 Hierarchical Active Perception

To efficiently locate visual evidence $\mathcal{E}$ without processing the entire high-resolution video, we design a coarse-to-fine retrieval mechanism involving Temporal Scouting and Spatial Focusing.

Temporal Scouting. We first identify the relevant temporal windows. We employ a lightweight, dual-encoder model (e.g., CLIP) to extract frame-level

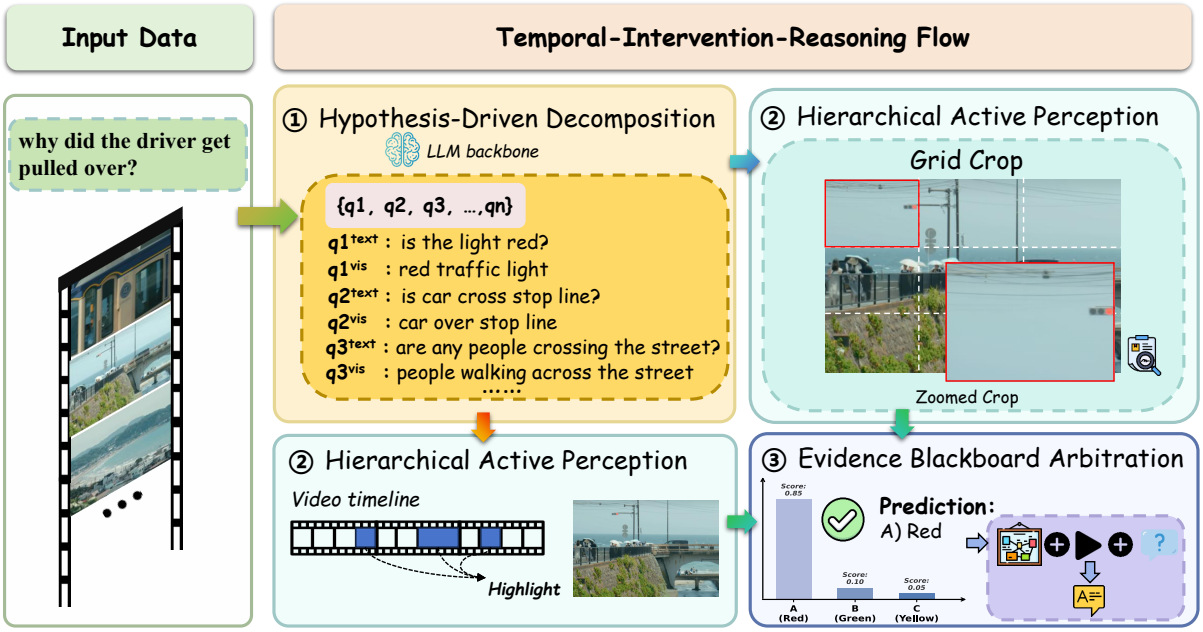

Figure 3: **Overall pipeline of TIR-Flow.** First, HDD breaks down the question into structured sub-queries. Then, HAP collects targeted evidence via coarse-to-fine spatial crops. Finally, an evidence blackboard arbitrates and scores the evidence, and the prediction is fused with the query to generate the final response.

features F_t from the video at a fixed sampling rate. For a visual query q_i^{vis}, we compute the cosine similarity scores

$$S_t = \cos(E_{img}(v_t), E_{txt}(q_i^{vis})) \quad (3)$$

Instead of selecting discrete frames, we apply a temporal smoothing window (kernel size $k = 5$) to capture continuous actions and select the top-K time intervals $\mathcal{W} = \{w_1, ..., w_K\}$ with the highest aggregated scores.

Grid Pyramid Strategy. Standard resizing of images often obliterates small but crucial details (e.g., a traffic light in the distance). To address this, we introduce a Spatial Grid-Pyramid strategy. For a candidate keyframe $I \in \mathcal{W}$, we construct a set of patches $\mathbb{P}$ containing: (i) The Global View: I_{global} (resized to preserve context). (ii) Local Grid Crops: We divide I into an $N \times N$ grid.

We then employ the CLIP visual encoder to score these patches against q_i^{vis}. The most relevant spatial region is selected as the visual evidence e_i:

$$e_i = \arg\max_{p \in \mathbb{P}} \left(\text{Sim}(E_{img}(p), E_{txt}(q_i^{vis}))\right) \quad (4)$$

This mechanism allows the model to zoom in on relevant regions dynamically, effectively increasing the effective resolution by a factor of N^2.

3.4 Evidence Blackboard & Arbitration

Unlike traditional approaches that treat reasoning as a linear chain, we model the evidence accumulation as a dynamic state maintenance process. The EBA module serves as the central memory unit, maintaining a Evidence Blackboard $\mathcal{B}_t$, which stores a structured history of verified visual facts in natural language form.

Evidence Arbitration Mechanism. At each reasoning step t, the VLM acts as an Arbitrator to validate the newly retrieved visual evidence e_t (from Section 3.3) against the current blackboard state $\mathcal{B}_{t-1}$ and the sub-query q_t^{text}. Formally, we define the arbitration function as:

$$(\mathcal{O}_t, s_t, \gamma_t) = \text{VLM}_{\text{arb}}(e_t, q_t^{text}, \mathcal{B}_{t-1}) \quad (5)$$

Here, $\mathcal{O}_t$ is the textual observation from e_t, $s_t \in [0, 1]$ is the confidence score, and $\gamma_t \in \{0, 1\}$ denotes a conflict indicator. We prompt the VLM to set $\gamma_t = 1$ if e_t contradicts $\mathcal{B}_{t-1}$ or appears ambiguous (see Appendix B for prompt details).

