See What You Need: Query-Aware Visual Intelligence through Reasoning-Perception Loops

Anonymous ACL submission

Abstract

Human video comprehension demonstrates dynamic coordination between reasoning and visual attention, adaptively focusing on query-relevant details. However, current long-form video question answering systems employ rigid pipelines that decouple reasoning from perception, leading to either information loss through premature visual abstraction or computational inefficiency through exhaustive processing. The core limitation lies in the inability to adapt visual extraction to specific reasoning requirementsdifferent queries demand fundamentally different visual evidence from the same video content. In this work, we present CAVIA, a training-free framework that revolutionizes video understanding through reasoning-perception coordination. Unlike conventional approaches where visual processing operates independently of reasoning, CAVIA creates a closed-loop system where reasoning continuously guides visual extraction based on identified information gaps. CAVIA introduces three innovations: (1) hierarchical reasoning-guided localization to precise frames; (2) cross-modal semantic bridging for targeted extraction; (3) confidence-driven iterative synthesis. CAVIA achieves state-of-the-art performance on challenging benchmarks: EgoSchema (65.7%, +5.3%), NExT-QA (76.1%, +2.6%), and IntentQA (73.8%, +6.9%), demonstrating that dynamic reasoning-perception coordination provides a scalable paradigm for video understanding.

1 Introduction

Video question answering (VideoQA) entails the comprehension of intricate temporal dynamics, spatial relationships, and semantic content within video sequences (Xiao et al., 2022). While substantial advancements have been made in analyzing short video clips, long-form video understanding presents unique challenges (Wu et al., 2024):
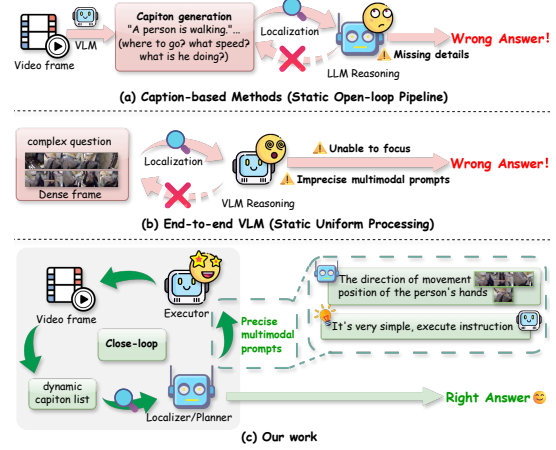

Figure 1: CAVIA addresses the limitations of static caption-based methods and unfocused VLMs through closed-loop coordination, dynamically generating precise multimodal prompts to extract query-specific spatiotemporal details for accurate video understanding.

the sparse distribution of query-relevant information across extensive sequences and the computational burden of processing lengthy content while preserving fine-grained details (Song et al., 2024; Li et al., 2023b). Given a natural language query and an extended video, systems must efficiently identify pertinent temporal segments while maintaining a holistic understanding of the video content (Fei et al., 2024).

Current methods achieve significant progress in addressing this challenge, typically through two predominant paradigms. Caption-based approaches (Jianqiao et al., 2025) convert visual information into textual descriptions, leverage the powerful reasoning capabilities of language models for answer generation. Conversely, end-to-end vision-language models (Youk et al., 2024) achieve comprehensive visual understanding by retaining and processing raw visual data uniformly. However, both paradigms exhibit a critical limitation stemming from the decoupling of reason-

ing and perception (Wang et al., 2024b; Ko et al., 2023a): perception passively feeds information to reasoning modules without adaptive feedback, preventing cognition from guiding perception. Consequently, captioning methods inevitably suffer from information loss due to premature abstraction of visual details, while end-to-end models incur computational overload from exhaustive, indiscriminate processing of all visual content (Shen et al., 2023).

Human cognition demonstrates a dynamic interplay between reasoning and visual attention, where perceptual focus continuously adapts based on evolving reasoning states (Santos et al., 2025), revealing a radically different architecture where reasoning and perception exist in constant dialogue, mutually informing and reshaping each other. Consider two distinct queries: answering "what tool was used?" demands detailed spatial analysis of objects and their interactions, while determining "why did the person stop?" necessitates understanding temporal sequences and broader contextual cues. This synergistic guidance of perception by higher-level cognition starkly contrasts with existing computational methods' fundamental limitation: the decoupling of reasoning and perception. Figure 1 illustrates these limitations and motivates our closed-loop approach.

Inspired by the profound interplay of reasoning and perception inherent to human cognition, we present **CAVIA** (**C**losed-loop **A**daptive **V**ideo **I**ntelligence **A**gent), which fundamentally rethinks video understanding by establishing dynamic coordination between reasoning and perception. Our key insight: effective comprehension requires neither exhaustive processing nor static captions, but adaptive mechanisms where reasoning continuously guides visual extraction—mirroring human cognitive processes. Unlike prior iterative approaches that perform simple caption corrections (Zhang et al., 2025), CAVIA introduces a sophisticated feedback loop that dynamically enriches visual descriptions based on identified reasoning gaps. CAVIA introduces three technical innovations:

- **Hierarchical Reasoning-Guided Localization**: Progressively narrows from caption clusters to precise frames through LLM-guided analysis.
- **Cross-Modal Semantic Bridging**: Translates reasoning gaps into targeted visual extraction directives for spatial-temporal analysis.
- **Confidence-Driven Iterative Synthesis**: Orchestrates progressive refinement through reasoning-perception cycles, with convergence determined by confidence metrics.

By coordinating pretrained VLMs and LLMs in a training-free framework, CAVIA achieves state-of-the-art performance: EgoSchema (65.7%, +5.3%), NExT-QA (76.1%, +2.6%), and IntentQA (73.8%, +6.9%). These results validate dynamic reasoning-perception coordination as a scalable paradigm for long-form video understanding

2 Related Work

2.1 Long-form Video Question Answering

Early VideoQA methods with CNN-RNN architectures faced context window limitations (Sharma and Jalal, 2022). Transformer-based approaches improved temporal modeling (Chen et al., 2024) but suffer from quadratic complexity. Recent works address efficiency through hierarchical processing: OptiGQA (Wang et al., 2025b) and VideoTree (Wang et al., 2025c) use multi-level structures, while MIST (Gao et al., 2023) decomposes spatial-temporal attention into cascaded selection modules. For extremely long videos, (Nguyen et al., 2024) integrates State Space Layers for global semantic encoding. However, these *static pipelines* process all queries uniformly regardless of complexity.

