Appendices

Appendix Contents

A. Explanation of UMIVR

Uncertainty Reduction Theory (URT) [1], originates from interpersonal communication studies and addresses how individuals seek information during interactions to alleviate uncertainty and enhance predictability. According to URT, uncertainty arises when communicators cannot accurately predict outcomes or interpret messages due to incomplete or ambiguous information. URT emphasizes that this uncertainty negatively impacts interactions, motivating communicators to actively gather additional information through questions, clarifications, or direct observations. Mathematically, the concept of uncertainty in URT aligns closely with Shannon's entropy in information theory, where greater entropy indicates higher unpredictability.

In interactive TVR, similar uncertainty challenges exist due to ambiguity in user queries and variability in the visual content of videos. Our UMIVR framework aligns naturally with URT principles by explicitly quantifying uncertainty through well-defined measures—semantic entropy [2, 3] for text ambiguity, Jensen–Shannon divergence for mapping ambiguity, and quality-based sampling for frame-level ambiguity—and systematically reducing these uncertainties via iterative interactions. Specifically, by leveraging these principled information-theoretic measures, UMIVR embodies URT's active information-seeking approach, enabling progressively clearer and more accurate text-to-video retrieval outcomes.

B. Reproduction of IVR Baselines

To ensure a rigorous and fair comparison with the proposed UMIVR approach, we faithfully reproduced the IVR baseline methods [4] within our experimental setting. This section elaborates on the implementation specifics, evaluation outcomes, and computational considerations of baselines.

Baseline Implementation We re-implement two variants of IVR: *ivrAuto*, which employs large language models for automatic clarifying question generation, and *ivrHeuristic*, which utilizes heuristic-based strategies for question generation. Besides, *ivrHeuristicWoAug* can be simply implemented by changing the config files. To align IVR variants closely with our proposed UMIVR framework, we re-implemented both variants using the unified multimodal model, VideoLLaVA. Specifically, the original ensemble-based components were replaced with the 4-bit quantized VideoLLaVA-7B model. Hyperparameters and experimental conditions were matched closely with the original setups to maintain fairness in comparisons.

Figure 3 provides a detailed illustration of the reproduced IVR heuristic pipeline. The pipeline systematically organizes interactions into iterative stages, initially focusing on the entire video, followed by granular interactions targeting the first and second halves separately. Additionally, general questions covering key objects, colors, and locations are integrated at the end of each interaction round, ensuring comprehensive query refinement. This structured reproduction closely adheres to the original IVR approach, effectively utilizing the capabilities of the VideoLLaVA.

Results Analysis Table 1 presents comprehensive performance comparisons between our VideoLLaVA-based IVR reproductions and the original IVR implementations across multiple evaluation metrics on the MSR-VTT dataset. Our reproduced models demonstrate clear improvements, especially in later interaction rounds. Notably, the *ivrHeuristic (VideoLLaVA)* model achieves significant gains across Hit@1, Hit@5, and Hit@10 metrics, confirming that leveraging a unified multimodal model like VideoLLaVA enhances query refinement and overall retrieval accuracy.

We also examine GPU memory consumption for various IVR implementations, as illustrated in Table 2. Original IVRAuto implementations incur substantial memory usage, primarily due to their reliance on multiple distinct models (e.g., T0++). By contrast, integrating the unified VideoLLaVA architecture with 4-bit quantization significantly reduces GPU memory usage by nearly sixfold compared to original implementations (e.g., 9,196 MB vs. 59,451 MB for *ivrAuto*), thereby greatly enhancing computational efficiency and practical deployment viability.

Overall, these reproduction efforts validate the robustness and effectiveness of employing a unified multimodal architecture, offering precise benchmarks for rigorous evaluation of the UMIVR framework.

