VideoMind: Thinking in Steps for Long Video Understanding

Shubhang Bhatnagar[1,*], Renxiong Wang[2], Kapil Krishnakumar[2],
Adel Ahmadyan[2], Zhaojiang Lin[2], Lambert Mathias[2], Xin Luna Dong[2],
Babak Damavandi[2], Narendra Ahuja[1], Seungwhan Moon[2]

1 : University of Illinois Urbana Champaign
2 : Meta Reality Labs
* : Work done as an intern at Meta Reality Labs

EACL 2026
MOROCCO
Palais Des Congres, Rabat
March • Mars • 2026 • مارس

The Long Video Bottleneck

The LVU Challenge: E.g., *"How many goals did Player #10 score in a 90-minute match?"*

Input Data:

- Typical long-form videos span tens of minutes to hours,
- Containing lots of redundant information interspersed with sparse, salient events.

Context Limits: Uniformly sampling these videos to fit into an MLLM generates tens of thousands of tokens, pushing models beyond their architectural limits due to quadratic attention complexity.

The Result: The model simply cannot attend to the details that matter, or the salient info gets completely lost during sampling

Current Approaches for Long Video Understanding

- **Efficient Attention:** Modifying MLLM architecture to manage very long contexts.

- **Token Compression & Memory:** Shrinking video's footprint while attempting to preserve salient content

- **Agentic Tool Calling:** Using an orchestrator to call specialized external tools (e.g., API models) to caption, retrieve, and analyze important sub-parts of the video.

Our Work: Can a base MLLM achieve better performance without relying on external modules ?

VideoMind uses *self-specialization* to actively parse through the video with different goals and reasoning specialties at different times, just as a human would

VideoMind: Method

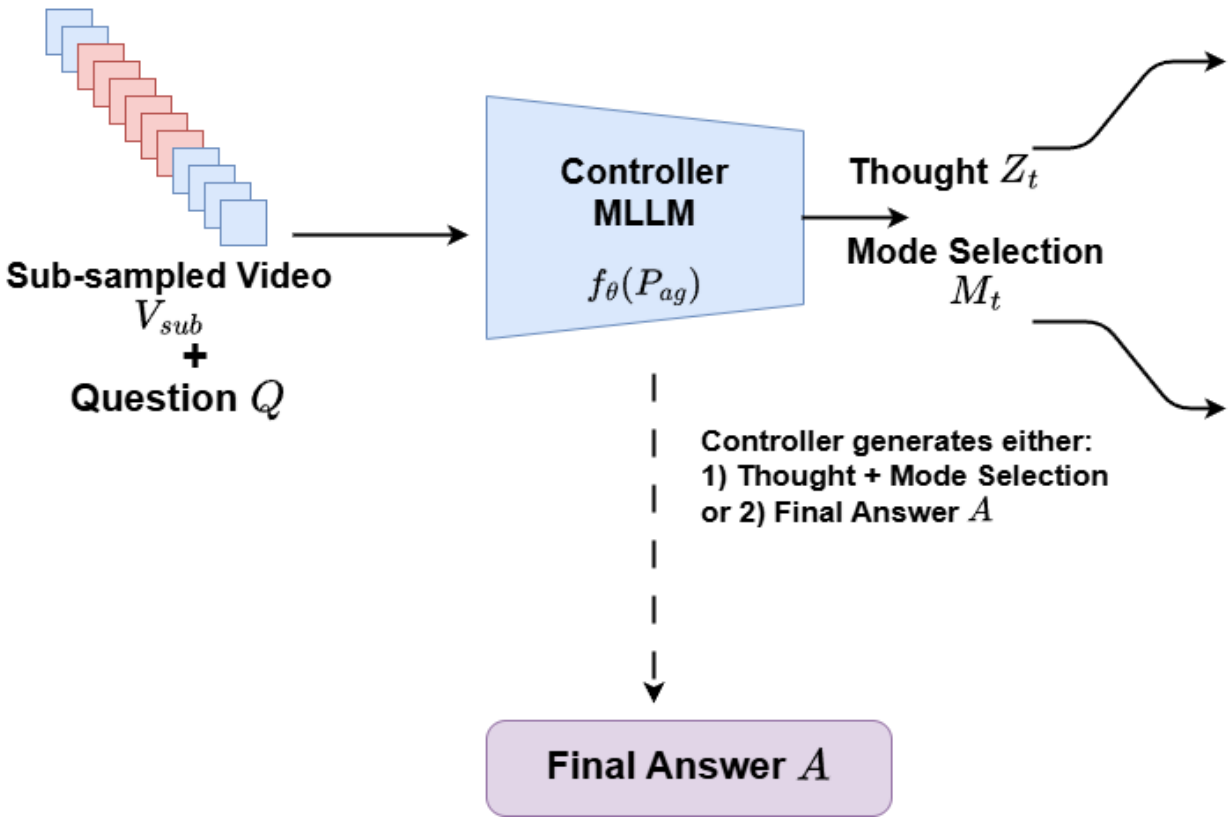
- **Training-Free & Zero-Shot:** A completely training-free, lightweight framework adaptable to off-the-shelf MLLMs.