2.2 Multimodal Architectures for Video Understanding

Video understanding systems coordinate LLMs and VLMs through various architectures (Ma et al., 2024; Jin et al., 2024). Recent efficiency improvements include sparse memory (MovieChat (Song et al., 2024), VideoINSTA (Liao et al., 2024)), streaming processing (Flash-VStream (Zhang et al., 2024b)), and memory-based reasoning (Glance-Focus (Bai et al., 2023)). SeViLA (Yu et al., 2023a) leverages image-language models with dual-module chaining for keyframe localization. While DrVideo (Ma et al., 2025) introduces iterative refinement, it lacks *fine-grained temporal localization* and *targeted multimodal prompting*, operating at coarse segments without precise spatial-temporal instructions.
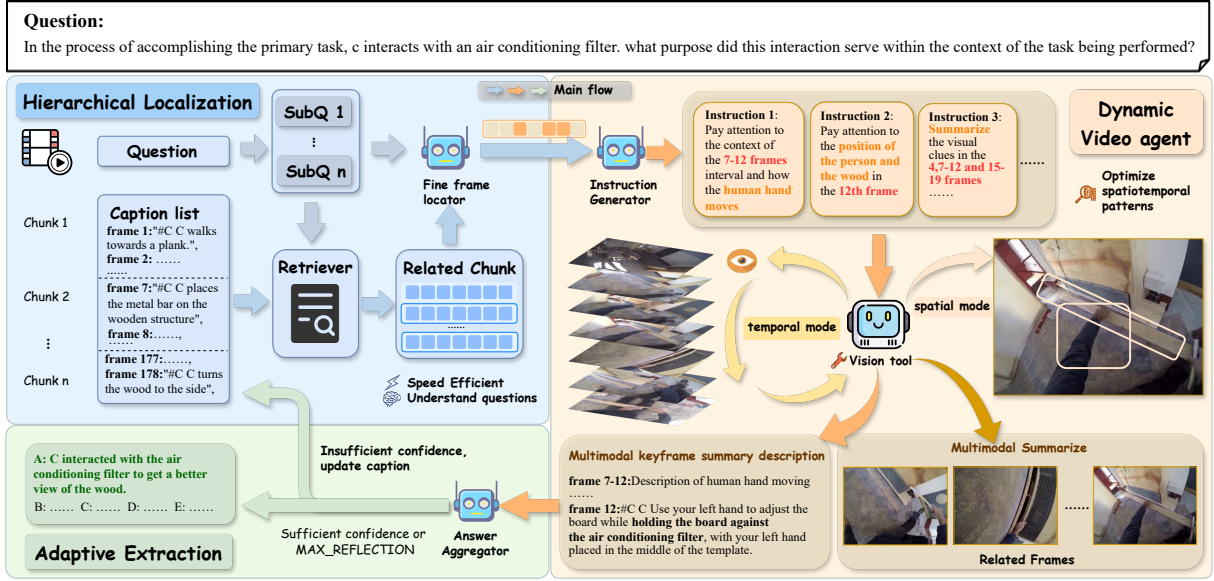
Figure 2: The CAVIA framework architecture. The system operates through closed-loop coordination between three components: Hierarchical Localization (blue) progressively narrows from caption chunks to specific frames via retrieval and LLM analysis; Dynamic Prompting (orange) generates targeted spatial-temporal instructions based on identified information gaps; Adaptive Extraction (green) employs VLMs in temporal or spatial mode to extract query-specific visual details. The iterative process continues until sufficient confidence is achieved.

2.3 Adaptive Reasoning and Cross-Modal Coordination

Adaptive frameworks implement query-specific processing through predefined selection (VideoAgent (Wang et al., 2025a)), text-visual alignment (LeAdQA (Dong et al., 2025)), or LLM-based temporal reasoning (LLaMA-VQA (Ko et al., 2023b)). However, these approaches operate via *open-loop reasoning* once visual features are extracted, models cannot request targeted information to address reasoning gaps, precluding iterative refinement crucial for complex video understanding.

3 Method

We present CAVIA, a closed-loop framework for long-form video question answering that establishes dynamic coordination between reasoning and perception. Unlike static pipelines that process videos uniformly, CAVIA introduces query-aware mechanisms where reasoning continuously guides visual extraction.

3.1 Problem Formulation and Framework Overview

Given a video V with uniformly sampled frames $\{f_1, f_2, \ldots, f_T\}$ and corresponding captions $\{c_1, c_2, \ldots, c_T\}$, we answer question q by dynamically identifying and extracting query-specific visual details rather than exhaustively processing all information.

As illustrated in Figure 2, CAVIA operates through four interconnected components forming a closed-loop system: (1) **Coarse-to-Fine Temporal Localization** that hierarchically narrows from caption blocks to specific frames, (2) **Reasoning Gap Identification** where LLMs detect missing information in current context, (3) **Targeted Multimodal Prompting** that generates precise spatial-temporal instructions for VLMs, and (4) **Iterative Caption Enhancement** that progressively enriches descriptions based on extracted details. This closed-loop architecture enables reasoning to actively guide perception throughout the process.

3.2 Coarse-to-Fine Temporal Localization

We introduce a coarse-to-fine strategy that adapts processing granularity to query complexity. For the initial localization, we partition the caption sequence into semantically coherent blocks and employ a retrieval mechanism that considers both the original query and decomposed sub-questions. This multi-granular matching captures both holistic relevance and specific detail alignment:

$$\mathcal{C}_{\text{rel}} = \arg\max_C \alpha S(q, C) + \beta \sum_{q_i \in D(q)} S(q_i, C)$$

where $S(q, C)$ measures semantic similarity between query q and caption block C, $D(q)$ represents decomposed sub-questions, and α, β are weighting parameters. This formulation captures both holistic relevance and specific detail alignment. Within selected caption blocks, our Fine Frame Locator employs LLM analysis to identify frames containing critical visual information. Unlike methods that uniformly sample frames, this query-driven selection ensures computational resources concentrate on discriminative content. The LLM evaluates each caption within relevant blocks, identifying those that potentially contain answer-critical information but lack sufficient detail.

