Quantifying and Narrowing the Unknown: Interactive Text-to-Video Retrieval via Uncertainty Minimization

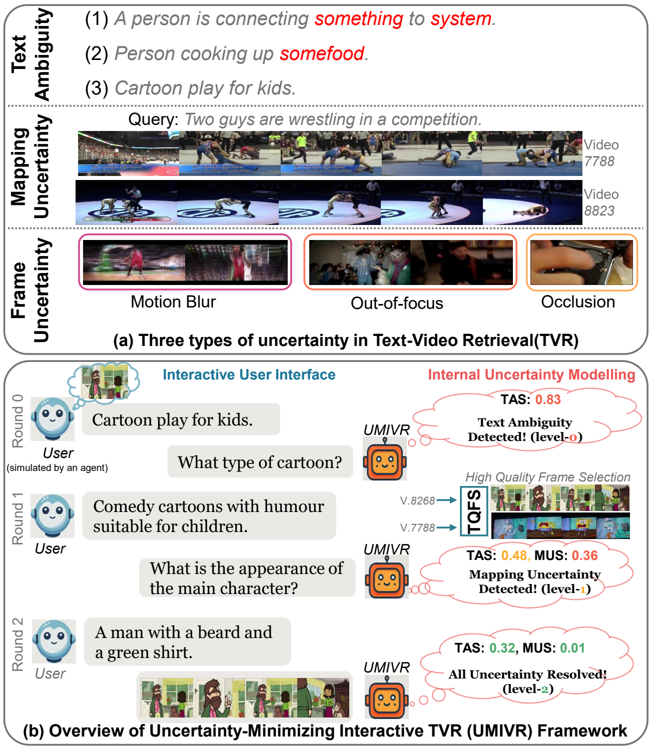
Bingqing Zhang[1,2] Zhuo Cao[1] Heming Du[1] Yang Li[2] Xue Li[1*] Jiajun Liu[2,1*] Sen Wang[1*]
[1] The University of Queensland, Australia
[2] CSIRO Data61, Australia
{bingqing.zhang, william.cao, heming.du}@uq.edu.au, yang.li1@csiro.au
xueli@eesc.uq.edu.au, jiajun.liu@csiro.au, sen.wang@uq.edu.au

Abstract

Despite recent advances, Text-to-video retrieval (TVR) is still hindered by multiple inherent uncertainties, such as ambiguous textual queries, indistinct text-video mappings, and low-quality video frames. Although interactive systems have emerged to address these challenges by refining user intent through clarifying questions, current methods typically rely on heuristic or ad-hoc strategies without explicitly quantifying these uncertainties, limiting their effectiveness. Motivated by this gap, we propose UMIVR, an Uncertainty-Minimizing Interactive Text-to-Video Retrieval framework that explicitly quantifies three critical uncertainties-text ambiguity, mapping uncertainty, and frame uncertainty-via principled, training-free metrics: semantic entropy-based Text Ambiguity Score (TAS), Jensen-Shannon divergence-based Mapping Uncertainty Score (MUS), and a Temporal Quality-based Frame Sampler (TQFS). By adaptively generating targeted clarifying questions guided by these uncertainty measures, UMIVR iteratively refines user queries, significantly reducing retrieval ambiguity. Extensive experiments on multiple benchmarks validate UMIVR's effectiveness, achieving notable gains in Recall@1 (69.2% after 10 interactive rounds) on the MSR-VTT-1k dataset, thereby establishing an uncertainty-minimizing foundation for interactive TVR. Code will be avaliable at https://github.com/bingqingzhang/umivr.

1. Introduction

Text-to-Video Retrieval (TVR) has emerged as a crucial task that bridges computer vision and natural language processing, aiming to retrieve relevant video content based on textual queries. Owing to its broad applicability in video search and recommendation, TVR has rapidly evolved from early attention-based mechanisms [14, 33] to vision–language pretraining models [27, 46]. This research paradigm is now diverse, encompassing streamlined en-


Text Ambiguity
(1) A person is connecting something to system.
(2) Person cooking up somefood.
(3) Cartoon play for kids.
Mapping Uncertainty
Query: Two guys are wrestling in a competition.
Video 7788
Video 8823
Frame Uncertainty
Motion Blur
Out-of-focus
Occlusion
Interactive User Interface
Internal Uncertainty Modelling
Round 0
User
(simulated by an agent)
Cartoon play for kids.
UMIVR
What type of cartoon?
TAS: 0.83
Text Ambiguity Detected! (level-0)
High Quality Frame Selection
Round 1
Comedy cartoons with humour suitable for children.
V.8268
V.7788
TQFS
What is the appearance of the main character?
TAS: 0.48, MUS: 0.36
Mapping Uncertainty Detected! (level-1)
Round 2
A man with a beard and a green shirt.
TAS: 0.32, MUS: 0.01
All Uncertainty Resolved! (level-2)

(a) Three types of uncertainty in Text-Video Retrieval(TVR)

(b) Overview of Uncertainty-Minimizing Interactive TVR (UMIVR) Framework

Figure 1. Illustration of uncertainty challenges in Text-to-Video Retrieval (TVR) and our proposed UMIVR framework. **(a)** Three types of uncertainty that commonly degrade retrieval performance. **(b)** UMIVR explicitly quantifies all three types of uncertainties per interaction round: Text Ambiguity Score (TAS), Mapping Uncertainty Score (MUS), and Frame Uncertainty (addressed via TQFS for selecting high-quality frames). UMIVR then iteratively generates adaptive clarifying questions, progressively reducing these uncertainties to achieve precise video retrieval.