- **Query Decomposition:** Decomposes a given complex user query into a sequence of manageable, actionable sub-queries.

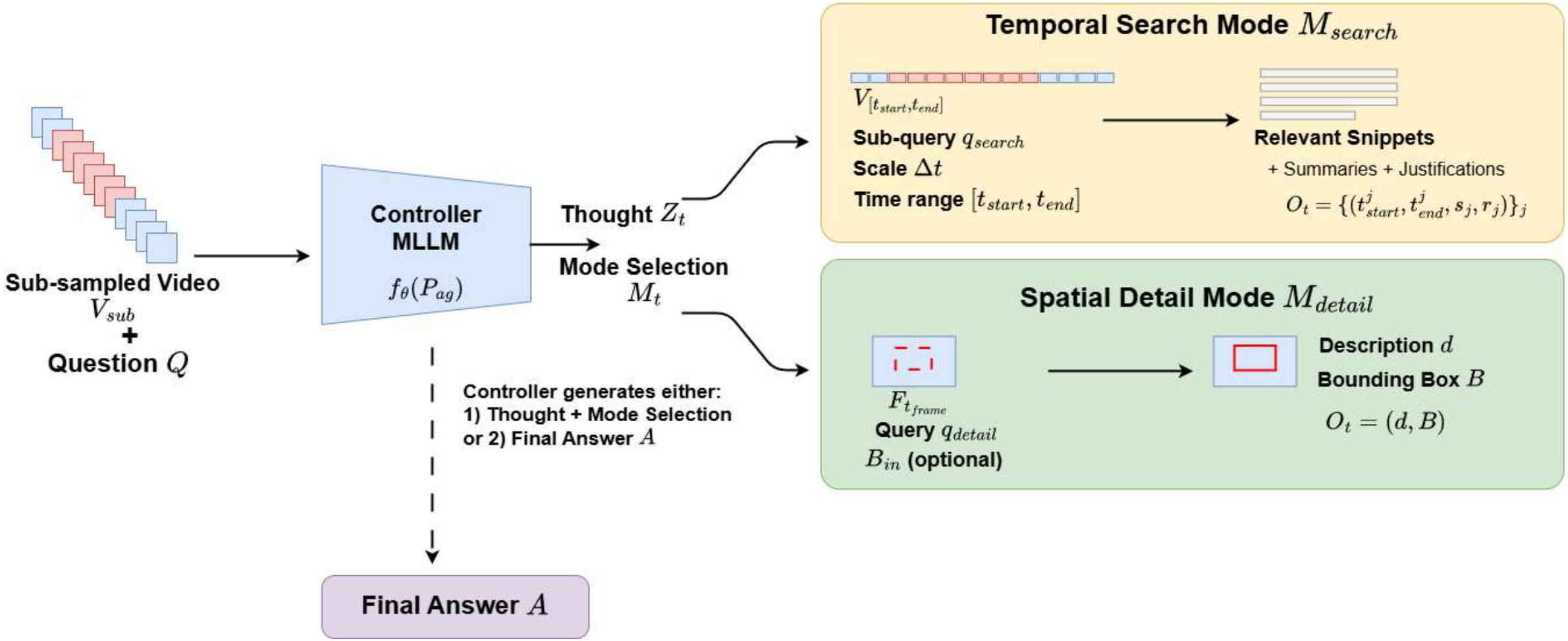
- **Intrinsic Reasoning Modes:** To oslve each sub query, model changes into appropriate mode, using ‘Mode’ prompts that unlock MLLM’s latent capabilities.

- **Tailored Context:** Modes receive minimal, task-specific video context (e.g., a few frames) allowing for focused, fine-grained analysis.