3.3 Targeted Multimodal Prompting

The core innovation of CAVIA lies in our iterative enhancement mechanism that dynamically enriches visual understanding based on reasoning requirements. When the LLM detects information gaps in the current context, it generates targeted instructions that guide the VLM to extract specific missing details. The visual tool is implemented using Qwen2.5-VL-7B[1](Bai et al., 2025), which provides strong visual-language understanding capabilities for both temporal and spatial analysis.

Our Instruction Generator operates in two complementary modes based on the nature of identified gaps:

Temporal Enhancement Mode: When reasoning requires understanding of actions, movements, or temporal transitions, the system selects coherent frame sequences that capture motion dynamics. Unlike static frame sampling, we adaptively determine the temporal span based on action complexity:

$$I_{\text{temp}} = \arg\max_{t_s, t_e} \text{AC}(f_{t_s:t_e}) \cdot \text{QR}(q, f_{t_s:t_e})$$

where $\text{AC}(f_{t_s:t_e})$ measures action coherence across frames t_s to t_e, and $\text{QR}(q, f_{t_s:t_e})$ evaluates query relevance. This ensures we capture complete action sequences rather than isolated frames. The VLM then processes these intervals with temporal-aware prompts, generating descriptions that capture motion patterns, action sequences, and temporal relationships absent from initial captions.

Spatial Enhancement Mode: For queries requiring fine-grained visual details about objects, scenes, or spatial relationships, prompts such as "Focus on the tool in the person's right hand" guide VLMs to extract fine-grained visual attributes from specific regions, ensuring focused analysis rather than generic description.

Algorithm 1 details this instruction-driven visual processing mechanism, showing how targeted prompts are generated based on identified reasoning gaps. This targeted enhancement mechanism represents a fundamental departure from existing approaches. Rather than relying on pre-computed visual features or exhaustive frame processing, we dynamically generate visual queries that address specific reasoning gaps, ensuring efficient yet comprehensive understanding.

Algorithm 1 Instruction-Driven Visual Processing

Require: Question q, relevant caption blocks $\mathcal{C}_{\text{rel}}$, video frames $\mathcal{F}$, frame-caption mapping $\mathcal{M}$
Ensure: Refined context $\mathcal{C}_{\text{ref}}$, final answer $\hat{y}$
1: $\mathcal{C}_{\text{ref}} \leftarrow \emptyset, \mathcal{I} \leftarrow \text{GENINST}(q, \mathcal{C}_{\text{rel}})$
2: **for** each instruction $i \in \mathcal{I}$ **do**
3:     **if** ISTEMP(i) **then**
4:         $I_t \leftarrow \text{SELTIME}(\mathcal{F}, i)$
5:         $V_t \leftarrow \text{FEATTIME}(I_t)$
6:         $\mathcal{C}_{\text{ref}} \leftarrow \mathcal{C}_{\text{ref}} \cup \text{DESCSEQ}(V_t, i)$
7:     **else if** ISSPAT(i) **then**
8:         $f \leftarrow \text{LOCFRAME}(\mathcal{F}, i)$
9:         $V_s \leftarrow \text{FEATOBJ}(f, i)$
10:        $\mathcal{C}_{\text{ref}} \leftarrow \mathcal{C}_{\text{ref}} \cup \text{DESCREG}(V_s, i)$
11:    **end if**
12: **end for**
13: $\mathcal{D} \leftarrow \text{SYNTH}(\mathcal{C}_{\text{ref}}, q)$
14: $\hat{y}_L \leftarrow \text{LLMPRED}(q, \mathcal{D})$
15: $\hat{y}_M \leftarrow \text{VLMPRED}(q, I_t, f)$
16: $\hat{y} \leftarrow \text{FUSE}(\hat{y}_L, \hat{y}_M)$
17: **if** LOWCONF($\hat{y}$) **then**
18:     $\mathcal{C}_{\text{ref}} \leftarrow \text{UPDATECAP}(\mathcal{C}_{\text{ref}}, I_t, f)$
19:     $\hat{y} \leftarrow \text{REFINEANSWER}(q, \mathcal{C}_{\text{ref}}, \mathcal{F}, \mathcal{M})$
20: **end if**
21: **return** $\mathcal{C}_{\text{ref}}, \hat{y}$

3.4 Iterative Caption Enhancement and Cross-Modal Synthesis

The final component orchestrates LLM reasoning and VLM perception through confidence-based iterative refinement. Unlike single-pass systems, our framework continues refinement until sufficient confidence is achieved or computational bud-

get is exhausted, enabling progressive improvement through query-specific processing.

In the synthesis stage, the system transcends simple prediction fusion by leveraging the LLM's linguistic reasoning capabilities in conjunction with the VLM's visual understanding. The VLM generates descriptive summaries based on questions and visual content, providing contextual references that enhance the LLM's judgment process.

The answer fusion mechanism adaptively weights contributions from both modalities based on query characteristics and confidence metrics:

$$\hat{y} = \arg\max_{y} \alpha P_L(y|q, \mathcal{C}^+) + (1 - \alpha)P_V(y|q, \mathcal{F})$$

where P_L and P_V represent probability distributions from LLM and VLM respectively, $\mathcal{C}^+$ denotes enhanced captions, and α dynamically adjusts based on query type and information completeness. This adaptive fusion ensures each modality contributes according to its strengths.

When available information remains insufficient, the framework initiates iterative refinement by updating captions to incorporate missing details. During updates, the system integrates temporal sequence information and previously overlooked visual details. The VLM dynamically adjusts frame sequence processing, incorporating both temporal context through frame intervals and spatial details based on enriched captions.

The iterative nature creates a virtuous cycle: initial coarse understanding guides targeted information extraction, which enables more precise reasoning. Each iteration refines the context through query-specific processing, transforming video understanding from exhaustive computation to intelligent, adaptive extraction.

4 Experiments

We conduct comprehensive experiments to validate CAVIA's effectiveness in long-form video question answering. Our evaluation addresses three key questions: (1) Does closed-loop reasoning-perception coordination outperform static pipelines? (2) How does dynamic prompt generation improve over fixed visual processing? (3) What are the contributions of each component in our iterative framework? We evaluate on three challenging benchmarks—EgoSchema, NExT-QA, and IntentQA—demonstrating that CAVIA achieves state-of-the-art performance while providing insights into the importance of adaptive visual information extraction.