Closed-Loop State Update and Re-planning. Based on the arbitration output, the system executes one of two atomic operations: (i) Update (Acceptance): If the confidence s_t exceeds a threshold τ and no conflict is detected ($\gamma_t = 0$), the observation is accepted. The blackboard state updates by appending the new fact: $\mathcal{B}_t = \mathcal{B}_{t-1} \cup \{\mathcal{O}_t\}$. (ii) Re-

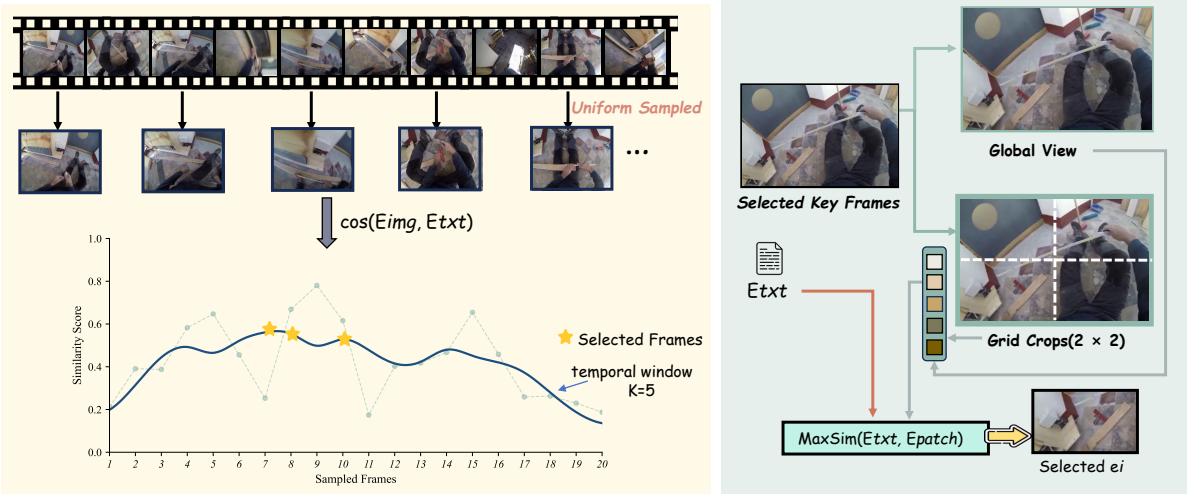

Figure 4: *Two Stage Grid Crop Pipeline:* Similarity Based Keyframe Scouting ($k = 5$) followed by Grid-Pyramid Patch Retrieval.

fine (Rejection): If $s_t < \tau$ or $\gamma_t = 1$, the evidence is deemed unreliable. The EBA triggers a Refinement Signal back to the HDD module (see Section 3.2). This signal carries the error type (e.g., object occlusion or temporal mismatch), prompting the HDD to regenerate a more granular sub-query $\hat{q}_t$ to re-examine the video.

This mechanism ensures that the final reasoning is grounded solely on high-confidence, non-contradictory visual evidence, significantly reducing hallucinations and effectively mitigating the fragility of the initial query decomposition.

4 Experiment

4.1 Settings

Baselines. We evaluate TIR-Flow against a diverse set of state-of-the-art methods, categorized into: (i) Proprietary Models: GPT-4o (OpenAI, 2024), Gemini (Comanici et al., 2025a). (ii) Open-source Models: VideoLLaMA3-7B (Zhang et al., 2025a), Video-R1-7B (Feng et al., 2025), Qwen3-VL-8B (Bai et al., 2025), and LLaVA-OneVision-7B (Li et al., 2024a). (iii) Training-Free Frameworks: We include SmartSight (Sun et al., 2025) (built on Video-R1-7B), VideoTree (Wang et al., 2025d), InfiniPot-V (Kim et al., 2025), and STTM (Hyun et al., 2025). Please refer to the Appendix C for more details.

Implementation Details. We separately use Qwen2.5-VL-7B and LLaVA-video-7B as our base models. We represent each video as a sequence with maximum T frames. T is initialized to 32 here. Regarding the hyperparameters, we set the window size $k = 5$, the Top-K selection to 3, and the EBA judgment threshold $\tau = 0.7$. The input frame I is spatially partitioned into an $N \times N$ grid (with $N = 3$). All experiments are conducted on NVIDIA V100 (32GB) GPUs.

4.2 Results

In this section, We evaluate TIR-Flow against Qwen2.5-VL and LLaVA-Video across reasoning and perception benchmarks. As shown in Table 1, our training-free framework demonstrates competitive performance against state-of-the-art baselines, effectively bridging the gap between raw visual input and complex reasoning.