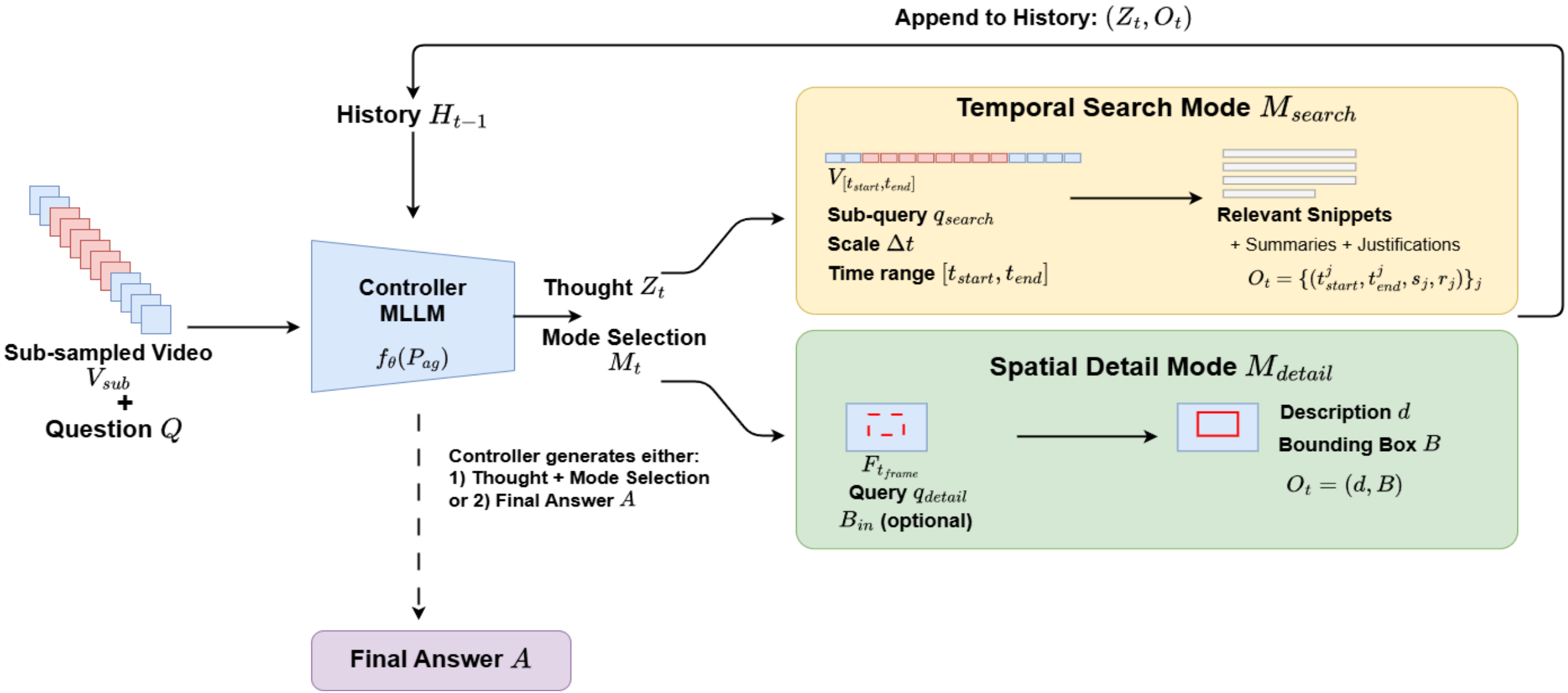
Method: Architecture Overview



Method: Architecture Overview



Method: Architecture Overview

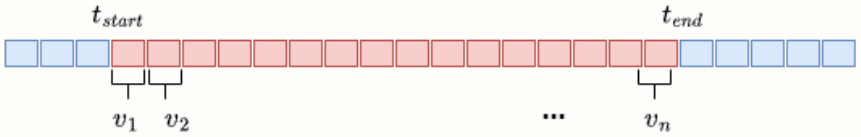


- A novel training-free framework designed to mimic a human reasoning process.
- Actively interrogates the video rather than passively consuming frames at once.
- Decomposes a complex user query into a sequence of focused, actionable sub-queries

Method: Mode 1

Multi-Granular Temporal Search Mode M_{search}

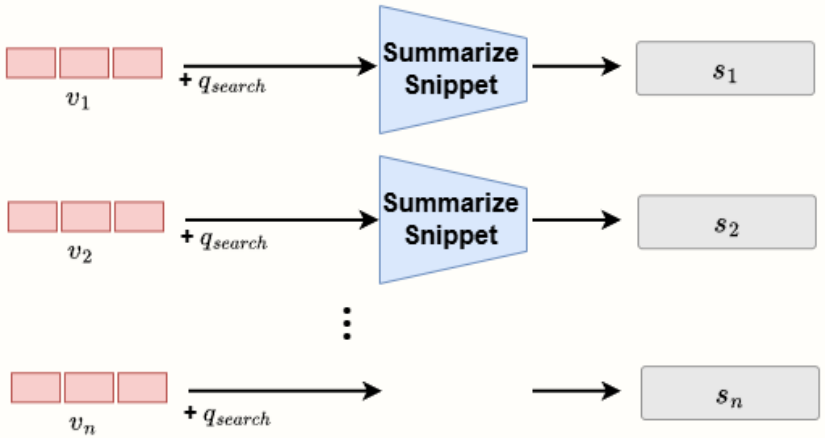
Video Segment $V_{[t_{start}, t_{end}]}$