[*] Corresponding authors

coder designs [35, 55], advanced training strategies [12], and improved feature alignment between text and video [37, 52, 58]. These innovations demonstrate significant performance gains in various evaluations.

Despite notable advances, TVR still remains challenging due to various forms of uncertainty that arise from both textual and visual sources, as illustrated in Fig. 1(a). First, **text ambiguity** poses a persistent obstacle: textual queries can often be vague, incomplete, or contain polysemous words (e.g.,"something," "somefood," or very generic phrases like "Cartoon play for kids"), leading to underspecified retrieval targets. Second, **mapping uncertainty** highlights that even clearly formulated textual queries (e.g., "Two guys are wrestling in a competition") can correspond to multiple plausible videos within a large and diverse dataset, making it difficult to pinpoint the most relevant candidate among visually similar alternatives. Third, **frame uncertainty** arises from deteriorated video frames—such as motion blur, out-of-focus shots, or occlusion of key objects—which obscure important visual cues essential for accurate retrieval. Taken together, these intertwined uncertainties largely degrade retrieval performance, underscoring the necessity for robust mechanisms that explicitly handle ambiguity and noise in both textual and visual domains.

In fact, uncertainty is by no means exclusive to TVR; it is a pervasive challenge in machine learning, encompassing both epistemic and aleatoric uncertainties [18]. To tackle such uncertainties, user-centric approaches such as active learning methods [25], interactive dialogue systems [47], and minimal human supervision [16] have shown remarkable effectiveness. Building on these insights, recent interactive TVR methods [17, 29, 38] also recognize the significance of uncertainty. These systems typically employ VideoQA models [27, 50] or large language models (LLMs) [19, 24, 49] to generate follow-up questions and simulate user responses, refining queries based on user feedback. However, a key limitation is that existing interactive methods do not explicitly model or quantify uncertainty, relying instead on heuristic or ad-hoc question generation strategies that may not optimally address uncertainties at play.

Motivated by this limitation, we propose a principled approach that directly addresses the uncertainty challenge in TVR. Rather than resorting to ad-hoc techniques, our method systematically quantifies three critical types of uncertainty—text ambiguity, mapping uncertainty, and frame uncertainty—by leveraging semantic entropy, JS divergence, and a novel Temporal Quality-based Frame Sampler (TQFS), respectively, all with *training-free* approaches. Building upon these quantified uncertainties, we further introduce the Uncertainty-Minimizing Interactive Text-to-Video Retrieval (UMIVR) framework, illustrated in Fig. 1(b). Specifically, UMIVR explicitly tracks and updates uncertainty scores (e.g., TAS for text ambiguity and MUS for mapping uncertainty) at each interaction round and adaptively generates clarifying questions tailored to progressively reducing these uncertainties. Through iterative and targeted user interactions, UMIVR effectively mitigates the negative impacts of ambiguous and noisy inputs, systematically refining query precision and ultimately enhancing retrieval accuracy.

We validate the effectiveness of our approach through extensive experiments on several challenging benchmarks. UMIVR consistently outperforms interactive baselines, notably surpassing the non-interactive leading methods by achieving a Hit@1 of 68.9% on MSR-VTT-1k after only 3 interaction rounds. Furthermore, extensive experiments also demonstrate substantial performance improvements across AVSD, MSVD, and ActivityNet datasets through iterative query clarifying. Beyond its empirical strength, UMIVR exhibits remarkable generalizability, as the proposed TQFS module can readily serve as a plug-in enhancement for existing TVR models, and the UMIVR architecture seamlessly extends to interactive text-to-image retrieval scenarios, underscoring its broad applicability across multimodal retrieval tasks.

In summary, our contributions are three-fold:

- We explicitly identify the uncertainty challenges in TVR and propose quantitative metrics tailored to distinct uncertainty types, thereby providing a more rigorous framework for understanding retrieval ambiguity;
- We introduce the UMIVR framework, which unifies video retrieval, captioning, and question answering into an integrated system that leverages uncertainty metrics to enhance query refinement;
- Through comprehensive experiments, we demonstrate the effectiveness and generalizability of our approach, setting new benchmarks and opening avenues for further research in interactive multimodal retrieval.