TIR-Flow significantly improves performance on reasoning-intensive tasks. Experimental results in Table 1 demonstrate our framework's effectiveness. While proprietary models (e.g., Gemini 1.5 Pro) achieve strong results via costly alignment, and training-free methods like VideoTree and STTM offer only marginal gains, TIR-Flow delivers substantial improvements without parameter updates. For instance, TIR-Flow boosts average accuracy by *6.7%* on multiple reasoning benchmarks. With this framework, Qwen2.5-VL-7B attains 69.2% on EgoSchema and consistent gains of 7.2% and 9.3% on VideoMMMU and VideoEspresso, respectively. These findings suggest that task-guided active search unlocks intrinsic reasoning potential of frozen VLMs, transforming latent knowledge into structured deductions.

Models	Frames	Reasoning				Perception		
		Egoschema	VideoEspresso	VideoMMMU	MMVU(mc)	MVBench	PerceptionTest	VideoMME
Proprietary models								
GPT-4o	64	-	-	61.2	**75.4**	-	-	-
Gemini-1.5-Flash	128	65.7	39.8	49.8	-	-	-	76.1
Gemini 1.5 Pro	128	**72.2**	44.2	53.9	-	-	-	78.6
Gemini 2.5 Pro	-	-	**83.6**	**74.9**	-	**69.9**	-	**85.1**
Training-Free models								
SmartSight†	32	-	-	51.9	-	-	-	60.2
VideoTree†	-	66.2	-	-	-	-	-	53.1
InfiniPot-V†	768	65.8	-	-	-	-	-	61.1
STTM†	1fps	58.6	-	-	-	-	-	62.6
Training models								
LLaVA-Mini-8B	1fps	51.2	-	-	-	44.5	-	-
VILA-40B	256	58.0	-	34.0	-	-	54.0	-
LLaVA-onevision-7B	32	60.1	56.1*	57.1	49.2	33.9		
VideoLLaMA3-7B	180	63.3	-	-	44.1	69.7	72.8	66.2
Video-R1-7B	32	-	-	52.3	64.2	63.9	-	59.3
Qwen3-VL-8B-Instruct	2fps	60.6*	52.0*	65.3	65.6	68.7	63.7*	71.4
LLaVA-Video-7B	32	57.3	48.8	34.4	48.8	58.6	67.9	63.3
LLaVA-Video-7B+TIR-Flow	32	67.8 (10.5↑)	54.2 (5.4↑)	41.3 (6.9↑)	56.4 (7.6↑)	63.9 (5.3↑)	71.5 (3.6↑)	68.9 (5.6↑)
Qwen2.5-VL-7B	32	65.6	47.0	51.2	61.3	59.0	69.1	62.4
Qwen2.5-VL-7B+TIR-Flow	32	69.2 (3.6↑)	56.3 (9.3↑)	58.4 (7.2↑)	64.3 (3.0↑)	63.6 (4.6↑)	**73.2** (4.1↑)	68.3 (5.9↑)

Table 1: Performance on reasoning and perception video QA multiple choice benchmarks. * indicates the result we reproduced using a 32-frame setting. † indicates training-free framework. The best performance for each metric is **bolded**, and the second best is underlined.

We attribute this to Perception-Reasoning Decoupling. Traditional models often succumb to the vanishing detail problem by entangling feature extraction and deduction. By offloading evidence search to HAP and EBA modules, TIR-Flow frees the LLM backbone to function purely as a cognitive orchestrator. This implies the primary bottleneck in VLMs is not reasoning capacity, but the high entropy of the input space, which TIR-Flow manages through structured deductions.

TIR-Flow notably enhances perception-heavy tasks. As shown in Table 1, our framework mitigates information dilution inherent in static sampling via a hierarchical zoom-in strategy. This targeted acquisition preserves critical visual semantics with high fidelity. Consequently, TIR-Flow boosts LLaVA-Video-7B by *5.6%* on VideoMME, capturing subtle details missed by base models. This gain stems from a Signal-to-Noise Ratio advantage: active search dynamically filters noise and amplifies salient patches. Relying on intrinsic semantics rather than memorized patterns, our approach remains robust across novel and complex scenes.

In summary, these results demonstrate that TIR-Flow offers a significant improvement in both reasoning and perception tasks by effectively shifting the computational burden from parameter optimization to active perception.

5 Ablation Studies

5.1 Effectiveness of different modules

To further reveal the underlying mechanism of complex video reasoning, we conduct an ablation study by evaluating three variants of TIR-Flow, each distinct in its component configuration: (i) TIR-Flow w/o HDD: Bypasses decomposition, directly utilizing the original query Q as the probe for HAP-based evidence retrieval. (ii) TIR-Flow w/o HAP: Eliminates the active spatiotemporal scouting mechanism. It directly feeds uniformly sampled frames alongside the sub queries into the EBA module for evidence extraction. (iii) TIR-Flow w/o EBA: Omits the filtering and arbitration mechanisms, directly accumulating all retrieved evidence for final inference without validity verification.

Modules	Ego	Espresso	MVBench	Per
Base	57.3/65.5	48.8/47.0	58.6/59.0	67.9/69.1
+HAP+EBA	+5.8/+0.5	+2.2/+4.5	+1.8/+0.6	+0.2/+0.3
+HDD+HAP	+8.0/+2.1	+4.5/+7.2	+3.5/+2.9	+2.6/+2.7
+HDD+EBA	+2.6/+1.2	+2.9/+2.5	+2.3/+2.2	+2.4/+1.9
+HDD+HAP+EBA	+10.5/+3.7	+5.4/+9.3	+5.3/+4.6	+3.6/+4.1

Table 2: **Ablation Study.** Base model is LLaVA/Qwen. **Ego**: Egoschema. **Per**: PerceptionTest.