4.1 Datasets and Evaluation Metrics

We evaluate CAVIA on three complementary datasets that test different aspects of video understanding:

EgoSchema(Mangalam et al., 2023): Contains 250+ hours of first-person videos with 5,000 multiple-choice questions. Provides both Subset and Fullset evaluations, covering diverse human activities and human-object interactions.

NExT-QA(Xiao et al., 2021): Comprises 5,440 videos (44 seconds each) with Temporal (Tem.), Causal (Cau.), and Descriptive (Des.) question types, emphasizing temporal and causal reasoning evaluation.

IntentQA(Li et al., 2023a): Focuses on intent-based reasoning requiring understanding of causal-temporal relationships and underlying motivations in sequential activities.

We use accuracy as the primary evaluation metric, reporting the percentage of correctly answered multiple-choice questions. For EgoSchema, we evaluate on both Subset and Fullset categories; for NExT-QA, we report Temporal, Causal, Descriptive, and Average performance; for IntentQA, we report overall average accuracy.

4.2 Experimental Results

4.2.1 Comparison with State-of-the-Art Methods

Table 1 presents comprehensive comparisons with existing approaches, categorized into two paradigms: Vision-Centric Multimodal Architectures that primarily rely on visual features, and LLM-Driven Cross-Modal Synergy methods that leverage language models for reasoning.

Overall Performance: CAVIA achieves new state-of-the-art results across all benchmarks. On EgoSchema, we attain 71.6% accuracy on Subset (+5.4% over VideoTree) and 65.7% on Fullset (+4.6%), demonstrating significant improvements in long-form video understanding. For NExT-QA, CAVIA reaches 76.1% average accuracy, surpassing the previous best by 2.6%. Most notably, on the challenging IntentQA dataset, our method achieves 73.8%, a substantial 4.6% improvement over MoReVQA.

Analysis by Question Type: The performance breakdown reveals CAVIA's strengths in complex

Method	LLM	EgoSchema		NExT-QA				IntentQA
		Sub.	Full	Tem.	Cau.	Des.	Avg.	Avg.
Vision-Centric Multimodal Architectures								
SeViLA (Yu et al., 2023b)	GPT-3.5	-	-	61.3	61.5	75.6	63.6	60.9
LongViViT (Papalampidi et al., 2024)	GPT-3.5	56.8	33.3	-	-	-	-	-
MC-ViT-L (Balažević et al., 2024)	GPT-3.5	65.8	52.9	-	-	-	-	-
InternVideo2 (Wang et al., 2024a)	GPT-4	66.4	55.8	-	-	-	-	-
LLM-Driven Cross-Modal Synergy								
MVU (Ranasinghe et al., 2025)	GPT-4	60.3	37.6	55.4	48.1	64.1	55.2	-
LLoVi (Zhang et al., 2024a)	GPT-4	57.6	50.3	61.0	69.5	75.6	67.7	64.0
VideoAgent (Wang et al., 2025a)	GPT-4	60.2	54.1	64.5	72.7	81.1	71.3	-
MoReVQA (Min et al., 2024)	GPT-3.5	-	-	64.6	70.2	-	-	69.2
OptiGQA (Wang et al., 2025b)	GPT-4	-	-	65.8	72.3	78.0	71.0	68.7
VideoTree (Wang et al., 2025c)	GPT-4	66.2	61.1	67.0	75.2	_81.3_	73.5	66.9
CAVIA(ours)	GPT-4	70.8	**65.7**	70.2	75.7	79.2	74.5	72.8
CAVIA(ours)	GPT-4.1	**71.6**	65.3	**71.0**	**78.2**	79.6	**76.1**	**73.8**

Table 1: Performance comparison on EgoSchema, NExT-QA, and IntentQA benchmarks. Methods are categorized into Vision-Centric Multimodal Architectures and LLM-Driven Cross-Modal Synergy paradigms. For NExT-QA, we report results on Temporal (Tem), Causal (Cau), and Descriptive (Des) question types. Bold indicates best performance; underlined indicates second best; "-" denotes results not reported in original papers.

reasoning tasks. For temporal questions on NExT-QA, we achieve 71.0% accuracy compared to VideoTree's 67.0%, demonstrating the effectiveness of our coarse-to-fine localization strategy. In causal reasoning (78.2% vs. 75.2%), the gains highlight how targeted multimodal prompting extracts crucial cause-effect information missed by static approaches. While VideoTree slightly outperforms on descriptive questions (81.3% vs. 79.6%), CAVIA's consistent superiority in temporal and causal categorieswhich require deeper understandingvalidates our focus on dynamic information extraction.

Vision-Centric vs. LLM-Driven Approaches: The results clearly demonstrate the limitations of pure vision-centric methods (e.g., SeViLA: 63.6%) when handling complex reasoning. Among LLM-driven approaches, CAVIA's closed-loop design consistently outperforms open-loop methods like LLoVi and VideoAgent, confirming that iterative refinement and reasoning-guided perception are crucial for comprehensive video understanding. The performance gap is particularly pronounced on IntentQA, where understanding human intentions requires both visual details and sophisticated reasoningprecisely what our dynamic coordination enables.

4.2.2 Ablation Study: Dissecting the Impact of Core Components

To validate the effectiveness of our closed-loop architecture, we conduct comprehensive ablation studies examining how each component contributes to overall performance.

Multimodal Coordination Analysis Figure 3 visualizes the performance comparison across different modalities on various reasoning dimensions. We evaluate text-only baselines (leveraging only caption information), video-only approaches (using VLM without iterative refinement), and our complete CAVIA framework across temporal reasoning (Why, How, When), spatial understanding (Loc, Bef&Aft), counting tasks (Cnt), and accuracy metrics for different question categories.