2. Related Work

2.1. Text-to-Video Retrieval

Text-to-Video Retrieval (TVR) aims at retrieving relevant video content given textual queries via cross-modal alignment. Early methods employed attention-based aggregation of multimodal features [14, 33], while subsequent approaches improved representation learning [8, 43]. Later, pretraining models [27, 46] advanced TVR by adapting pretrained image-text encoders to videos [35, 59] and refining alignment with fine-grained contrastive learning [34, 37] and auxiliary captioning tasks [55, 60, 61].

Recent studies also highlight the critical challenge posed by uncertainty in TVR. TAM [31] and UATVR [11] approached mapping uncertainty through adaptive visual prototypes and probabilistic embeddings, respectively; PAU [26] focused on modeling text ambiguity via evidential theory. Nevertheless, existing approaches consider only single uncertainty aspects, thus providing limited performance gains. In contrast, our approach explicitly identifies and systematically quantifies three key types of uncertainty (text ambiguity, mapping uncertainty, and frame uncertainty) within a unified, training-free interactive retrieval framework, significantly mitigating ambiguity and enhancing retrieval effectiveness.

2.2. Interactive Vision Retrieval

Interactive retrieval historically aimed at bridging the semantic gap by leveraging iterative user feedback. Early content-based image retrieval relied on low-level features and relevance feedback to incrementally refine user queries [48, 53]. Deep learning advances subsequently introduced reinforcement learning or zero-shot learning [5, 6] for adaptive question generation and query refinement [3, 9, 39]. More recently, LLMs such as ChatGPT [42] have greatly enhanced interactive retrieval by generating context-aware clarifying questions without additional training [17, 22, 23], achieving remarkable generalization. Despite their success, existing methods primarily rely on heuristic or context-driven question generation, overlooking explicit modeling and quantification of underlying uncertainties. In contrast, we propose UMIVR, an explicitly uncertainty-aware interactive retrieval framework that systematically quantifies and reduces multiple uncertainty types (text ambiguity, mapping uncertainty, frame uncertainty), improving retrieval robustness and accuracy.

3. Method

In this section, we present our proposed UMIVR framework, explicitly designed to address the three types of uncertainty identified in Sec. Specifically, we introduce principled metrics for quantifying *text ambiguity* (Sec. 3.1) and *mapping uncertainty* (Sec. 3.2), and propose a Temporal Quality-based Frame Sampler (TQFS, Sec. 3.3) to mitigate *frame-level uncertainty*. Finally, we integrate these components into a unified interactive retrieval pipeline (Sec. 3.4), which generates adaptive clarifying questions based on quantified uncertainties, enabling iterative refinement of user queries and enhancing retrieval precision.

3.1. Text Ambiguity Score via Semantic Entropy

Text ambiguity arises when queries are vague, incomplete, or permit multiple semantic interpretations. Unlike conventional token-level heuristics that often overestimate ambiguity by treating lexical variants separately, semantic entropy [13, 21] more accurately captures genuine semantic variability by analyzing distributions of meanings.

Motivated by this insight, we introduce a *Text Ambiguity Score* (TAS) to quantify the semantic uncertainty associated with textual queries. Specifically, we first apply a captioning model to videos in the retrieval database, resulting in a corpus of textual descriptions $\mathcal{C} = \{s_i\}_{i=1}^{N}$, each describing video i. These captions are encoded into normalized embeddings $\mathbf{e}_{s_i}$, which we store offline.

Given a query x, we compute its embedding $\mathbf{e}_x$ and retrieve the top-K most similar captions from $\mathcal{C}$. To reduce redundancy and avoid inflated entropy, we cluster these captions into M coherent groups $\{c_j\}_{j=1}^{M}$. For each cluster c_j, the aggregated probability is computed as $p(c_j \mid x) = \sum_{c \in c_j} \text{sim}(x, c) / \sum_{k=1}^{M} \sum_{c \in c_k} \text{sim}(x, c)$, reflecting its share of the total similarity mass. The semantic entropy is then defined as:

$$SE(x) = -\sum_{j=1}^{M} p(c_j \mid x) \log p(c_j \mid x), \tag{1}$$

which is further transformed into the final TAS value in $[0, 1]$ through a normalization function. This step can also incorporate additional adjustments, such as accounting for the structural complexity of the query, ensuring that more complex, well-specified queries are assigned lower TAS values. Higher TAS indicates greater semantic uncertainty, reflecting more ambiguous textual queries.

3.2. Mapping Uncertainty Score via JS Divergence

Mapping uncertainty arises when similarity scores between a text query and candidate videos lack a clear peak, leading to ambiguous mappings [11]. To quantify this uncertainty, we propose a Mapping Uncertainty Score (MUS) based on Jensen–Shannon (JS) divergence [40], measuring the deviation of the similarity distribution from an ideal, perfectly certain scenario.

Given top-k similarity scores between a query text and video candidates $[s_1, s_2, \ldots, s_k]$ (sorted in descending order), we first transform them into a normalized probability distribution p:

$$p_i = \frac{\max(s_i - \bar{s}, 0)^2}{\sum_{j=1}^{k} \max(s_j - \bar{s}, 0)^2}, \tag{2}$$

where $\bar{s}$ denotes the mean similarity score. This transformation emphasizes high-confidence candidates while suppressing low-confidence noise. If all scores fall below $\bar{s}$, we default to a uniform distribution.