+ Query q_{search}

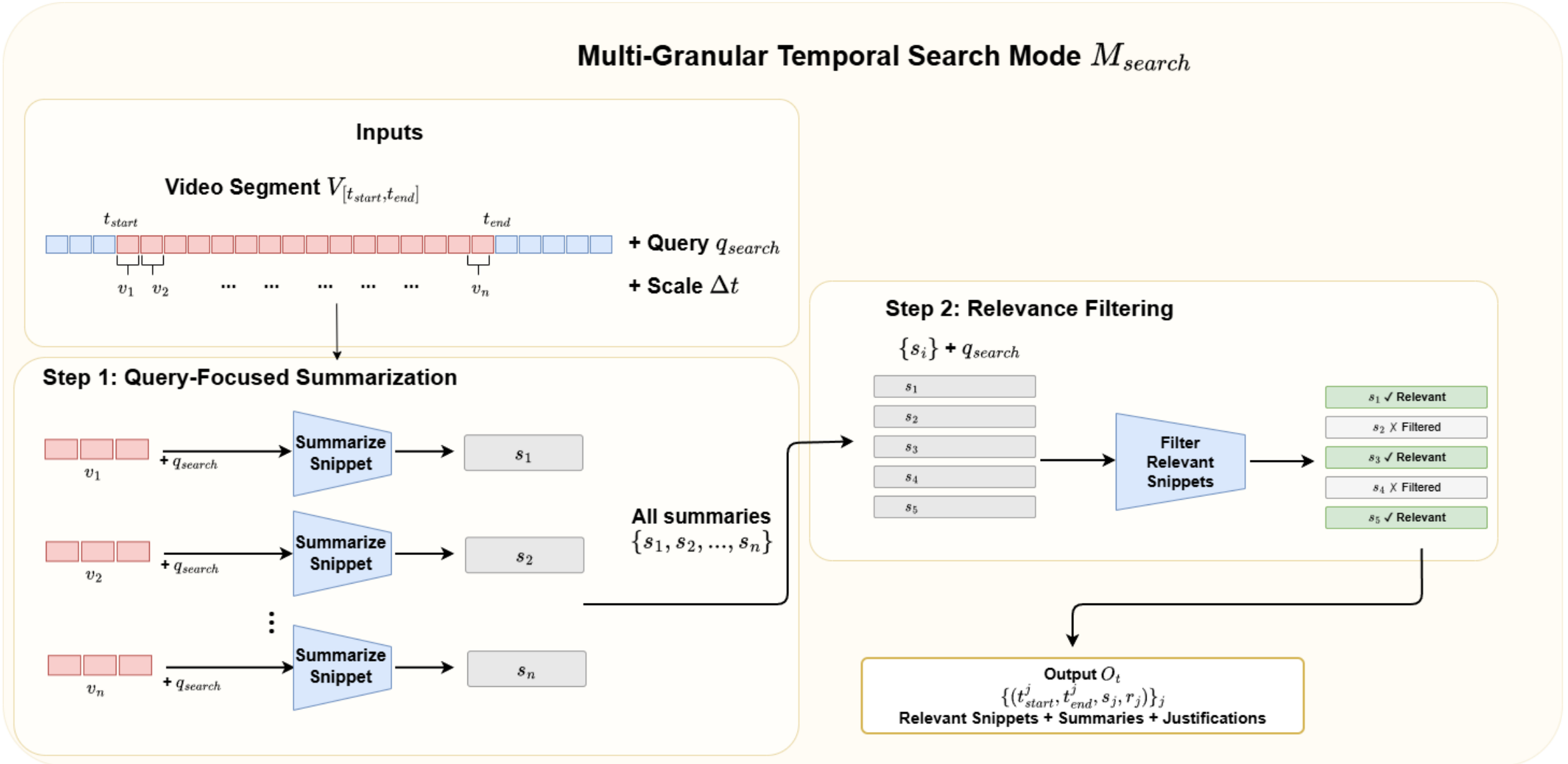
+ Scale Δt

Step 1: Query-Focused Summarization

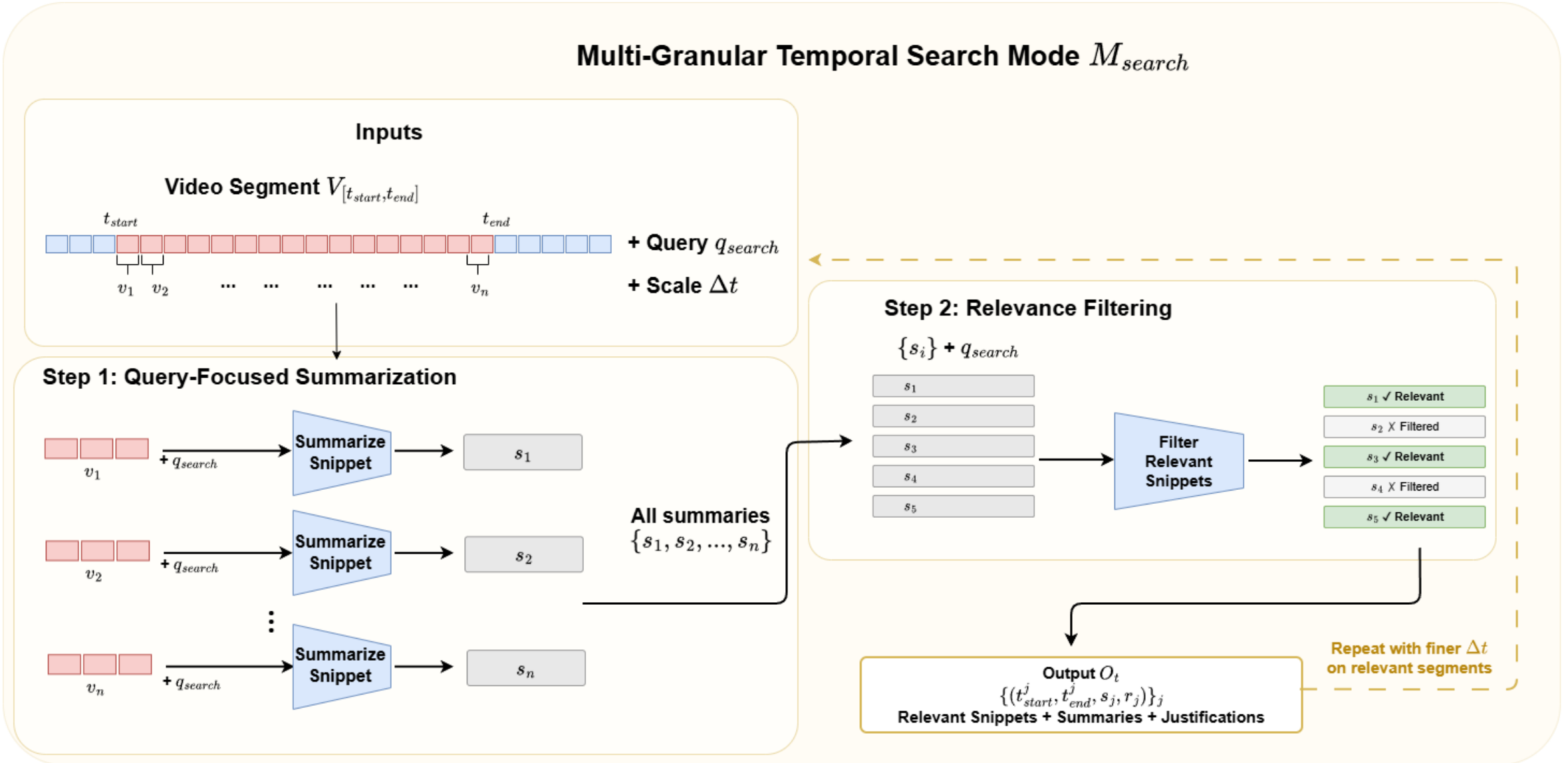