As shown in Table 2, the full TIR-Flow framework achieves the best performance across all benchmarks, validating the synergy of the proposed modules. We analyze the contribution of each com-

ponent based on the quantitative drops observed when they are removed:

Impact of Hierarchical Decomposition. The HDD module proves indispensable for reasoning-intensive tasks, serving as the cognitive roadmap for complex queries. When removing HDD, we observe the sharpest performance decline on EgoSchema, dropping from 10.5% to 5.8%. This empirical evidence confirms that without structured task decomposition, the model struggles to bootstrap the complex logical deductions required for long-form video reasoning, effectively reverting to shallower understanding.

Impact of Active Perception. The HAP module is critical for handling fine visual details that are typically lost in standard downsampling processes. The removal of HAP leads to a noticeable drop on perception benchmarks like MVBench (from 5.3% down to 3.3%) and EgoSchema (decreasing by 3.3%). This validates that HAP effectively addresses the perceptual resolution bottleneck. By employing an active strategy, it increases the visual signal-to-noise ratio, ensuring subsequent reasoning is based on high-fidelity evidence.

Impact of Evidence Arbitration. Finally, the exclusion of EBA results in consistent performance degradation, exemplified by a 2.5% drop on EgoSchema and a 1.8% drop on MVBench. This widespread decline underscores the critical importance of a structured memory buffer for cross verifying disparate evidence. Video streams inherently contain redundant and occasionally ambiguous information; without EBA's filtering and arbitration mechanisms, the model accumulates noisy or logically contradictory observations. This error propagation diminishes the model's capacity to resolve visual ambiguities, leading to lower confidence and accuracy during the final inference stage.

5.2 Micro-Ablation

Spatiotemporal Synergy within HAP. The HAP module is built upon two pillars: *Temporal Scouting* for identifying critical moments and *Spatial Focusing* (Grid Pyramid) for resolving fine-grained details. To disentangle their individual contributions, we evaluate two sub-variants on representative benchmarks: EgoSchema and MVBench. **w/o Spatial Focusing:** This variant retains the temporal scouting mechanism to retrieve relevant time windows but disables the Grid Pyramid strategy. The selected frames are processed solely via the global view, removing the model's ability to zoom in. **w/o Temporal Scouting:** This variant replaces the relevance based temporal scouting with standard uniform sampling. While it retains the Grid Pyramid for spatial enhancement, the model blindly distributes its high-resolution budget across the video, potentially missing sparse temporal cues.

Configuration	Mechanism		EgoSchema	VideoMME
	Temporal	Spatial		
w/o S&T	×	×	59.9/66.7	64.2/63.1
w/o Spatial	✓	×	65.2/67.5	67.0/66.9
w/o Temporal	×	✓	63.4/65.6	66.8/66.5
Full HAP	✓	✓	**67.8/69.2**	**68.9/68.3**

Table 3: **Component Analysis within HAP.** Base model is LLaVA/Qwen. We report accuracy (%) on EgoSchema and VideoMME.

Analysis. Table 3 reveals the distinct roles of each component: Temporal Scouting dominates long-horizon reasoning. On EgoSchema benchmark, which relies on locating sparse cues in long videos, removing the temporal mechanism causes a more severe drop (69.2% $\rightarrow$ 65.6%) compared to removing spatial focusing (69.2% $\rightarrow$ 67.5%). This confirms that blindly applying high resolution to the wrong timeframe is futile, locating the *critical window* is the prerequisite for success. Spatial Focusing resolves information bottlenecks. On VideoMME dataset, even when the correct temporal window is identified, the performance drops from 68.9% to 67.0% without spatial focusing. This indicates that standard resolution even at the right moment fails to capture the fine-grained details required for precise answering. The best performance is achieved only when the model looks at the *right time* with the *right resolution*.

5.3 Sensitivity Analysis of Hyperparameters.

To balance model performance with computational efficiency, we conducted a comprehensive analysis of the key hyperparameters in TIR-Flow: the Sliding Window Size k. Additionally, we evaluate the sensitivity of the confidence threshold τ in the EBA module, which is detailed in Appendix E.

Effectiveness of Sliding Window Size k. To investigate the impact of temporal granularity on active perception, we evaluate the performance of the HAP module across various sliding window sizes, specifically $k \in \{1, 3, 5, 7, 9\}$. As illustrated in Figure 5, we observe that performance follows a bell shaped trend as k increases. The exclusion of the sliding window (equivalent to $k = 1$) leads to an average degradation of 2.75%, confirming

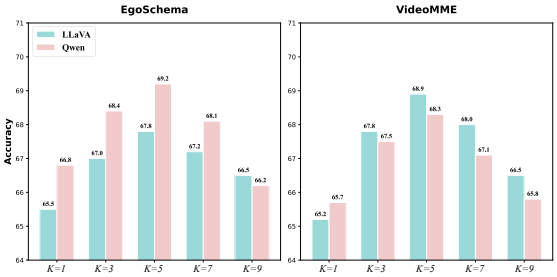

Figure 5: Impact of Diverse Sample Size (k) on EgoSchema and VideoMME.

that localized temporal anchoring is essential for mitigating interference from redundant frames. Notably, $k = 5$ emerges as the optimal configuration. Smaller window sizes ($k \leq 3$) tend to oversegment the video, potentially fracturing the continuity of complex actions, while excessively large windows ($k \geq 9$) suffer from information dilution, where the model's attention is insufficiently focused on salient events. These results suggest that an appropriately sized sliding window acts as a crucial temporal filter, allowing TIR-Flow to strike an optimal balance between local visual precision and global temporal context.