Next, we define an ideal one-hot distribution q, representing complete certainty:

$$q_i = \begin{cases} 1, & \text{if } i = 1, \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

We then compute the JS divergence between the distributions p and q:

$$JSD(p \parallel q) = \frac{1}{2}\text{KL}(p \parallel m) + \frac{1}{2}\text{KL}(q \parallel m), \tag{4}$$

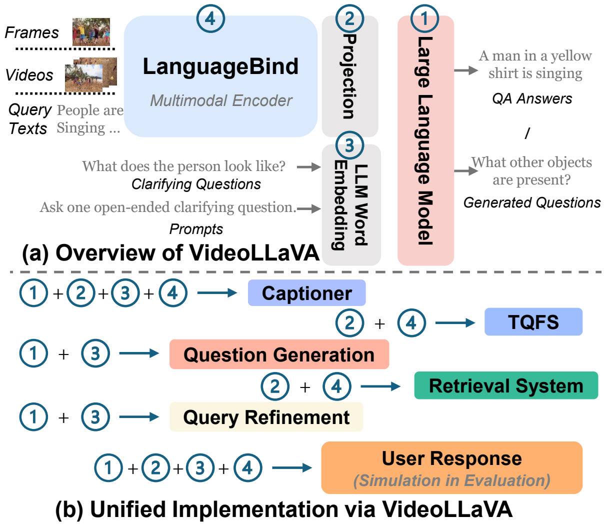

Frames
Videos
Query Texts
People are Singing ...
4
LanguageBind
Multimodal Encoder
2
Projection
3
LLM Word Embedding
1
Large Language Model
A man in a yellow shirt is singing
QA Answers
/
What other objects are present?
Generated Questions
What does the person look like?
Clarifying Questions
Ask one open-ended clarifying question.
Prompts
(a) Overview of VideoLLaVA
1+2+3+4 → Captioner
2 + 4 → TQFS
1 + 3 → Question Generation
2 + 4 → Retrieval System
1 + 3 → Query Refinement
1+2+3+4 → User Response (Simulation in Evaluation)
(b) Unified Implementation via VideoLLaVA

Figure 2. **A unified implementation of UMIVR with VideoLLaVA. (a)** VideoLLaVA integrates LanguageBind and LLM into a unified architecture, enabling simultaneous handling of video-text tasks. **(b)** A summary illustrating how UMIVR leverages the single, unified VideoLLaVA model to compactly realize all its core functionalities, significantly simplifying the system architecture compared to prior ensemble or hybrid approaches.

where $m = \frac{1}{2}(p+q)$ and $\mathrm{KL}(\cdot)$ indicates Kullback–Leibler divergence [7].

Finally, we normalize the JS divergence by its theoretical maximum $JSD_{\max}$ to obtain MUS within $[0, 1]$:

$$\mathrm{MUS}(x) = \frac{JSD(p \parallel q)}{JSD_{\max}}. \tag{5}$$

A higher MUS indicates greater ambiguity in the mapping between query and video candidates, while a lower MUS reflects a more confident retrieval scenario.

3.3. Temporal Quality-based Frame Sampler

Existing video-based methods, such as video retrieval [35, 55] and video QA models [27, 30], typically sample frames uniformly, which may inadvertently include low-quality frames affected by defocus or blur [57, 63]. To address this issue, we propose a plug-and-play Temporal Quality-based Frame Sampler (TQFS) that adaptively selects high-quality frames while ensuring sufficient temporal coverage.

Given a video V of length T and original frame rate r, we first uniformly sample frames at a reduced frame rate r', resulting in $N = \lfloor T \times r / r' \rfloor$ frames $\{F_1, \ldots, F_N\}$, each associated with a timestamp t_i. Next, we evaluate the visual clarity of each frame F_i using a no-reference image quality assessment (NR-IQA) algorithm $Q(\cdot)$, such as simple Laplacian-variance measures [45] or more advanced NR-IQA methods (e.g., BRISQUE[41]), assigning each frame a quality score $Q(F_i)$.

To maintain temporal coverage, TQFS divides the video into M uniform temporal bins. In each bin $\mathcal{I}_m$, we select the highest-quality frame:

$$F_m^* = \arg\max_{F_i \in \mathcal{I}_m} Q(F_i), \tag{6}$$

resulting in candidate frames $\{F_1^*, \ldots, F_M^*\}$.

To further reduce redundancy and ensure semantic diversity, we extract semantic embeddings $\phi(F_m^*)$ for each candidate frame, forming the embedding matrix $\Phi = [\phi(F_1^*), \ldots, \phi(F_M^*)]$. We then apply K-means clustering on these embeddings, selecting the highest-quality frame within each cluster. Finally, the selected frames are chronologically ordered, yielding the final K high-quality frames.