The results reveal striking patterns that validate our design choices. Text-based methods excel at semantic reasoning (Acc_D: 79.5%) but struggle with fine-grained spatial-temporal understanding, particularly in localization tasks (60.5%). Conversely, video-based approaches demonstrate strong spatial awareness (Loc: 88.81%) but fail to maintain consistent performance across reasoning-intensive categories. CAVIA achieves balanced superiority across all dimensions, with particularly notable improvements in temporal reason-
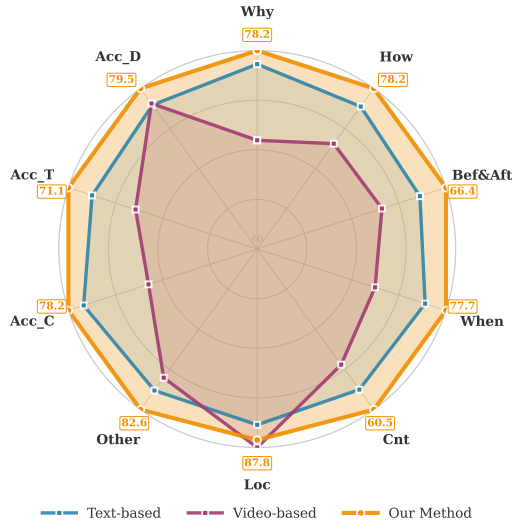
Figure 3: Performance comparison of text-based methods, video-based methods, and CAVIA across multiple NExT-QA reasoning dimensions including temporal (Why, How, When), spatial (Loc, Bef&Aft), counting (Cnt), and accuracy metrics for different question categories.

Method	NExT-QA	IntentQA
Baseline Methods		
LLoVi[α]	67.7	64.0
VideoTree[β]	73.5	66.9
Efficient LLMs		
COT (Llama-3.1-8B)	55.7	41.5
CAVIA (Llama-3.1-8B)	61.2	53.8
Gain	**+5.5**	**+12.3**
COT (Llama-4-Scout)	64.7	62.6
CAVIA (Llama-4-Scout)	70.6	67.9
Gain	**+5.9**	**+5.3**
Large-Scale LLMs		
COT (DeepSeek-V3)	71.8	67.3
CAVIA (DeepSeek-V3)	74.9	69.1
Gain	**+3.1**	**+1.8**
COT (GPT-4)	70.4	64.9
CAVIA (GPT-4)	74.5	72.8
Gain	**+4.1**	**+7.9**
COT (GPT-4.1)	71.8	65.6
CAVIA (GPT-4.1)	76.1	73.8
Gain	**+4.3**	**+8.2**

Table 2: Ablation study examining CAVIA's performance gains across different base LLMs. We compare our framework against caption-only chain-of-thought (COT) baselines using models ranging from Llama-3.1-8B to GPT-4.1. α (Zhang et al., 2024a), β (Wang et al., 2025c).

ing (Why: 78.2%, How: 78.2%, When: 77.7%) where the synergy between iterative caption enhancement and targeted visual extraction proves most valuable. This comprehensive advantage confirms that our closed-loop coordination successfully bridges the complementary strengths of both modalities.

Impact of Dynamic Prompt Generation Across LLM Scales Table 2 examines how CAVIA's benefits scale across different base LLMs, from efficient models like Llama-3.1-8B (Vavekanand and Sam, 2024) to state-of-the-art systems including Llama-4-Scout (Meta AI, 2025) (a MoE architecture with selective parameter activation), DeepSeek-V3 (Liu et al., 2025), GPT-4 and GPT-4.1[2].

CAVIA consistently outperforms caption-only COT baselines across all model scales. The improvements are particularly substantial with smaller models (Llama-3.1-8B: +5.5% on NExT-QA, +12.3% on IntentQA), where targeted multimodal prompting compensates for limited reasoning capacity. Even with powerful LLMs, CAVIA maintains significant gains (GPT-4.1: +4.3% on NExT-QA, +8.2% on IntentQA), demonstrating that our iterative refinement addresses a fundamen-

tal limitation: the inability to dynamically request missing visual details during reasoning.

The stronger gains on IntentQA across all models highlight how intent understanding critically benefits from query-specific visual extraction, as inferring human motivations requires nuanced integration of temporal dynamics and spatial context that static captions cannot capture. These model-agnostic improvements validate that closed-loop reasoning-perception coordination represents a fundamental advancement in video understanding methodology, offering a principled solution that enhances any language model's ability to reason about visual content.

Impact of Iterative Refinement Figure 4 demonstrates the effectiveness of our iterative refinement mechanism. The strong positive correlation (r=0.932) between iteration count and accuracy validates our closed-loop design: performance steadily improves from 67% (single itera-

---

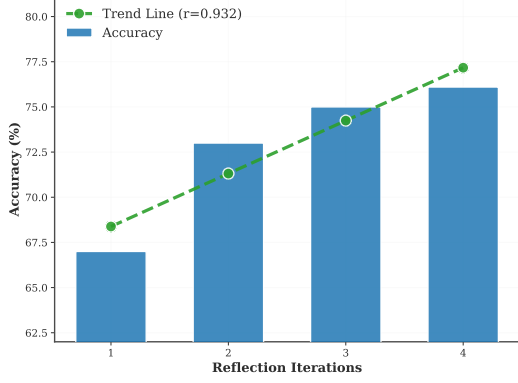[2]Both GPT-4 and GPT-4.1 can be accessed at https://openai.com/

Figure 4: Accuracy as a function of refinement iterations on NExT-QA dataset. The plot shows performance changes from iteration 1 to 4, with a trend line indicating correlation (r=0.932).
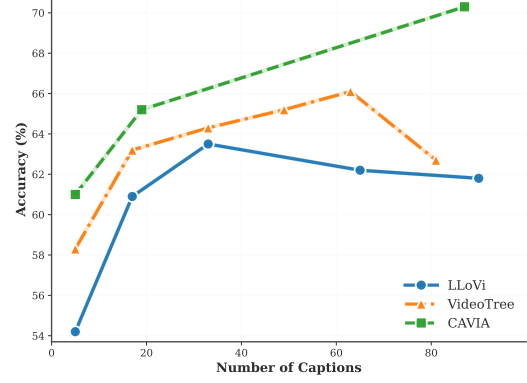

Figure 5: Accuracy versus number of captions processed on EgoSchema Subset. The plot compares CAVIA with LLoVi and VideoTree across different computational budgets.

tion) to 76% (four iterations), representing a substantial 9% absolute gain.

This consistent improvement pattern reveals how each refinement cycle serves a distinct purpose. Early iterations typically identify and address coarse-grained temporal gaps, while later cycles focus on fine-grained spatial details and causal relationships. During each iteration, the LLM identifies specific missing information and generates targeted prompts (e.g., "describe the tool manipulation sequence in frames 45-52"), ensuring that newly extracted information directly addresses reasoning gaps rather than adding redundant details. This targeted approach explains the sustained performance gains across all refinement stages.