5.4 Oracle-Based Evidence Quality Assessment

To rigorously validate the information density of the retrieved evidence, we conduct a cross-model evaluation using Qwen3-VL-32B-Instruct as an external Oracle Judge.

Settings. We randomly sample 200 instances from VideoMME. For each instance, we compare the *Average Relevance Score (ARS)* of the visual inputs provided by two methods: *Baseline (Uniform Sampling):* We uniformly sample $N = 16$ frames from the video. To quantify the signal-to-noise ratio, we feed these frames individually to the Oracle Judge. The final score for the baseline is calculated as the mean of the scores of all 16 frames. *Ours (Active HAP):* We take the single, final zoomed-in image crop (e_i) selected by the HAP module. This frame is scored once by the Oracle.

The Oracle Judge is prompted to rate each image on a scale of 1-5 based on whether it provides critical visual evidence to answer the user's query. Prompt details are shown in Appendix B.

The results in Table 4 reveal a substantial gap in evidence quality between passive sampling and our active approach. (1) Superior Information Density:

Method	Avg. Score (1-5)	High-Sufficiency Rate
Uniform Sampling	2.88	39.5%
HAP (Ours)	**4.15**	**71.0%**

Table 4: *Oracle Evaluation of Visual Evidence Quality.* We compare the sufficiency of evidence provided by 16 uniform frames versus a single HAP-selected crop. The High-Sufficiency Rate denotes the percentage of samples receiving a score ≥ 4.

While the uniform baseline processes 16 frames to achieve a modest average score of 2.88, our HAP module achieves a significantly higher score of *4.15* using only a single, strategically selected crop. This indicates that HAP effectively filters out temporal redundancy and spatial noise, concentrating the visual signal into a highly informative region. (2) Decisive Evidence Retrieval: The High-Sufficiency Rate serves as a proxy for answering confidence. The dramatic improvement from 39.5% to *71.0%* demonstrates that standard downsampling often misses fine-grained details required for complex queries. In contrast, TIR-Flow's active zooming mechanism successfully locates and verifies decisive visual clues that are otherwise lost, validating the necessity of active perception. Furthermore, to demonstrate the practical advantages of our framework, a detailed efficiency analysis regarding inference latency and token overhead is provided in Appendix F.

6 Conclusion

In this work, we propose TIR-Flow, a novel framework that shifts the paradigm from passive processing to active video searching and reasoning without additional data or parameter updating. TIR-Flow empowers frozen VLMs to plan, look, and verify before answering, effectively shifting the focus from static parameter optimization to active perception. Extensive experiments on seven benchmarks demonstrate that TIR-Flow significantly outperforms recent strong baselines, delivering an average performance boost of 5.9%. By decoupling reasoning planning from visual perception, TIR-Flow frees the LLM backbone to function purely as a cognitive orchestrator, proving that the primary bottleneck in VLMs is often the high entropy of the input space rather than an intrinsic lack of reasoning capacity. Ultimately, our analysis confirms that empowering frozen VLMs with System-2-like active perception is a scalable path toward solving long-horizon video reasoning.

Limitations

This study has a few limitations. Firstly, due to resource constraints, our evaluation focuses on 7B-parameter models (LLaVA-Video and Qwen2.5-VL); investigating scalability on larger foundation models remains future work. Secondly, while proven highly effective on standard long-form benchmarks, handling extremely long, unedited continuous streams may require further optimization in context window management. Finally, while TIR-Flow significantly elevates reasoning ceilings, the iterative nature of System-2 active perception entails a modest trade-off in inference speed compared to static, single-pass baselines.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, and so on. 2025. Qwen3-vl technical report. Preprint, arXiv:2511.21631.

Maciej Besta, Nils Blach, Aleš Kubíček, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, H. Niewiadomski, P. Nyczyk, and Torsten Hoefler. 2023. Graph of thoughts: Solving elaborate problems with large language models. In AAAI Conference on Artificial Intelligence.

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. 2025. Scaling rl to long videos. Preprint, arXiv:2507.07966.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, and 1 others. 2025b. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. Trends in cognitive sciences, 7(10):454–459.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. 2024. Videoagent: A memory-augmented multimodal agent for video understanding. In European Conference on Computer Vision, pages 75–92. Springer.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. Preprint, arXiv:2503.21776.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. 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.

Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaojian Ma, Tao Yuan, Yue Fan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. 2025. Multi-modal agent tuning: Building a vlm-driven agent for efficient tool usage. Preprint, arXiv:2412.15606.

Google. 2024. Introducing gemini 1.5, Google's next-generation AI model. Accessed: 2024-06-10.

Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2025. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 26181–26191.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. Preprint, arXiv:2310.01798.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. arXiv preprint arXiv:2410.21276.

Jeongseok Hyun, Sukjun Hwang, Su Ho Han, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Joon-Young Lee, Seon Joo Kim, and Minho Shim. 2025. Multi-granular spatio-temporal token merging for training-free acceleration of video llms. Preprint, arXiv:2507.07990.