Overall, TQFS reduces frame-level uncertainty by emphasizing visually clear and semantically diverse frames, significantly enhancing the robustness of subsequent video retrieval tasks.

3.4. UMIVR: Uncertainty-Minimizing Interactive Text-to-Video Retrieval Framework

Fig. 2 and Fig. 3 illustrate the detailed architecture and workflow of our proposed UMIVR framework. UMIVR seamlessly integrates text-video retrieval, video captioning, and video question answering into a unified multimodal system, with two key innovations: (1) the adoption of a unified Video-LLM architecture for efficient multi-task integration, and (2) a principled uncertainty-minimizing interactive retrieval pipeline that adaptively generates clarifying questions based on explicitly quantified uncertainties.

Unified Video-LLM Architecture. Existing interactive TVR methods typically rely on either multi-model ensemble architectures (e.g., combining BLIP [27] for retrieval with T0++ [49] for dialogue generation) or hybrid local/cloud schemes (e.g., PlugIR [22] via ChatGPT [42]), both of which lead to substantial memory overhead or inference latency. To overcome these limitations, UMIVR leverages VideoLLaVA [30], a unified multimodal model that integrates LanguageBind [62]—a robust multimodal encoder—and a LLM via an efficient *align-before-projection* design, as depicted in Fig. 2(a). Specifically, visual inputs (frames and videos) and textual queries are first encoded by LanguageBind, projecting visual modality information into the shared language embedding space for accurate cross-modal alignment. In contrast, textual inputs intended for LLM generation tasks, such as clarifying questions and prompts, are directly encoded by the LLM's word embedding layer, enabling effective language generation and comprehension within the model. Fig. 2(b) summarizes this unified implementation, highlighting how UMIVR compactly realizes captioning, response simulation, clarifying question generation, and retrieval functionalities within a single model. This unified design not only eliminates cross-model

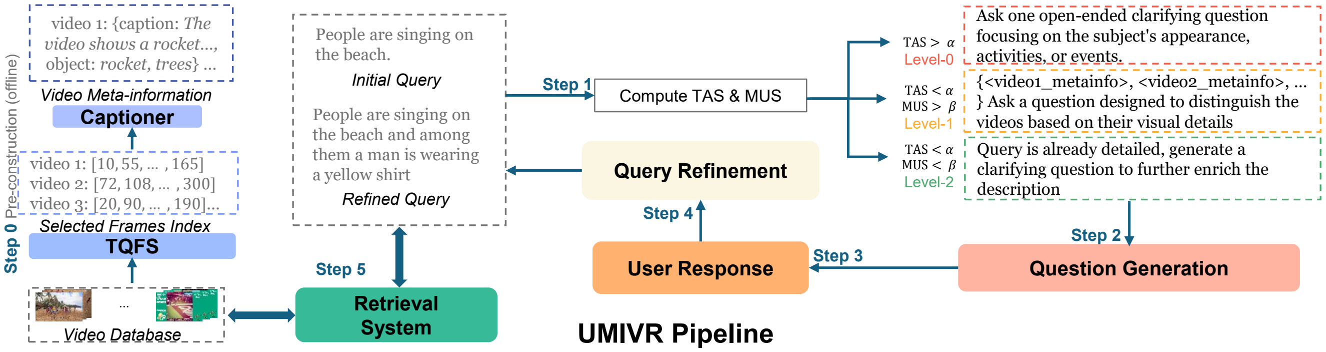

video 1: {caption: *The video shows a rocket...*, object: *rocket, trees*} ...
Video Meta-information
Captioner
Step 0 Pre-construction (offline)
video 1: [10, 55, ... , 165]
video 2: [72, 108, ... , 300]
video 3: [20, 90, ... , 190]...
Selected Frames Index
TQFS
...
Video Database
People are singing on the beach.
Initial Query
People are singing on the beach and among them a man is wearing a yellow shirt
Refined Query
Step 5
Retrieval System
Step 1
Compute TAS & MUS
TAS > α
Level-0
Ask one open-ended clarifying question focusing on the subject's appearance, activities, or events.
TAS < α
MUS > β
Level-1
{<video1_metainfo>, <video2_metainfo>, ... } Ask a question designed to distinguish the videos based on their visual details
TAS < α
MUS < β
Level-2
Query is already detailed, generate a clarifying question to further enrich the description
Query Refinement
Step 4
Step 2
User Response
Step 3
Question Generation
UMIVR Pipeline

Figure 3. **Pipeline of the UMIVR framework** Videos are first preprocessed offline (Step 0) by TQFS for high-quality frame selection and captioning to generate meta-information. Given an initial user query, UMIVR quantifies textual and mapping uncertainties (TAS & MUS, Step 1), adaptively generates clarifying questions at different uncertainty levels (Level-0, 1 and 2, Step 2), and iteratively refines queries based on user responses (Steps 3–4), ultimately retrieving the most relevant videos (Step 5).

compatibility issues but also significantly reduces memory usage compared to traditional ensemble-based approaches.