All summaries
$\{s_1, s_2, ..., s_n\}$

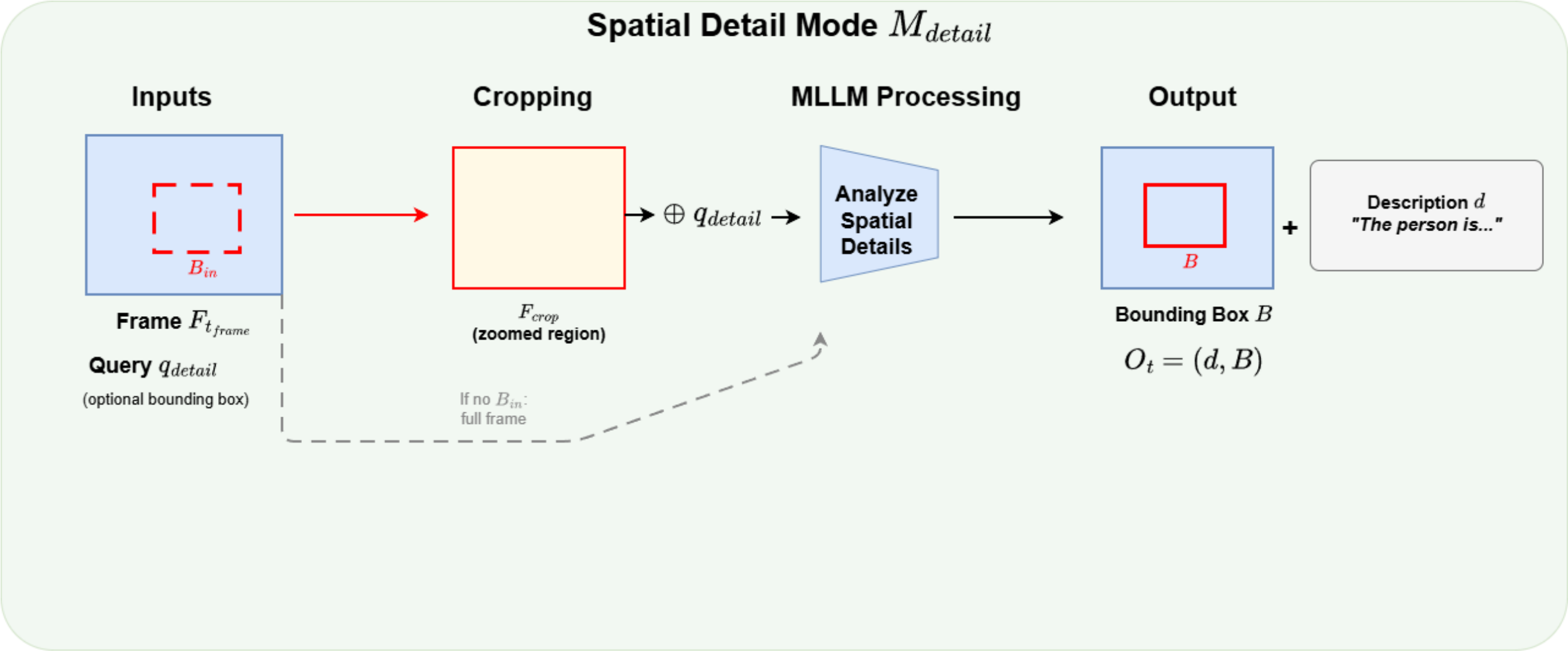
Method: Mode 1



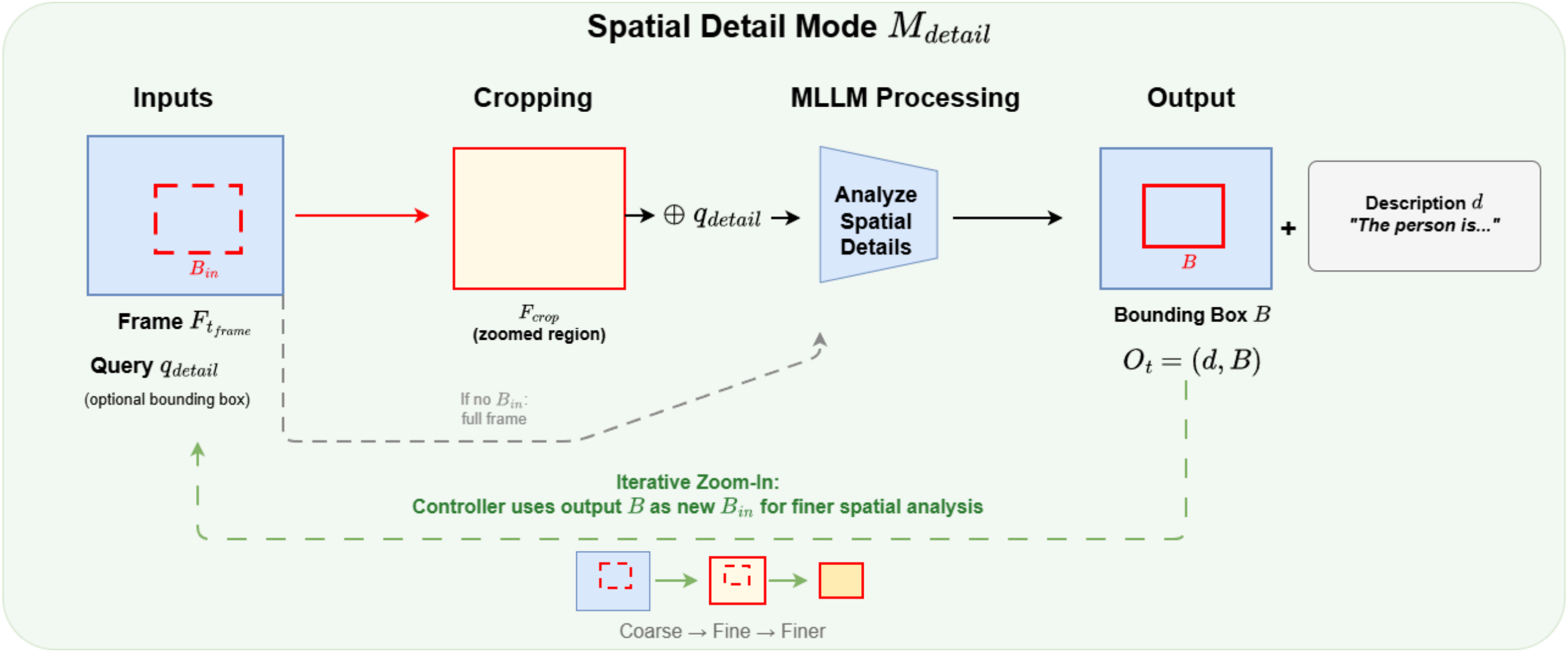
Method: Mode 1



Method: Mode 2



Method: Mode 2



Experimental Setup

Benchmarks:

- Video-MME,
- Long Video Bench,
- MLVU

Backbones:

- Qwen 2.5 72B
- Llama 4 Scout (17B).