Hongbo Jin, Kuanwei Lin, Wenhao Zhang, Yichen Jin, and Ge Li. 2025a. Videocurl: Video curriculum reinforcement learning with orthogonal difficulty decomposition. arXiv preprint arXiv:2601.00887.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. 2025b. Videomem: Enhancing ultra-long video understanding via adaptive memory management. arXiv preprint arXiv:2512.04540.

Hongbo Jin, Rongpeng Zhu, Jiayu Ding, Wenhao Zhang, and Ge Li. 2026a. Himac: Hierarchical macro-micro learning for long-horizon llm agents. arXiv preprint arXiv:2603.00977.

Hongbo Jin, Rongpeng Zhu, Zhongjing Du, Xu Jiang, Jingqi Tian, Qiaoman Zhang, and Jiayu Ding. 2026b. Dgpo: Distribution guided policy optimization for fine grained credit assignment. arXiv preprint arXiv:2605.03327.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. 2025. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. Preprint, arXiv:2506.15745.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. Preprint, arXiv:2310.06452.

Ang Li, Charles Wang, Deqing Fu, Kaiyu Yue, Zikui Cai, Wang Bill Zhu, Ollie Liu, Peng Guo, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. 2025a. Zebra-cot: A dataset for interleaved vision language reasoning. Preprint, arXiv:2507.16746.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. Preprint, arXiv:2408.03326.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024b. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22195–22206.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025b. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024a. Video-llava: Learning united visual representation by alignment before projection. Preprint, arXiv:2311.10122.

Hao Lin, Kunyang Lv, Xu Jiang, Jingqi Tian, Zhongjing Du, Jiayu Ding, Qiaoman Zhang, and Hongbo Jin. 2026. Visd: Enhancing video reasoning via structured self-distillation. arXiv preprint arXiv:2605.06094.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024b. Vila: On pre-training for visual language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 26689–26699.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. In NeurIPS.

OpenAI. 2024. Hello GPT-4o.

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, and 5 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. In Advances in Neural Information Processing Systems.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. 2025. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. Preprint, arXiv:2504.07956.

Yiming Sun, Mi Zhang, Feifei Li, Geng Hong, and Min Yang. 2025. Smartsight: Mitigating hallucination in video-llms without compromising video understanding via temporal attention collapse. Preprint, arXiv:2512.18671.

Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. Preprint, arXiv:2303.08128.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. Preprint, arXiv:2506.13654.

Chenyu Wang, Weixin Luo, Sixun Dong, Xiaohua Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. 2025a. Mllm-tool: A multimodal large language model for tool agent learning. Preprint, arXiv:2401.10727.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, and so on. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. Preprint, arXiv:2508.18265.

Xiaodong Wang and Peixi Peng. 2025. Open-r1-video. https://github.com/Wang-Xiaodong1899/Open-R1-Video.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, William Wang, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025c. Multimodal chain-of-thought reasoning: A comprehensive survey. Preprint, arXiv:2503.12605.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025d. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 3272–3283.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025a. Llava-cot: Let vision language models reason step-by-step. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2087–2098.

Mingze Xu, Mingfei Gao, Shiyu Li, Jiasen Lu, Zhe Gan, Zhengfeng Lai, Meng Cao, Kai Kang, Yinfei Yang, and Afshin Dehghan. 2025b. Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding. Preprint, arXiv:2503.18943.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. Preprint, arXiv:2303.11381.

Fanglong Yao, Changyuan Tian, Jintao Liu, Zequn Zhang, Qing Liu, Li Jin, Shuchao Li, Xiaoyu Li, and Xian Sun. 2023a. Thinking like an expert: Multimodal hypergraph-of-thought (hot) reasoning to boost foundation modals. arXiv preprint arXiv:2308.06207.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023b. Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems, 36:11809–11822.

Huaying Yuan, Zheng Liu, Junjie Zhou, Hongjin Qian, Yan Shu, Nicu Sebe, Ji-Rong Wen, and Zhicheng Dou. 2025. Videoexplorer: Think with videos for agentic long-video understanding. Preprint, arXiv:2506.10821.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? Preprint, arXiv:2504.13837.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, and 1 others. 2025a. Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. arXiv preprint arXiv:2401.02582.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025b. LLaVA-mini: Efficient image and video large multimodal models with one vision token. In The Thirteenth International Conference on Learning Representations.

Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. 2025c. Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. Preprint, arXiv:2506.08817.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, and 1 others. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 8475–8489.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, and 1 others. 2025. Apollo: An exploration of video understanding in large multimodal models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 18891–18901.

A Failure Case Analysis

While TIR-Flow significantly elevates the reasoning ceiling of Video-LLMs by introducing an iterative Plan-Look-Verify mechanism, we observe performance boundaries on tasks that rely heavily on atomic visual perception rather than logical deduction. In this section, we analyze two representative failure cases from the VideoMME benchmark to delineate the distinction between reasoning failures and perception bottlenecks.

Dataset: *VideoMME*
Video: *24i4ncHuf6A*
Question: *According to the video, how many individuals were in the car when Archduke Franz Ferdinand was assassinated?*
Answer: *A. Three*
Candidates:
- *A. Three*
- *B. Two*
- *C. One*
- *D. Four*

Analysis: TIR-Flow successfully locates the correct temporal segment (Temporal Scouting). However, counting is a fundamental System-1 capability. Even with the correct keyframe and a zoomed-in view, the base VLM struggles with dense object detection and occlusion reasoning .