UMIVR Framework Pipeline. The overall pipeline of UMIVR is depicted in Fig. 3. Initially, UMIVR preprocesses the video database offline (Step 0), applying the Temporal Quality-based Frame Sampler (TQFS) to select temporally representative and visually high-quality frames. These selected frames are then passed through VideoLLaVA to generate textual meta-information—including video descriptions and salient object annotations—that will be stored and used during online interactions.

Upon receiving an initial textual query from a user, UMIVR first quantifies the query's uncertainty by computing two complementary uncertainty metrics (Step 1): TAS (Sec. 3.1) and MUS (Sec. 3.2). According to these scores, UMIVR adaptively generates clarifying questions at three uncertainty levels (Step 2), effectively guiding user interactions toward reducing ambiguity. Specifically, if TAS is high (greater than a predefined threshold α), indicating significant semantic ambiguity in the textual query, the framework instructs VideoLLaVA to generate an open-ended clarifying question, prompting the user for additional context about appearance, activities, or events. Conversely, if TAS is low but MUS remains high (greater than threshold β), indicating clearly expressed but visually indistinguishable queries, UMIVR leverages the retrieved candidate videos' meta-information (captions, objects, etc.) to generate targeted clarifying questions that explicitly distinguish visually similar videos. Finally, if both uncertainty measures are below their respective thresholds (low uncertainty), UMIVR generates enrichment-oriented questions merely to further enrich the query's descriptive power.

After generating the clarifying question in Step 2, UMIVR expects a user response to refine the query. This interaction can be conducted with real users or approximated for evaluation purposes using simulated responses derived from VideoQA modules (Step 3). The collected answer is then combined with the initial query via a standard query refinement strategy, yielding a more precise query (Step 4). This refined query significantly reduces uncertainties, enabling a focused and accurate video retrieval (Step 5).

This entire interactive retrieval process is inherently iterative and uncertainty-driven. Each subsequent interaction round benefits from progressively reduced uncertainty scores, systematically refining the query until the retrieved videos closely align with the user's refined intent. By explicitly quantifying uncertainty and adaptively generating clarifying questions, UMIVR robustly addresses ambiguity and noise inherent in text-to-video retrieval, significantly enhancing retrieval accuracy and user satisfaction.

4. Experiments

4.1. Dataset and Evaluation

Datasets We perform experiments on four established TVR datasets: **MSR-VTT** [54], featuring 10,000 short videos each with 20 captions, using the standard 1K evaluation split (we also conduct our ablation studies on this dataset); **AVSD** [1], providing dialogues grounded in videos, evaluated on a standard 1,000-sample test subset following prior works [29, 36, 38]; **MSVD** [4], consisting of around 2,000 videos annotated by multilingual captions, evaluated using the widely-adopted 670-video test subset; and **ActivityNet** [10], a large-scale dataset of untrimmed videos capturing 200 activity categories, evaluated on the commonly-used validation set (4,917 videos) [51, 55, 56].

Evaluation Metrics We adopt three metrics for comprehensive evaluation: **Recall**, **Hit@k**, and the recent **Best log Rank Integral (BRI)** [22] metric. *Recall* as a standard metric measures retrieval accuracy at fixed ranks. *Hit* commonly used in interactive retrieval, reflects whether the tar-

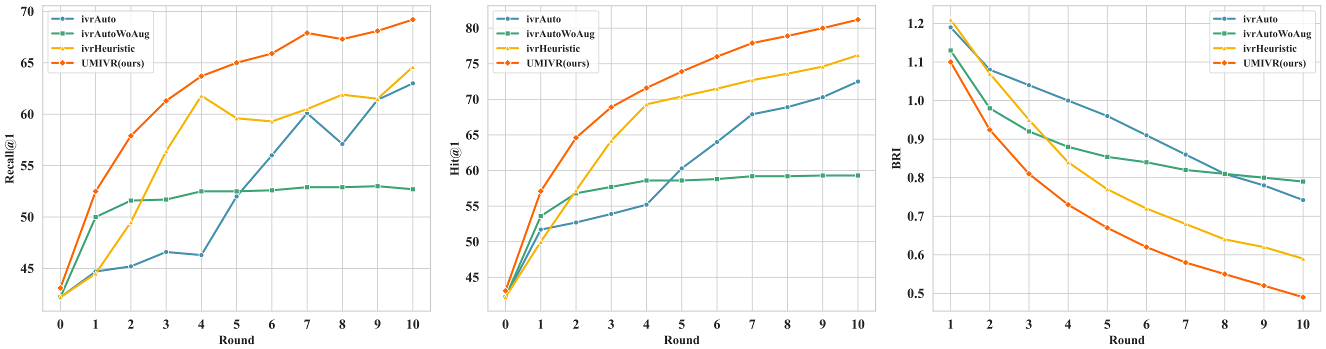


Figure 4. **Comparison of UMIVR with interactive baseline methods on MSR-VTT-1k across multiple interaction rounds.** From left to right, we illustrate Recall@1, Hit@1, and BRI scores, respectively. UMIVR consistently outperforms competing interactive baselines by achieving higher Recall@1 and Hit@1, as well as lower BRI values, highlighting its superior efficiency and effectiveness in iterative query refinement and retrieval accuracy improvement.