	Rounds	0	1	2	3	4	5	6	7	8	9	10
Hit@1	ivrAuto (original)	42.5	50.1	54.2	56.4	58.4	61.2	62.4	64.0	64.8	66.4	67.5
	ivrAuto (VideoLLaVA)	42.2	51.7	52.7	53.9	55.2	60.3	64.0	67.9	68.9	70.3	72.5
	ivrHeuristic (original)	42.5	49.0	56.0	63.0	67.2	68.9	71.5	73.0	74.0	74.9	75.5
	ivrHeuristic (VideoLLaVA)	42.2	50.0	57.2	64.2	69.3	70.4	71.5	72.7	73.6	74.6	76.2
Hit@5	ivrAuto (original)	65.3	73.9	81.1	83.4	84.2	84.9	85.5	85.9	86.0	86.1	86.3
	ivrAuto (VideoLLaVA)	65.8	75.6	76.6	76.9	79.2	83.2	87.2	89.0	89.3	90.2	91.2
	ivrHeuristic (original)	65.3	72.0	78.9	84.3	86.3	88.3	89.3	90.7	91.4	91.7	92.2
	ivrHeuristic (VideoLLaVA)	65.8	73.2	79.9	86.9	89.2	89.8	90.2	90.9	91.0	91.7	92.4
Hit@10	ivrAuto (original)	74	81.7	83.1	84.1	86.0	88.6	90.1	90.8	91.1	91.5	91.9
	ivrAuto (VideoLLaVA)	75.3	83.9	84.4	84.8	87.5	90.3	92.2	93.3	93.3	93.7	94.3
	ivrHeuristic (original)	74	81.1	86.8	90.5	92.0	93.2	93.8	94.7	95.1	95.1	95.8
	ivrHeuristic (VideoLLaVA)	75.3	82.3	84.9	87.6	90.1	90.6	91.8	93.0	94.1	96.2	96.5

Table 1. **Reproduction results of IVR [4] with VideoLLaVA.** This table compares the original IVR method with its VideoLLaVA-based implementation across multiple interaction rounds. We evaluate two IVR variants—*ivrAuto* and *ivrHeuristic*—on standard retrieval metrics (Hit@1, Hit@5, and Hit@10). The results demonstrate that integrating VideoLLaVA into IVR leads to improved retrieval performance, particularly in later interaction rounds.

Model	ivrAuto (original)	ivrAuto(4bit) (VideoLLaVA)	ivrHeuristic (original)	ivrHeuristic(4bit) (VideoLLaVA)	UMIVR (4bit)	UMIVR (8bit)	UMIVT (woquant)
GPU-Memory-Usage(MB)	59451	9196	14934	10010	10210	15605	19297

Table 2. GPU memory usage comparison of different models. We evaluate the GPU memory consumption of various interactive retrieval models on a subset of MSR-VTT-1kA (first 10 samples) during inference. The results show that **ivrAuto (original)** exhibits the highest memory usage due to its reliance on T0++[7] for captioning, significantly increasing computational overhead. In contrast, the **VideoLLaVA-based 4-bit quantized models** (ivrAuto (4bit), ivrHeuristic (4bit), and UMIVR (4bit)) achieve substantial memory savings while maintaining competitive performance, making them efficient alternatives. The proposed **UMIVR framework** effectively balances performance and memory efficiency, with its 4-bit quantized version consuming only a fraction of the GPU memory required by non-quantized baselines.

C. Analysis of Uncertainty Score Distributions

In this section, we conduct a detailed analysis of the distributions of the two uncertainty metrics introduced in our proposed UMIVR framework: the Text Ambiguity Score (TAS) and the Mapping Uncertainty Score (MUS). These analyses highlight the effectiveness and interpretability of our uncertainty metrics across iterative rounds.

Figure 1 shows the distributions of TAS values across multiple interaction rounds during the interactive retrieval process. At the initial stage (Round 0), queries exhibit notably high textual ambiguity, predominantly concentrated above a TAS value of 0.6. As the interactive retrieval proceeds and clarifying questions iteratively refine user queries, the TAS distributions consistently shift leftward, signaling a clear reduction in ambiguity. Specifically, from Rounds 3 to 5, the majority of queries already fall below our defined TAS uncertainty threshold of 0.5, indicating substantial resolution of textual ambiguity. By Round 7, the distribution further tightens around lower TAS values, underscoring the robustness of UMIVR's adaptive clarification strategy in progressively refining textual queries and reducing ambiguity.

D. Discussion on Quantization Strategies

We evaluate the effectiveness and computational trade-offs of different quantization strategies in UMIVR, comparing three configurations: UMIVR with 4-bit quantization (UMIVR-4bit), 8-bit quantization (UMIVR-8bit), and without quantization (UMIVR-woquant).

Table 3 reports retrieval performance across multiple interaction rounds. UMIVR-woquant consistently achieves the best results, particularly in later rounds, highlighting the advantages of full-precision models. However, the improvements over quantized versions remain modest. UMIVR-8bit slightly outperforms UMIVR-4bit, though the latter remains highly competitive with minimal degradation.