Insight: TIR-Flow acts as a cognitive amplifier; it can guide the model where to look, but it cannot rectify the model's intrinsic inability to recognize or count objects if the visual encoder's resolution or the LLM's numeracy is insufficient. The Planner cannot decompose Count objects into simpler text-based sub-steps effectively.

Dataset: *VideoMME*
Video: *LCtOpCi5r2s*
Question: *Which item was not featured in the video?*
Answer: *A. Three*
Candidates:
- *A. Balance scale*
- *B. Traffic light*
- *C. Gavel*
- *D. Magnifying glass*

Analysis: Unlike causal questions (e.g., "Why did the car stop?"), which have strong semantic cues to guide the Grid-Pyramid search (e.g., look for traffic lights), existence verification requires scanning the entire video for all candidates. The Planner attempts to verify the existence of each item. However, if an object (e.g., a magnifying glass) appears briefly in a cluttered background, the VLM's lack of fine-grained discriminative power leads to false negatives.

Insight: TIR-Flow is optimized for hypothesis-driven verification (confirming a specific clue exists), not exhaustive open-set scanning. When the semantic link between the query and the visual evidence is weak, the active search mechanism degenerates to the base model's passive scan, failing to compensate for the blind spots in low-level perception.

B Prompt Engineering Details

HDD Planner Prompt: Hypothesis Decomposition

```
You are a video reasoning planner.

Given a complex video question, your task
    is to decompose it into a set of
    independent, visually verifiable
    sub-questions.

For each sub-question:
1. Provide a concise textual question
    (q_text).
2. Provide a short visual query phrase
    (q_vis) suitable for visual matching.

Guidelines:
- Each sub-question should test a single
    hypothesis.
- q_vis must refer to observable entities
    or actions.
- Avoid abstract concepts such as
    intentions or motivations.
- Do NOT answer the questions.

Output format (JSON list):
[
  {"q_text": "...", "q_vis": "..."},
  ...
]

Original Question:
<QUESTION>
```

HDD Refinement Prompt

```
HDD Refinement Prompt:
You are a strategic planner refining a
    failed visual search.
The previous search for the sub-question
    failed.
Input:
1. Original sub-question:
   q_text: <q_text>
   q_vis: <q_vis>
2. Error Diagnosis (from Arbitrator):
   <error_type> (e.g., "Object Occlusion",
    "Temporal Mismatch", "Low Confidence",
    "Contradictory Evidence")

Your task is to regenerate a more granular
    sub-query to overcome this specific
    error.

Guidelines:
- If Error is "Object Occlusion/Blur":
    Switch to looking for associated
    static objects or surrounding context.
- If Error is "Temporal Mismatch": Look for
    temporal landmarks (e.g., "before X
    happens").
- If Error is "Contradiction": Break the
    question into two smaller verification
    steps.
- q_vis must be concrete and observable.

Output format (JSON):
{"q_text": "...", "q_vis": "..."}
```

EBA Prompt: Evidence Arbitration

```
You are an evidence arbitrator for video
    reasoning.

You are given:
1. A visual observation from the video.
2. A question about the observation.
3. A list of previously verified facts
    (Evidence Blackboard).

Your task:
- Decide whether the visual evidence
    supports the question.
- Produce a concise factual observation.
- Assign a confidence score between 0 and 1.

Rules:
- Base your judgment ONLY on the visual
    evidence.
- If the evidence is ambiguous, lower the
    confidence.
- If the observation contradicts existing
    evidence, mark a conflict.

Evidence Blackboard:
<BLACKBOARD>

Question:
<q_text>

Output format (JSON):
{
  "observation": "...",
  "confidence": 0.xx,
  "conflict": true/false
}
```

Final Prompt: Answer Synthesis

```
You are a video question answering
    assistant.

You are given:
1. The video frames.
2. The original question.
3. A set of verified visual facts extracted
    from the video.

Use the verified facts to answer the
    question.

Video Frames:
<VIDEO>

Original Question:
<QUESTION>

Verified Evidence:
<BLACKBOARD>

Only select the best option.
```

Oracle Evidence Quality Judge

System Instruction:
You are an expert visual evidence evaluator. Your task is to assess the quality and relevance of a specific image in the context of answering a user query about a video. You must strictly follow the scoring criteria provided.

User Input Template:
I will provide you with a Query and a Single Image Frame extracted from a video. Please evaluate the **Information Sufficiency** of this image for answering the query.
Query: `{q_txt}`
Image Context: `{Context_Note}` *(Inserted only for HAP: "Note: This image is a zoomed-in crop focusing on specific fine-grained details.")*
Scoring Criteria (1-5):

- **5 (Critical Evidence):** The image contains the exact visual information needed. No guessing required.
- **4 (Strong Evidence):** Provides strong cues. Answer can be inferred with high confidence.
- **3 (Partial/Related):** Relevant to the topic/scene, but specific key details are missing, blurry, or occluded.
- **2 (Weak Relevance):** Shows general setting but not helpful for the specific question.
- **1 (Irrelevant/Noise):** Completely unrelated, blurry beyond recognition, or misses the event.