Methods \ Metrics	R@1 ↑	R@5 ↑	R@10 ↑	MnR ↓	Hit@1 ↑	Hit@10 ↑
			Non-interactive TVR			
CLIP4Clip [35]	44.5	71.4	81.6	15.3	44.5	81.6
ProST [28]	49.5	75.0	84.0	11.7	49.5	84.0
UCOFIA [52]	49.4	72.1	-	-	49.4	-
EERCF [51]	54.1	78.8	86.9	-	54.1	86.9
CLIP-ViP [55]	57.7	80.5	88.2	-	57.7	88.2
HunYuan(SOTA) [20]	62.9	84.5	90.8	9.3	62.9	90.8
			UMIVR (Interactive TVR)			
round 0	43.1	66.1	75.8	22.4	43.1	75.8
round 2	57.9	81.2	86.6	10.4	57.2	89.9
round 3	61.3	84.1	89.0	8.1	68.9	92.7
round 6	65.9	87.7	91.8	5.9	76.0	95.3
round 8	**67.3**	**88.3**	**92.8**	**5.7**	**78.9**	**96.5**

Table 1. **Comparison with non-interactive TVR methods on MSR-VTT-1k.** UMIVR significantly improves retrieval performance through multiple rounds of interaction. Notably, after at most three rounds of interaction, UMIVR surpasses most competing methods. Underlined values indicate the earliest interaction round where UMIVR exceeds existing approaches, while bold values denote the best overall performance across all methods.

Methods	Round	R@1 ↑	R@10 ↑	Hit@1 ↑	Hit@10 ↑	BRI ↓
D2V[36]	0	8.8	32.1	8.8	32.1	-
	2	22.9	54.0	-	-	-
	4	22.5	58.5	-	-	-
	6	23.9	61.0	-	-	-
VIRED [38]	3	24.9	60.8	-	-	-
ivrAuto	0	29.6	61.3	29.6	61.3	-
	2	30.7	62.2	35.9	66.9	1.74
	4	32.6	66.7	39.4	71.6	1.67
	6	37.3	74.3	45.1	78.9	1.57
ivrAutoWoAug	0	29.6	61.3	29.6	61.3	-
	2	34.6	67.2	39.4	71.6	1.64
	4	34.6	67.3	40.4	72.6	1.56
	6	34.6	67.2	40.8	72.8	1.53
ivrHeuristic	0	29.6	61.3	29.6	61.3	-
	2	35.2	69.2	41.6	73.5	1.61
	4	43.2	80.3	51.1	82.7	1.38
	6	43.4	80.8	53.5	86.2	1.23
UMIVR(ours)	0	30.6	61.6	30.6	61.6	-
	2	44.8	78.9	50.8	81.5	1.40
	4	47.9	81.4	58.8	86.7	1.16
	6	**49.9**	**82.2**	**63.3**	**88.2**	**1.02**

Table 2. Comparison results on AVSD dataset.

get appears in the top-k candidates at any interaction step.

BRI [22] is a new metric for interactive retrieval, integrating three essential aspects: (1) *user satisfaction*, checking if the target video is eventually retrieved; (2) *retrieval efficiency*, encouraging successful retrieval with fewer interactions; and (3) *ranking improvement significance*, rewarding substantial improvements at higher ranks. By synthesizing these factors into one unified score, BRI aligns better with real user interaction scenarios, enabling nuanced performance comparisons among interactive retrieval methods.

4.2. Implementation Details

Our implementation is based on the open-source VideoLLaVA codebase[1] with Python-3.10 and Pytorch-2.0.1. Specifically, we adopt VideoLLaVA-7B as our backbone model, leveraging 4-bit quantization to significantly reduce GPU memory consumption. To ensure consistency and reproducibility, we set the generation temperature to 0.1 for internal modules such as Captioner, Question Generation, and Query Refinement. Conversely, for the VideoQA module simulating diverse user responses, we set the generation temperature to 0.7. All other generation settings follow the default configurations provided by the codebase.