Table 2 shows GPU memory usage on the MSR-VTT dataset. Traditional IVR models incur high memory overhead due to ensemble-based architectures, whereas VideoLLaVA-based UMIVR significantly reduces memory consumption. Notably, UMIVR-4bit requires only 10,210 MB, less than half of UMIVR-8bit and one-third of UMIVR-woquant, demonstrating substantial efficiency.

Given the trade-off between accuracy and computational cost, we adopt UMIVR-4bit as the final configuration, offering near-optimal performance while drastically reducing GPU memory consumption, making it well-suited for real-world deployment.

	Rounds	0	1	2	3	4	5
R@1	UMIVR-4bit	43.1	52.5	57.9	61.3	63.7	65.0
	UMIVR-8bit	43.1	53.7	58.8	61.6	64.0	65.1
	UMIVR-woquant	**43.8**	**54.2**	**58.9**	**63.3**	**65.4**	**67.1**
Hit@10	UMIVR-4bit	75.8	86.7	89.9	92.7	93.8	94.8
	UMIVR-8bit	75.8	**87.6**	91.6	**94.3**	95.2	**95.9**
	UMIVR-woquant	**76.3**	87.4	**92.1**	93.8	**95.6**	95.8
BRI	UMIVR-4bit	-	1.10	0.92	0.81	0.73	0.67
	UMIVR-8bit	-	1.10	**0.89**	**0.77**	**0.69**	0.63
	UMIVR-woquant	-	**1.09**	0.90	0.78	**0.69**	**0.62**

Table 3. **Performance comparison of UMIVR with different quantization strategies.** We evaluate UMIVR under three settings: 4-bit quantization (UMIVR-4bit), 8-bit quantization (UMIVR-8bit), and non-quantized (UMIVR-woquant) across multiple interaction rounds. The results indicate that while UMIVR-woquant achieves the best overall performance, especially in later interaction rounds. However, the non-quantized version significantly increases GPU memory consumption, making it less practical. Meanwhile, UMIVR-8bit offers slightly better performance than UMIVR-4bit but at the cost of higher GPU usage. Given the trade-off between computational efficiency and retrieval performance, we adopt 4-bit quantization as the final configuration, as it achieves near-optimal results while maintaining significantly lower memory requirements.

	Hit@1	Hit@5	Hit@10	Rounds Min	Max	Mean	Median	25%	75%
UMIVR-EarlyStop	65.8	84.7	90.7	1	10	3.05	3.0	2.0	3.0

Table 4. **Evaluation of UMIVR with an automatic early stopping mechanism.** Traditional interactive TVR methods typically run for a fixed number of interaction rounds (e.g., $n = 10$), but real-world user interactions require a more dynamic approach to optimize user experience. To achieve this, we set a Text Ambiguity Score (TAS) threshold $\alpha = 0.4$ and a Mapping Uncertainty Score (MUS) threshold $\beta = 0.2$, allowing the system to automatically terminate interactions when the query is sufficiently refined. The results show that UMIVR-EarlyStop maintains strong retrieval performance while significantly reducing the average number of interaction rounds (**Mean = 3.05, Median = 3.0**). This validates the effectiveness of TAS and MUS as uncertainty indicators and demonstrates the practical benefits of adaptive interaction termination in improving efficiency and user experience.

E. Early Stopping Strategy for Interactive Retrieval

In practical interactive retrieval scenarios, optimizing the number of user interactions is crucial for enhancing user experience and efficiency. Traditional interactive text-to-video retrieval (TVR) methods often rely on a fixed number of interaction rounds, potentially leading to unnecessary or redundant interactions. To address this, we introduce an automatic early stopping strategy into our UMIVR framework, utilizing uncertainty metrics—specifically, the Text Ambiguity Score (TAS) and the Mapping Uncertainty Score (MUS)—as termination criteria.

As shown in Table 4, by setting thresholds of $\alpha = 0.4$ for TAS and $\beta = 0.2$ for MUS, UMIVR effectively determines when further interactions become unnecessary. This adaptive early stopping mechanism achieves competitive retrieval performance (Hit1 = 65.8%, Hit5 = 84.7%, Hit10 = 90.7%) while significantly reducing the average interaction rounds to approximately 3.05 (median of 3 rounds). The distribution of interaction rounds, ranging from a minimum of 1 to a maximum of 10, underscores the system's flexibility in dynamically adapting to varying query difficulties.