4.3 Efficiency Analysis

Figure 5 illustrates CAVIA's efficiency characteristics by examining the relationship between computational cost (number of captions processed) and performance on EgoSchema Subset. Our framework demonstrates strong performance across the entire computational spectrum, maintaining a consistently superior position in the efficiency-effectiveness space.

This advantage stems from two key design choices. First, our hierarchical localization strategy prioritizes the most informative segments through coarse-to-fine filtering, maximizing early-stage performance gains. Second, the closed-loop architecture ensures intelligent resource utilization: the LLM evaluates information sufficiency at each step and generates targeted prompts only when additional details would improve reasoning,

preventing redundant processing while ensuring comprehensive understanding.

CAVIA's stable scaling behavior and flexible computational profile make it particularly suitable for real-world deployment. Whether operating under tight resource constraints or with abundant computational budget, the framework maintains effectiveness through adaptive information extraction. This allows practitioners to dynamically adjust the accuracy-computation trade-off based on specific application requirements, providing a practical solution for diverse video understanding scenarios.

5 Conclusion

We presented CAVIA, a closed-loop framework that fundamentally transforms video understanding through dynamic reasoning-perception coordination. By establishing iterative feedback between LLMs and VLMs, CAVIA addresses the critical limitation of static pipelines: the inability to adapt visual processing to query-specific needs. Our hierarchical localization and targeted multimodal prompting enable the system to progressively identify and extract precisely the visual information required for each unique query, transforming video understanding from exhaustive computation to intelligent, adaptive extraction. Extensive experiments across multiple benchmarks validate CAVIA's effectiveness and efficiency, demonstrating that dynamic coordination between reasoning and perception represents a fundamental advancement in video understanding methodology with strong potential for real-world applications.

8

Limitations

While CAVIA demonstrates strong performance across multiple benchmarks, several limitations merit discussion. First, our iterative refinement mechanism increases computational overhead compared to single-pass methods, particularly when multiple iterations are required for complex queries. Although we show efficiency gains through intelligent caption selection, the cumulative cost of repeated LLM-VLM interactions may limit deployment in latency-sensitive applications. Second, CAVIA's effectiveness relies on the quality of initial captionsvideos with poor textual descriptions or highly ambiguous visual content may not benefit fully from our dynamic prompting approach. Third, our current implementation uses fixed confidence thresholds for terminating iterations, which may not be optimal across all query types and video domains. Future work could explore adaptive termination criteria and more efficient architectures for iterative reasoning to address these limitations while maintaining the benefits of closed-loop coordination.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ziyi Bai, Ruiping Wang, and Xilin Chen. 2023. Glance and focus: Memory prompting for multi-event video question answering. In *Advances in Neural Information Processing Systems*, volume 36.

Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. 2024. Memory consolidation enables long-context video understanding. *arXiv preprint arXiv:2402.05861*.

Zhengxuan Chen, Shuo Wang, Deyang Yan, and Yushi Li. 2024. A spatio-temporl deepfake video detection method based on timesformer-cnn. In *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, pages 1–6. IEEE.

Xinxin Dong, Baoyun Peng, Haokai Ma, Yufei Wang, Zixuan Dong, Fei Hu, and Xiaodong Wang. 2025. Leadqa: Llm-driven context-aware temporal grounding for video question answering. *Preprint*, arXiv:2507.14784.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition.

Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14773–14783.

Sun Jianqiao, Su Yudi, Zhang Hao, Cheng Ziheng, Zeng Zequn, Wang Zhengjue, Chen Chunhui Qu, Bo, and Yuan Xin. 2025. Snapcap: efficient snapshot compressive scene captioning. *Advanced Imaging*, 2(1):011003. Accessed: 2025-07-22.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, and 1 others. 2024. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023a. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023b. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316.

Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. 2023a. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11963–11974.

Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, and Tat-Seng Chua. 2023b. Discovering spatio-temporal rationales for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13869–13878.

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. 2024. Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. *arXiv preprint arXiv:2409.20365*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2025. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. 2024. Robust visual question answering: Datasets, methods, and future challenges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5575–5594.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. 2025.

Drvideo: Document retrieval based long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18936–18946.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-04-05.

Juhong Min, Shyamal Buch, Arsha Nagrani, Minsu Cho, and Cordelia Schmid. 2024. Morevqa: Exploring modular reasoning models for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13235–13245.

Thong Thanh Nguyen, Zhiyuan Hu, Xiaobao Wu, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. Encoding and controlling global semantics for long-form video question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7109–7126.

Pinelopi Papalampidi, Skanda Koppula, Shreya Pathak, Justin Chiu, Joe Heyward, Viorica Patraucean, Jiajun Shen, Antoine Miech, Andrew Zisserman, and Aida Nematzdeh. 2024. A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14386–14397.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S Ryoo. 2025. Understanding long videos in one multimodal language model pass. *arXiv preprint arXiv:2403.16998*, 3.

Saul Santos and 1 others. 2025. ∞-video: A training-free approach to long video understanding via continuous-time memory consolidation. *arXiv preprint arXiv:2501.19098*.

Himanshu Sharma and Anand Singh Jalal. 2022. Convolutional neural networks-based vqa model. In *Proceedings of International Conference on Frontiers in Computing and Systems: COMSYS 2021*, pages 109–116. Springer.

Yaojie Shen, Xin Gu, Kai Xu, Heng Fan, Longyin Wen, and Libo Zhang. 2023. Accurate and fast compressed video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15558–15567.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others.

2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Raja Vavekanand and Kira Sam. 2024. Llama 3.1: An in-depth analysis of the next-generation large language model.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2025a. Videoagent: Long-form video understanding withǎlarge language model asǎagent. In *Computer Vision – ECCV 2024*, pages 58–76, Cham. Springer Nature Switzerland.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Jilan Xu, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024a. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision (ECCV)*, volume 15143 of *Lecture Notes in Computer Science*, pages 396–416. Springer.

Yueqian Wang, Yuxuan Wang, Kai Chen, and Dongyan Zhao. 2024b. Stair: Spatial-temporal reasoning with auditable intermediate results for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19215–19223.

Yufei Wang, Baoyun Peng, Zixuan Dong, Jia Fu, Xinxin Dong, Fei Hu, and Xiaodong Wang. 2025b. Optigqa: Llm-driven query optimization for efficient visual grounding in adaptive video question answering. In *Advanced Intelligent Computing Technology and Applications*, pages 200–211, Singapore. Springer Nature Singapore.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025c. 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.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer.