C Baseline Details

Baseline Setting. We compare TIR-Flow against a wide range of state-of-the-art methods: Proprietary Models: GPT-4o (Hurst et al., 2024), Gemini-1.5-Flash (Google, 2024), Gemini 1.5 Pro (Google, 2024) and Gemini 2.5 pro (Comanici et al., 2025b) Open-source Models: LLaVA-Mini-8B (Zhang et al., 2025b), VILA-40B (Lin et al., 2024b), LLAVA-onevision-7B (Li et al., 2024a), VideoLLaMA3-7B (Zhang et al., 2025a), Video-R1-7B (Feng et al., 2025), Qwen3-VL-8B (Bai et al., 2025). Training-Free Frameworks: SmartSight (Sun et al., 2025) (built on Video-R1-7B). VideoTree (Wang et al., 2025d) (built on GPT-4): A more computationally intensive baseline, utilizing an average of 62.4 frames for EgoSchema and 53.1 frames for Video-MME. InfiniPot-V (Kim et al., 2025) (built on LLaVA-Next-7B). STTM (Hyun et al., 2025) (built on LLaVA-Video 7B).

D Data

To comprehensively evaluate the effectiveness of TIR-Flow across different cognitive dimensions, we adopt seven representative multiple-choice video benchmarks. We categorize these benchmarks into two groups based on their primary challenges: Perception Benchmarks: We use MVBench (Li et al., 2024c), Perception-Test (Pătrăucean et al., 2023), and Video-MME (Fu et al., 2025). These benchmarks demand high-fidelity spatial-temporal recognition, testing the model's ability to discern fine-grained actions, object attributes, and long-range temporal consistency. Reasoning Benchmarks: We evaluate on EgoSchema (Mangalam et al., 2023), VideoEspresso (Han et al., 2025), VideoMMMU (Hu et al., 2025), and MMVU (Zhao et al., 2025). These datasets require advanced cognitive abilities beyond simple recognition, such as causal induction, intentionality analysis, and the integration of complex multi-modal context.

E Sensitivity Analysis

Threshold	EgoSchema *(LLaVA / Qwen)*	VideoMME *(LLaVA / Qwen)*
$\tau = 0.5$	66.3 / 67.5	66.9 / 66.4
$\tau = 0.6$	67.2 / 68.1	68.0 / 67.5
$\tau = 0.7$	**67.8 / 69.2**	**68.9 / 68.3**
$\tau = 0.8$	67.0 / 68.5	68.1 / 67.5
$\tau = 0.9$	66.2 / 67.4	67.3 / 66.7

Table 5: **Impact of Confidence Threshold τ.** Analysis on EgoSchema and VideoMME.

Impact of Confidence Threshold τ. We further investigate the sensitivity of the Evidence Blackboard & Arbitration (EBA) module to the confidence threshold τ. Table 5 demonstrates the accuracy curves across different benchmarks as τ varies from 0.5 to 0.9. We observe that an overly low τ introduces noise into the reasoning chain (false positives), while an excessively high τ leads to conservative predictions and unnecessary backtracking. Empirically, the model achieves robust performance peaks across all datasets when τ is set to approximately **0.7**.

F Efficiency Analysis

To further demonstrate TIR-Flow's practical advantages, we compare its computational efficiency against representative agent frameworks. TIR-Flow establishes a superior efficiency-performance trade-off through three key aspects:

Table 6: Efficiency Analysis and Comparison with Agent-based Methods. We report Average Latency (s) and Token consumption (k) per query alongside Accuracy (%). "Baseline" refers to the standard, single-pass Qwen2.5-VL-7B model without active perception.

Dataset	Method	Latency (s)	Tokens (k)	Accuracy (%)
LVBench	Baseline	18.7	8.2	45.3
	VideoAgent (Fan et al., 2024)	–	266.9	34.9
	VideoExplorer (Yuan et al., 2025)	–	19.6	50.6
	TIR-Flow (Ours)	**60.2**	**12.8**	**51.8**
	Ego-R1 (Tian et al., 2025)	–	532.6	39.6
	GPT-4o	–	23.0	48.9
MLVU	Baseline	19.4	8.2	68.2
	VideoAgent (Fan et al., 2024)	–	61.5	38.6
	VideoExplorer (Yuan et al., 2025)	–	20.5	55.4
	TIR-Flow (Ours)	**62.8**	**13.6**	**71.9**
	Ego-R1 (Tian et al., 2025)	–	122.2	40.9
	GPT-4o	–	98.3	54.9

Exceptional Token Efficiency: TIR-Flow operates on a minimal token budget (averaging 13k) while significantly outperforming heavy agents like Ego-R1 (which consumes up to 532k tokens). By aggressively filtering temporal redundancy via active perception, our framework successfully mitigates attention dilution in LLMs.

Robustness vs. Tool-Use: Unlike traditional agents that suffer from error propagation across external tool pipelines (e.g., VideoExplorer), TIR-Flow leverages the VLM's native encoder for active search. This unified approach yields substantial accuracy gains, such as a +16.5% boost over VideoExplorer on MLVU.

Controllable Latency: While the active searching mechanism introduces moderate latency (60s) compared to single-pass baselines, it avoids the protracted and unpredictable execution times of multi-round agent loops, making TIR-Flow a highly practical controller for long-horizon reasoning.