These results validate the practical effectiveness of using TAS and MUS as uncertainty-driven early stopping indicators, highlighting substantial improvements in retrieval efficiency and user interaction quality.

F. Impact of NR-IQA Methods in TQFS

The Temporal Quality-based Frame Sampler (TQFS) introduced in our UMIVR framework relies on No-Reference Image Quality Assessment (NR-IQA) methods to select high-quality video frames effectively. Here, we briefly examine the impact of two NR-IQA methods—BRISQUE [5] and Laplacian Variance [6]—on retrieval performance and runtime efficiency.

Table 5 compares the retrieval performance and computational costs associated with these NR-IQA methods. Although BRISQUE achieves marginally superior performance, it incurs significantly higher computational overhead (4.09 s/video) compared to the simpler Laplacian Variance method (0.29 s/video). Given this minimal difference in retrieval accuracy and substantial computational advantage, we select Laplacian Variance for TQFS to balance efficiency and retrieval effectiveness in practical deployments.

G. Prompt Design for UMIVR

We carefully design structured prompts at each stage of the UMIVR framework to systematically address different uncertainty scenarios and improve retrieval accuracy. Specifically, we first introduce detailed yet precise prompts for offline extraction of video meta-information, including video captioning, primary object identification, and semantic scene classification (Table 6). These prompts enforce clear, evidence-based visual descriptions and explicitly discourage speculative content generation.

Subsequently, in the interactive retrieval phase, we propose a hierarchical, uncertainty-guided prompting system to dynamically generate clarifying questions appropriate to query ambiguity levels (Tables 7, 8, 9). Our prompts progressively transition from open-ended inquiries addressing high textual ambiguity, to targeted visual-distinguishing questions under high mapping uncertainty, and enrichment-oriented queries when uncertainty is minimal. Additionally, simulated user responses are generated through structured prompts that incorporate diverse visual details, enabling realistic iterative refinement and further enhancing retrieval precision (Tables 10, 11).

In summary, our prompt design provides a principled, flexible, and effective mechanism for managing uncertainty throughout the interactive retrieval process, facilitating clear communication between user and model, and ultimately improving overall retrieval performance.

NR-IQA Methods	R@1	R@5	R@10	MdR	MnR	Runtime (s/video)
BRISQUE [5]	43.3	66.4	75.9	2	22.3	4.09
Laplacian Variance [6]	43.1	66.1	75.8	2	22.8	0.29

Table 5. **Comparison of TQFS with different No-Reference Image Quality Assessment (NR-IQA) methods.** We evaluate the impact of different NR-IQA methods on retrieval performance and runtime efficiency. While BRISQUE achieves slightly better results, it comes at a significant computational cost, requiring 4.09 seconds per video compared to 0.29 seconds for Laplacian Variance. Given the minimal performance difference but substantial speed advantage, we adopt Laplacian Variance in TQFS to ensure computational efficiency without sacrificing retrieval effectiveness.