Geunhyuk Youk, Jihyong Oh, and Munchurl Kim. 2024. Fma-net: Flow-guided dynamic filtering and

iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 44–55.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023a. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, volume 36.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. 2023b. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. A simple llm framework for long-range video question-answering. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. 2024b. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*.

Junyang Zhang, Mu Yuan, Ruiguang Zhong, Puhan Luo, Huiyou Zhan, Ningkang Zhang, Chengchen Hu, and Xiang-Yang Li. 2025. A-vl: Adaptive attention for large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22461–22469.

A System Implementation Details

A.1 Framework Architecture

The CAVIA framework implements closed-loop coordination between reasoning and perception through three interconnected modules: Hierarchical Localization for coarse-to-fine filtering, Dynamic Prompting for targeted instruction generation, and Adaptive Extraction for multimodal analysis. Given video $\mathcal{V} = \{f_t\}_{t=1}^{T}$ with captions $\mathcal{C} = \{c_t\}_{t=1}^{T}$, we iteratively refine understanding to find optimal answer a^*:

$$a^* = \arg\max_{a \in \mathcal{A}} P(a|q, \mathcal{V}, \mathcal{C}^{(k)}) \qquad (1)$$

where $\mathcal{C}^{(k)}$ denotes enhanced captions at iteration k.

A.2 Algorithm Description

Algorithm 2 presents the complete framework with simplified notation for column constraints.

Algorithm 2 CAVIA Framework

Require: Video $\mathcal{V}$, Captions $\mathcal{C}$, Question q
Ensure: Answer a
1: Init: $\phi \leftarrow 0$, $k \leftarrow 0$, $\mathcal{C}^{(0)} \leftarrow \mathcal{C}$
2: $Q_s \leftarrow$ Decompose(q)
3: **while** $\phi < \theta$ **and** $k < K$ **do**
4:     $B_r \leftarrow$ Retrieve($Q_s, \mathcal{C}^{(k)}$)
5:     $G, F_g \leftarrow$ Analyze(q, B_r)
6:     **if** $|G| > 0$ **then**
7:         $P \leftarrow$ GenPrompts(G)
8:         **for** $(p_i, f_i) \in$ zip(P, F_g) **do**
9:             $d_i \leftarrow$ VLM($\mathcal{V}, p_i, f_i$)
10:            $\mathcal{C}^{(k+1)}[f_i] \leftarrow$ Merge($\mathcal{C}^{(k)}[f_i], d_i$)
11:         **end for**
12:     **end if**
13:     $a_L, \phi_L \leftarrow$ LLM($q, \mathcal{C}^{(k+1)}$)
14:     $a_V, \phi_V \leftarrow$ VLM($q, \mathcal{V}$)
15:     $a, \phi \leftarrow$ Fuse(a_L, a_V, ϕ_L, ϕ_V)
16:     $k \leftarrow k + 1$
17: **end while**
18: **return** a

The algorithm operates through five carefully orchestrated phases, each contributing to the closed-loop refinement process:

Phase 1 - Query Decomposition: The Decompose() function transforms complex, multi-faceted queries into atomic sub-questions $Q_s = \{q_1, q_2, ..., q_n\}$. This decomposition targets three primary aspects: temporal relations (e.g., action sequences and timing), spatial configurations (e.g., object locations and movements), and causal dependencies (e.g., action-consequence relationships). By breaking down complex queries, we enable more precise retrieval and focused analysis in subsequent phases.

Phase 2 - Hierarchical Retrieval: The Retrieve() function implements a two-stage filtering mechanism. First, caption sequences are partitioned into semantically coherent blocks B_i of size w (typically 10 frames). Then, relevance scoring is computed as:

$$s(B_i) = \alpha \cdot \text{sim}(q, B_i) + \beta \sum_{q_j \in Q_s} \text{sim}(q_j, B_i) \quad (2)$$

where sim() computes cosine similarity in the embedding space, B_i represents caption blocks, and α, β are weighting parameters balancing global query relevance against sub-question specificity. To ensure high recall in the coarse localization stage, we set the top-K value relatively large (typically $K = 5$), preventing potentially relevant segments from being prematurely excluded from the reasoning pipeline during early iterations. This conservative approach proves crucial for complex queries where relevant information may be distributed across multiple non-adjacent segments.

Phase 3 - Gap Analysis: The Analyze() function performs structured examination of retrieved content against query requirements. It outputs a gap set $G = \{g_1, g_2, ..., g_m\}$ where each gap g_i is characterized by: (i) type $\in \{temporal, spatial, causal\}$, (ii) associated frames $F_{g_i} \subset \{1, ..., T\}$, and (iii) specific information requirements. This systematic categorization enables targeted prompt generation in the next phase.

Phase 4 - Dynamic Prompt Generation: Based on identified gaps, GenPrompts() creates specialized instructions $P = \{p_1, p_2, ..., p_m\}$. For temporal gaps, prompts focus on action sequences and state transitions across frame intervals. Spatial prompts direct attention to specific regions or object attributes within individual frames. The prompt generation process adapts to both the gap type and the specific visual content characteristics, ensuring efficient information extraction.