Baselines: Fed 768 frames following previous work

VideoMind:

1. MLLM is limited to engaging in modes a maximum of 20 times per question.
2. 5 Time scale options are used Multi granular temporal search

Results

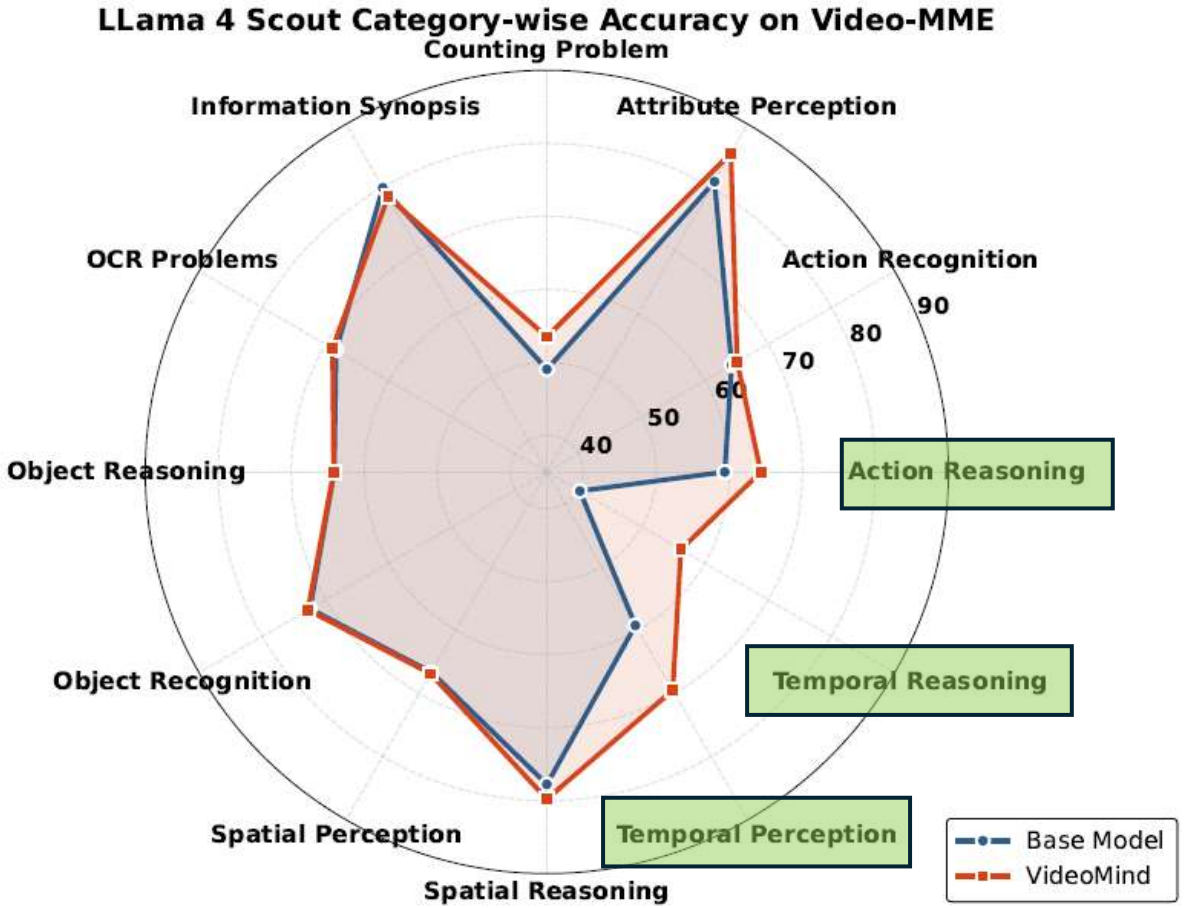
Consistent gains across all three benchmarks