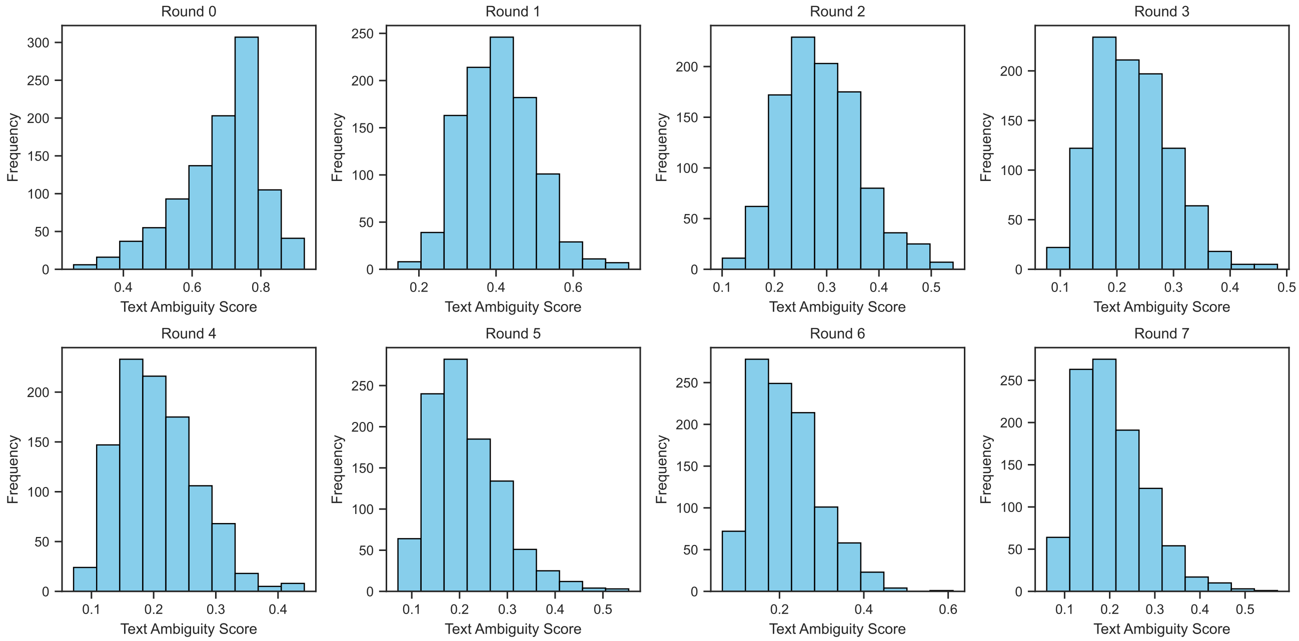

Figure 1. **Distribution of Text Ambiguity Score (TAS) across interaction rounds.** The histograms illustrate the progressive reduction in TAS as the interactive retrieval process advances from Round 0 to Round 7. Initially, the majority of queries exhibit high ambiguity, with a strong concentration above 0.6 in Round 0. As clarifying questions iteratively refine the queries, the TAS distribution shifts leftward, indicating reduced textual ambiguity. By Round 3–5, most queries fall below the TAS threshold of 0.5 (marked as our uncertainty resolution threshold), and by Round 7, ambiguity is significantly minimized, demonstrating the effectiveness of UMIVR's adaptive clarification strategy in refining textual queries.

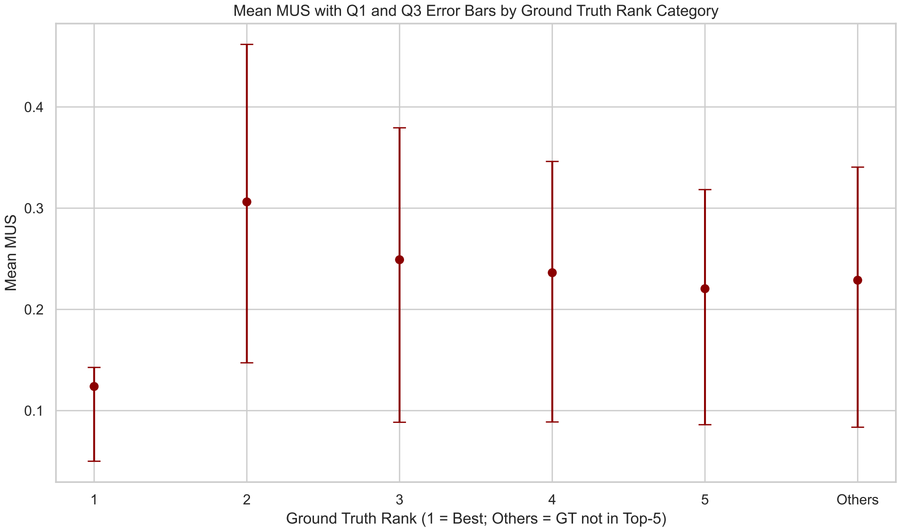

Figure 2. **Mapping Uncertainty Score (MUS) across different Ground Truth (GT) ranks.** The figure shows the relationship between MUS and the rank position of the ground truth video in the retrieval results. When GT is ranked 1, MUS is consistently low, indicating high confidence in retrieval. However, when GT is ranked 2, MUS is noticeably higher, reflecting the system's difficulty in distinguishing between the top-ranked candidates. This trend confirms that MUS effectively identifies ambiguous retrieval scenarios, particularly when the correct video is close but not yet ranked first. We set the MUS threshold to 0.2 in UMIVR to trigger interactive refinement in cases of high mapping uncertainty.