Phase 5 - Multimodal Fusion: The Fuse() function implements adaptive combination of LLM and VLM predictions through confidence-
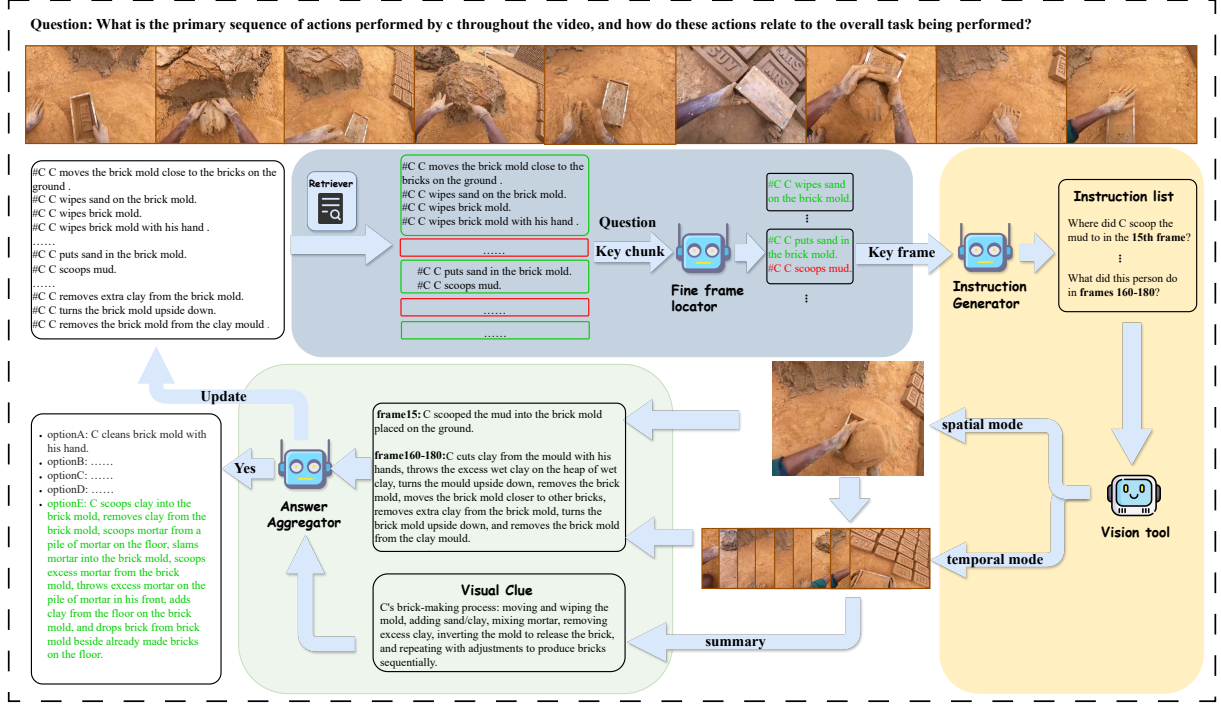
Figure 6: CAVIA's iterative refinement on a brick-making task. The system progressively identifies missing clay manipulation details through targeted prompts, demonstrating how closed-loop coordination enables comprehensive understanding of complex sequential activities.

weighted aggregation:

$$a = \arg\max_{a_i \in \mathcal{A}}[\gamma(q) \cdot P_L(a_i) + (1 - \gamma(q)) \cdot P_V(a_i)] \quad (3)$$

where $\gamma(q) \in [0, 1]$ is a query-dependent weighting function learned from question characteristics, P_L and P_V represent probability distributions from LLM and VLM respectively. The weighting adapts based on query type: reasoning-heavy questions favor LLM contributions ($\gamma > 0.5$), while visually-grounded queries emphasize VLM outputs ($\gamma < 0.5$). This adaptive fusion ensures each modality contributes according to its strengths for the specific query at hand.

B Experimental Analysis

B.1 Case Study: Iterative Refinement Process

Figure 6 demonstrates CAVIA's iterative process on an EgoSchema brick-making task. Starting with fragmented captions like "C moves brick mold" and "C wipes sand", the system identifies critical gaps in understanding the complete workflow.

In the first iteration, the Retriever identifies relevant caption chunks while the Fine Frame Locator pinpoints frames 15 and 160-180 as containing key information. The gap analysis reveals missing details about tool usage and action sequencing.

The second iteration employs targeted prompting. For spatial understanding, the system queries "Where did C scoop the mud to in the 15th frame?", extracting contextual information about workspace layout. For temporal dynamics, it generates "What did this person do in frames 160-180?", capturing the detailed sequence of clay cutting, mold manipulation, and brick formation.

The final synthesis combines enhanced captions with visual analysis. The spatial mode identifies tools and materials positioning, while temporal mode captures the complete action sequence: scooping mud into molds, removing excess clay, and systematic brick arrangement. This progressive refinement transforms fragmented observations into coherent understanding of the brick-making process.

B.2 Experimental Environment

Our experiments utilize a heterogeneous computing environment. Llama-3.1-8B-Instruct and Qwen2.5-VL runs on 2 * NVIDIA RTX 4090 GPU-24GB VRAM each, while Llama-4-Scout requires 4 * NVIDIA A800 GPU-80GB VRAM due to its MoE architecture. All other models in-

cluding GPT-4, GPT-4.1 and DeepSeek-V3 are accessed through API endpoints. The implementation uses PyTorch 2.1.0 with CUDA 12.1, maintaining consistent hyperparameters across all experiments with confidence threshold 0.85, maximum iterations 5, and temperature settings of 0 for LLMs and VLMs.

B.3 Ablation Analysis: Component Contributions

To validate the effectiveness of our closed-loop design, we conduct comprehensive ablation studies examining both individual components and their interactions.

B.3.1 Component Isolation Analysis

We systematically remove each component to measure its contribution:

$$\Delta_i = \text{Acc}(\text{Full}) - \text{Acc}(\text{Full} \setminus \{i\}) \qquad (4)$$

where i represents individual components. Beyond quantitative metrics, we analyze qualitative behavioral changes to understand each component's role in the overall system.

B.3.2 Interaction Effects and Synergy

The interaction between components reveals interesting patterns. When combining hierarchical localization with dynamic prompting, the joint improvement exceeds the sum of individual contributions:

$$\Delta_{HL+DP} > \Delta_{HL} + \Delta_{DP} \qquad (5)$$

This synergistic effect stems from the complementary nature of our design: precise localization enables more targeted prompts, while better prompts improve the quality of retrieved information. The closed-loop architecture amplifies these benefits through iterative refinement.

C Prompt Engineering

C.1 Prompt Design Philosophy

Our prompt system follows principles of clarity, specificity, and adaptability across five functional categories:

Question Decomposition: Transforms complex queries into atomic sub-questions addressing temporal ("when did X happen?"), spatial ("where is Y located?"), and causal ("why did Z occur?") aspects.

Gap Identification: Structures missing information analysis with explicit categorization, enabling precise targeting of visual extraction efforts.

Spatial Extraction: Directs VLM attention to specific regions with clear attribute specifications, focusing on object identities, positions, and fine-grained features.

Temporal Extraction: Guides sequential analysis across frame intervals, emphasizing state changes, motion patterns, and action continuity.

Cross-Modal Synthesis: Coordinates information fusion between modalities with confidence-based weighting, ensuring balanced contributions based on query characteristics.