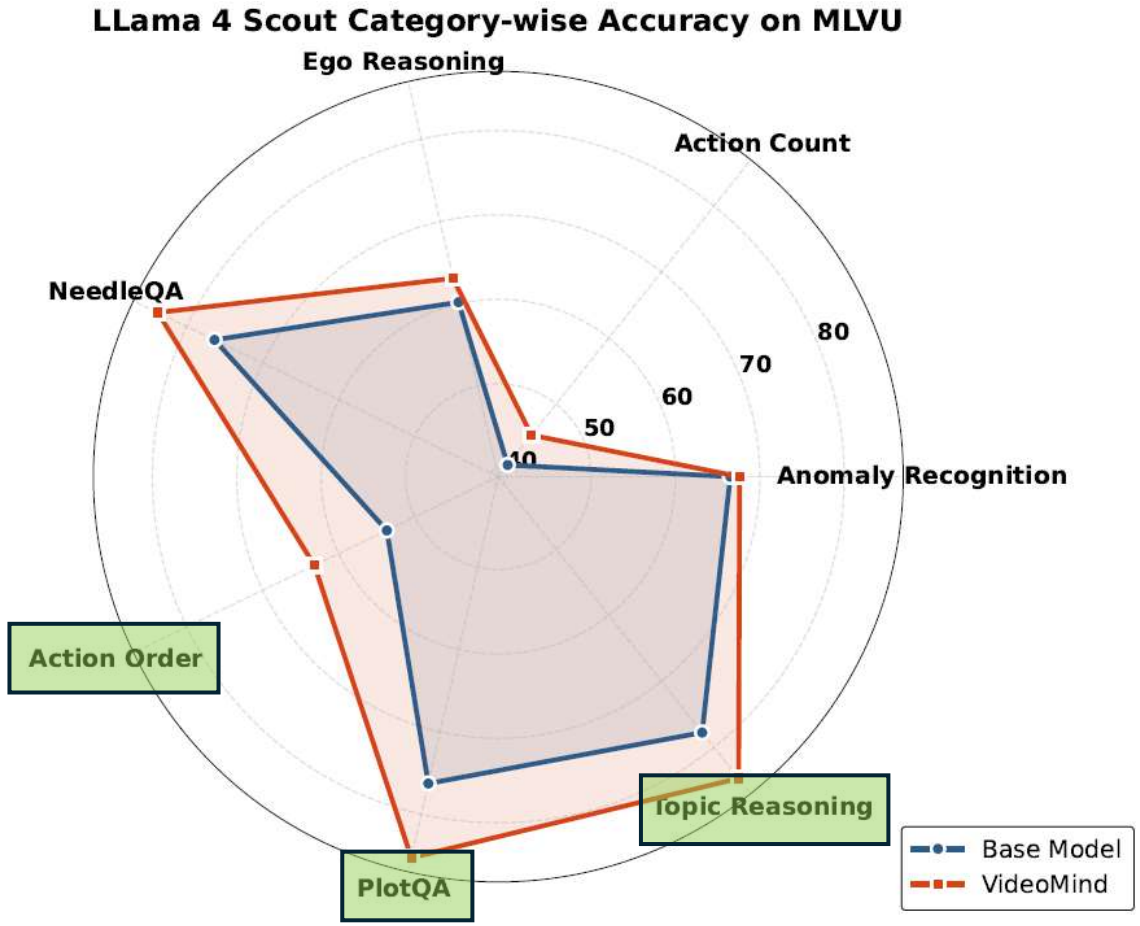
Method	Params	LongVideo Bench	VideoMME				MLVU
		Overall	Overall	Short	Medium	Long	Overall
GPT-4o (OpenAI, 2023)	-	66.7	71.9	-	-	65.3	64.6
Gemini-1.5-Pro (Gemini et al., 2024)	-	64.0	75.0	-	-	67.4	64.0
InternVL2.5 72B (Chen et al., 2024)	78B	63.6	72.1	-	-	62.6	75.7
LLaVA-OneVision (Li et al., 2024a)	72B	61.3	66.2	-	-	-	68.0
Qwen2.5-VL+AdaReTaKe (Ma et al., 2024)	72B	67.0	73.5	-	-	65.0	78.1
Base Model (Llama 4 Scout)	17B	49.5	62.8	76.5	67.1	54.2	67.4
VideoMind (Llama 4 Scout)	17B	53.1	67.8	79.7	76.7	59.5	73.6
Base Model (Qwen-2.5-VL 72B)	72B	60.5	72.8	**82.1**	70.1	60.2	74.6
VideoMind (Qwen-2.5-VL 72B)	72B	63.1	77.6	81.5	77.8	64.5	77.2

Provides Major Performance Boosts over the base model, e.g.:

- VideoMind with Qwen 2.5 achieves SoTA performance of 77.6% on Video-MME, boosting accuracy by 4.8%
- Similarly Improves Llama 4 Scout (17B) accuracy by 5.0% on VideoMME to 67.8%,

Category Wise Results





Biggest Improvements on tasks requiring complex temporal understanding

Results: Ablation

Contribution of Individual Modes:

- Temporal Search captures majority of performance gains
- Spatial Detail adds a smaller improvement

Method	VideoMME		LongVideoBench
	Long	Overall	Overall
Base Model	54.2	62.8	49.5
+ M_{detail} only	54.2	64.5	50.1
+ M_{search} only	58.7	66.9	52.4
VideoMind	**59.5**	**67.8**	**53.1**

Design of Spatial Detail Mode:

- Using External Detectors in place of the spatial detail mode lowers performance by a lot
- Removing crop-and-zoom significantly degrades fine-grained spatial recognition

Method	Video MME	LongVideo Overall	LongVideo S2A
VideoMind	**67.8**	**53.5**	**72.5**
w/ Grounding DINO	56.1	44.6	56.1
w/o Crop	66.4	52.2	64.3

Conclusion

VideoMind:

- A training-free framework for LVU mimicking human reasoning
- Overcomes context limits via self-specialization modes of the base MLLM without retraining or external tools.
- Achieves SoTA benchmark performance (e.g., 77.6% on Video-MME) using these modes

Limitations: Performance is strictly bound by the base model's instruction-following fidelity (smaller 7B models fail), and iterative steps add latency.

Future Work: Exploring RL fine-tuning to optimize the agent's planning and mode-selection strategy.