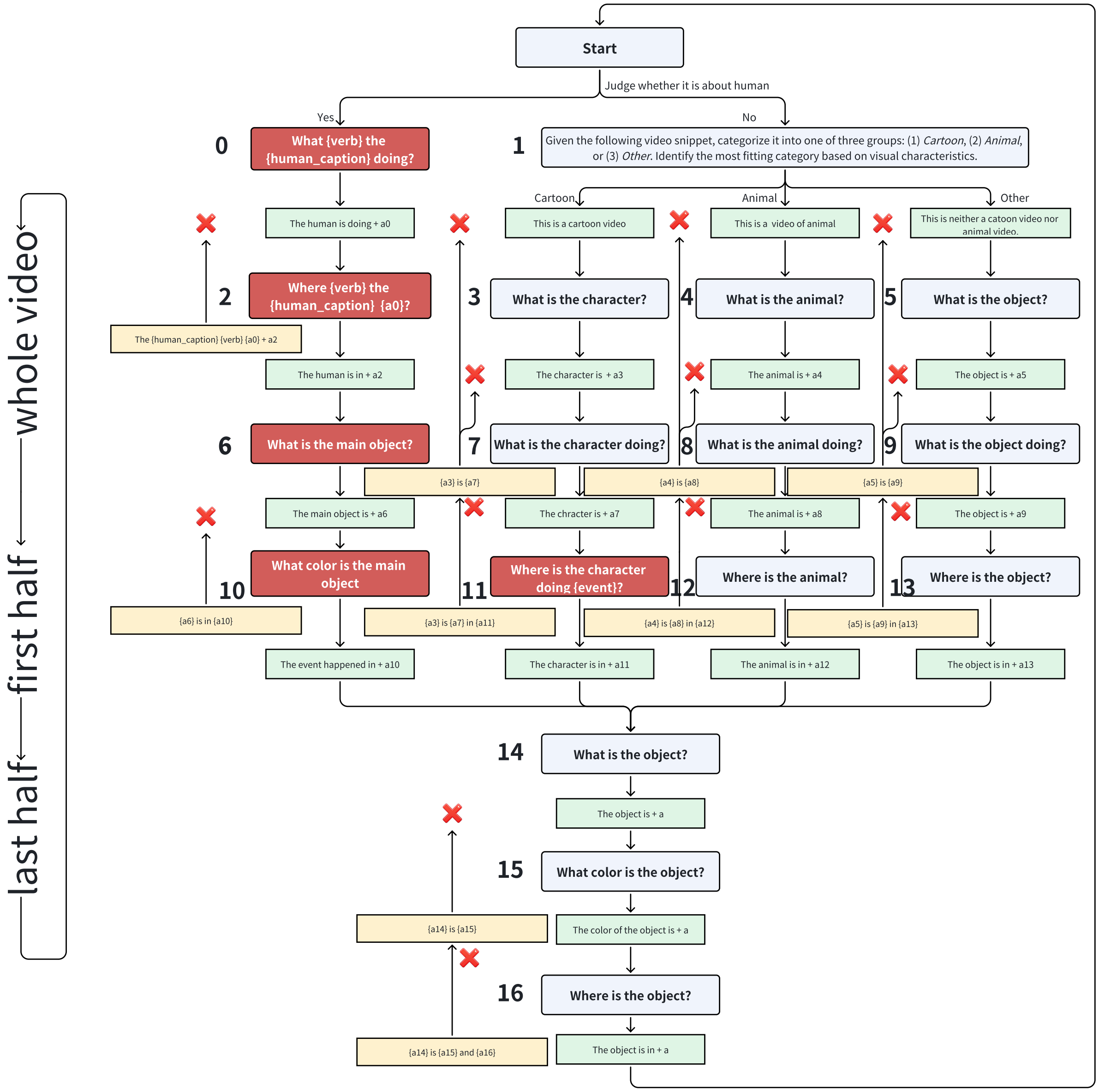

Figure 3. **Overview of the Reproduced IVR-Heuristic Pipeline Structure.** The pipeline follows a structured, heuristic-driven interactive retrieval approach for text-to-video retrieval. It first categorizes videos into four types: human, cartoon, animal, and other, tailoring its questioning strategy accordingly to extract key video features. To improve retrieval effectiveness, the pipeline segments video interactions into three stages: whole video, first half, and last half, enabling more granular query refinement. Additionally, to prevent missing critical details, it incorporates general questions at the end of each round, covering main objects, colors, and locations. The process iterates through multiple rounds until reaching a predefined maximum, progressively refining the query and enhancing retrieval accuracy.

H. Limitations

Although our proposed UMIVR framework demonstrates superior performance and strong generalization across multiple retrieval benchmarks, several limitations remain open for future exploration.

First, UMIVR explicitly quantifies uncertainty using external metrics (TAS, MUS, TQFS) rather than relying directly on the inherent uncertainty-awareness capabilities of the underlying Large Language Model (LLM). In other words, current LLMs cannot intrinsically perceive the uncertainty within user queries effectively. If LLMs could internally recognize and quantify these uncertainties, it would enable a more tightly coupled and contextually adaptive generation of clarifying interactions, potentially leading to further improvements in retrieval performance.

Second, our experiments rely on a simulated question-answering mechanism that mimics user responses. While human-simulating question-answering significantly reduces the practical costs of evaluating interactive systems, it inevitably differs from real human interactions in nuanced aspects, such as response variability, hesitation, or misunderstanding. Thus, real-world performance may differ from simulated scenarios, necessitating future validation through human-in-the-loop studies.

Third, the effectiveness of our approach inherently depends on the performance of the underlying multimodal LLM (VideoLLaVA). As current multimodal LLMs still struggle with certain challenging scenarios, such as fine-grained action recognition, temporal understanding, or handling complex textual semantics, improvements in the intrinsic capabilities of VideoLLMs would directly enhance the reliability and overall retrieval performance of our UMIVR framework.

Finally, previous works such as IVR [4] have expressed concern regarding potential information leakage when utilizing the same backbone model (e.g., BLIP) for retrieval and VideoQA. However, we found this concern unwarranted in our experiments. Since our retrieval and VideoQA processes are both stateless and executed through independent inference calls, we observed no such leakage effect. Nevertheless, researchers adopting similar strategies should remain cautious and explicitly verify the absence of leakage in different model architectures.

Details about video meta-information generation prompt in UMIVR
System Prompt: A conversation between a curious human and an AI assistant. The assistant is specialized in analyzing video content and provides detailed, precise, and evidence-based descriptions. Follow these guidelines strictly: - **Precision**: Describe only what is directly observable from the video. - **Detail**: Include all readily visible details while keeping responses focused. - **No Speculation**: If any part of the content is uncertain, explicitly state the uncertainty instead of guessing.
Caption Prompt: {video_features} Please provide a detailed and highly accurate caption that fully describes the overall scene or main activity in this video. Make sure your caption includes all relevant visual details and does not exceed 80 words. Do not add any information that is not clearly supported by the video content.
Main Objects Prompt: {video_features}Based solely on the visible content of the video, list up to five primary objects or characters you can clearly identify. Each item should be provided as a single word or a brief noun phrase (e.g., 'man', 'tree', 'couch'). Only include items that are explicitly visible and avoid any speculation.
Scene Type Prompt: {video_features} Based on the visual content of the video, identify the primary setting, scene type, or dominant visual theme by listing up to five concise keywords (e.g., 'underwater', 'indoor', 'black'). Only include keywords that are directly evident from the video, and do not include any speculative information.
Max New Tokens: 1024
Temperature: 0.1

Table 6. **Prompts used in UMIVR for generating video meta-information.** These prompts guide the video LLM in producing accurate and detailed video descriptions, identifying primary objects, and categorizing scene types. The system prompt enforces strict adherence to precision, detail, and avoidance of speculation. The caption prompt ensures comprehensive yet concise descriptions, while the main objects and scene type prompts extract key visual elements.

Details about level-0 question generation prompt in UMIVR
System Prompt: You are an advanced AI specialized in asking clarifying questions for vague queries. Your task is to extract details—such as appearance, activities, or events—to enable precise retrieval.
User Prompt: Query: {text_query} Ask one open-ended clarifying question focusing on the subject's appearance, activities, or events.Return ONLY the question.
Max New Tokens: 1024
Temperature: 0.1

Table 7. **Prompt design for Level-0 question generation in UMIVR.** This prompt is used when the initial text query exhibits high text ambiguity, as determined by the Text Ambiguity Score (TAS). The LLM generates an open-ended clarifying question aimed at refining vague queries by eliciting additional details about the subject's appearance, activities, or events. This process helps reduce uncertainty in the retrieval task.

Details about level-1 question generation prompt in UMIVR
System Prompt: You are a clarifying question generator for text-video retrieval. Given a user query and multiple video info, your task is to generate one question that focuses on visual differences. **User Prompt**: Query: {text_query} Videos: {video_meta_info_list} Ask one question starting with What, Where, or Who to distinguish these videos based on visual details. Return ONLY the question. **Max New Tokens**: 1024 **Temperature**: 0.1

Table 8. **Prompt design for Level-1 question generation in UMIVR.** This prompt is used when the query has low text ambiguity but exhibits high mapping uncertainty, as determined by the Mapping Uncertainty Score (MUS). Given a user query and multiple retrieved video candidates, the LLM generates a clarifying question that highlights visual distinctions between them. The question is structured to start with ”What,” ”Where,” or ”Who,” ensuring a focus on differentiating key visual elements.

Details about level-2 question generation prompt in UMIVR

System Prompt:
You are an advanced AI specialized in asking clarifying questions for queries. Your task is to extract details—such as appearance, activities, or events—to enable precise retrieval.

User Prompt:
You need to ask a question based on a user query.
1. First you need to evaluate whether the user's query includes sufficient visual details (such as characters, colors, objects, or locations).
User Query: {cur_text_query}

2. Ask a question
- If details are missing, generate one question to gather them.
- If the query is already detailed, generate a clarifying question to further enrich the description (e.g., 'What other objects are present?', 'What is the main color?', or 'Where is the event taking place?').

Return ONLY the question, nothing else.

Max New Tokens:
1024

Temperature:
0.1

Table 9. **Prompt design for Level-2 question generation in UMIVR.** This prompt is used when both text ambiguity and mapping uncertainty are low, but further query enrichment is beneficial. The UMIVR evaluates whether the user's query contains sufficient visual details (e.g., characters, colors, objects, locations). If key details are missing, it generates a question to obtain them; otherwise, it formulates an enrichment-oriented question to enhance the query's specificity. This iterative refinement helps maximize retrieval accuracy.

Details about human-simulation answer generation prompt in UMIVR

System Prompt:
You are a video question answering assistant. When provided with a video and a question, your task is to provide a concise, one-sentence answer. Your answer should clearly state the key visual details such as people, objects, scenes, and events. Keep it clear, direct, and focused on essential information.

User Prompt:
{video_features}

Question: {question}

Provide a one-sentence answer that clearly identifies the key visual details in the video, such as people, objects, scenes, and events.

Max New Tokens:
1024

Temperature:
0.7

Table 10. **Prompt design for human-simulation answer generation in UMIVR.** This prompt is used to simulate user responses in interactive retrieval by generating concise, one-sentence answers to clarifying questions based on video content. The video LLM extracts key visual details, including people, objects, scenes, and events, ensuring clarity and relevance. A higher temperature setting (0.7) is used to introduce variability, better mimicking the natural diversity in human responses. This simulation enables iterative query refinement, improving retrieval accuracy through realistic user interactions.

Details about query refinement prompt in UMIVR
System Prompt: You are an expert in query refinement for interactive text-video retrieval. Your task is to synthesize and update a previous query with new details from the current answer. Ensure the new query includes key information (e.g., characters, events, objects, colors, locations) and does not exceed 60 words.) **User Prompt**: Previous Query: {pre_query} Current Answer (includes new information to enhance video retrieval): {cur_answer} Combine the above into one concise, positive declarative sentence that includes key details (characters, events, objects, colors, locations, etc.). Ensure the new query leverages the new information from the current answer for better retrieval and is no longer than 60 words. Only return the refined query, nothing else. **Max New Tokens**: 1024 **Temperature**: 0.1

Table 11. **Prompt design for query refinement in UMIVR.** This prompt is used to iteratively improve user queries by incorporating new details extracted from simulated user answers. The LLM synthesizes the previous query with newly provided information (e.g., characters, events, objects, colors, locations), ensuring a more precise and informative query for video retrieval.

References

[1] Charles R Berger and Richard J Calabrese. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human communication research*, 1(2):99–112, 1974. 1

[2] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630, 2024. 1

[3] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 1

[4] Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11057–11067. IEEE, 2023. 1, 2, 7

[5] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. Blind/referenceless image spatial quality evaluator. In *Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers, ACSCC 2011, Pacific Grove, CA, USA, November 6-9, 2011*, pages 723–727. IEEE, 2011. 4

[6] Maria MP Petrou and Costas Petrou. *Image processing: the fundamentals*. John Wiley & Sons, 2010. 4

[7] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, et al. Multitask prompted training enables zero-shot task generalization, 2021. 2