For visual encoding, we utilize the default LanguageBind encoder integrated within VideoLLaVA-7B, a post-pretrained CLIP ViT-L/14 model. This visual encoder processes RGB inputs with spatial dimensions of 224 × 224 and employs a patch size of 14 pixels, consisting of 24 transformer layers with temporal attention enabled. Each transformer layer features a hidden dimensionality of 1024 and 16 attention heads, providing high representational capabil-

[1] https://github.com/PKU-YuanGroup/Video-LLaVA

	Methods	MSVD				ActivityNet			
		0	1	3	5	0	1	3	5
R@1	ivrAuto	50.3	52.8	54.4	61.7	32.9	33.8	33.9	35.1
	ivrAutowoAug	50.3	55.5	58.9	59.2	32.9	36.3	36.4	36.5
	ivrHeuristic	50.3	53.4	64.6	67.3	32.9	34.0	37.8	39.9
	UMIVR(ours)	**51.9**	**61.2**	**67.0**	**69.7**	**33.1**	**38.3**	**40.4**	**41.8**
R@10	ivrAuto	85.0	87.7	88.2	92.8	73.4	74.5	74.7	76.4
	ivrAutowoAug	85.0	88.9	89.8	90.4	73.4	77.0	77.2	77.3
	ivrHeuristic	85.0	87.1	90.5	91.3	73.4	76.1	79.2	80.5
	UMIVR(ours)	**86.4**	**90.9**	**93.9**	**94.8**	**73.7**	**78.2**	**80.0**	**80.7**
Hit@1	ivrAuto	50.3	57.7	59.8	68.5	32.9	36.9	37.3	39.7
	ivrAutowoAug	50.3	61.0	65.6	66.1	32.9	39.2	40.5	40.6
	ivrHeuristic	50.3	58.0	70.8	77.0	32.9	35.5	41.5	44.8
	UMIVR(ours)	**51.9**	**65.4**	**74.6**	**79.3**	**33.1**	**42.0**	**47.8**	**50.1**
Hit@10	ivrAuto	85.0	91.3	91.7	95.2	73.4	77.3	77.7	79.6
	ivrAutowoAug	85.0	91.4	92.3	92.9	73.4	79.1	79.8	79.9
	ivrHeuristic	85.0	89.5	93.1	94.5	73.4	77.2	81.3	83.0
	UMIVR(ours)	**86.4**	**92.5**	**95.5**	**96.7**	**73.7**	**79.9**	**83.0**	**84.1**
BRI	ivrAuto	-	0.85	0.75	0.68	-	1.40	1.35	1.32
	ivrAutowoAug	-	0.82	0.68	0.64	-	1.37	1.28	1.26
	ivrHeuristic	-	0.86	0.64	0.51	-	1.42	1.30	1.22
	UMIVR(ours)	-	**0.78**	**0.58**	**0.49**	-	**1.35**	**1.20**	**1.13**

Table 3. Comparison results on MSVD and ActivityNet Dataset.

ity for both spatial and temporal information. The encoder supports video inputs of 8 frames and projects final vision embeddings to a dimensionality of 768.

Baselines. For baseline comparisons, we primarily compete with IVR [29], a recent influential approach in interactive text-to-video retrieval that introduces an LLM-based training-free paradigm and demonstrates strong performance on MSR-VTT, MSVD, and AVSD datasets. IVR encompasses three variants: *ivrHeuristic* (using manually-defined question templates), *ivrAuto* (generating questions from top-k similar video captions assisted by heuristic-based augmentation), and *ivrAutoWoAug* (identical to *ivrAuto* but without heuristic augmentation). To ensure fair and rigorous comparisons, we faithfully reproduce these three IVR variants within our UMIVR framework, integrating the IVR codebase[2] into the Video-LLaVA system. Additionally, since IVR's original implementation generates multiple questions concurrently without iterative interaction, we adapt it to strictly follow standard interactive retrieval conventions [17, 22], allowing only one question-answer exchange per interaction round. After this adaptation, our IVR baselines and UMIVR framework are fully aligned and directly comparable. Consistent with prior interactive retrieval studies and realistic application scenarios, we limit the maximum number of interaction rounds to 10, as exceeding this typically results in poor user experience.

4.3. Comparison Results

Fig. 4 and Tab. 1 summarize retrieval performance on the MSR-VTT dataset. UMIVR consistently surpasses interactive baselines across interaction rounds, demonstrating clear advantages in retrieval accuracy and efficiency. Moreover, UMIVR quickly outperforms the leading non-interactive method (HunYuan_tvr [20]). Specifically, after only three rounds, UMIVR already surpasses the HunYuan_tvr in Hit@1 (68.9 vs. 62.9), highlighting its effectiveness in leveraging iterative interactions to explicitly address retrieval uncertainty.

Comp.				
	+TAS	✔	✔	✔
	+MUS		✔	✔
	+TQFS			✔
R@1	1	51.6	52.2	**52.5**
	3	61.0	**62.1**	61.3
	5	63.4	64.2	**65.0**
Hit@1	1	56.6	**57.4**	57.1
	3	67.0	68.6	**68.9**
	5	73.0	72.8	**73.9**
Hit@10	1	86.1	86.4	**86.7**
	3	91.1	92.5	**92.7**
	5	93.7	94.4	**94.8**
BRI	5	0.69	**0.67**	**0.67**

Table 4. Ablation study on different components.

Tab. 2 and Tab. 3 further confirm UMIVR's robust performance and generalization capabilities. Specifically, UMIVR exhibits substantial improvements over baselines on AVSD, MSVD, and ActivityNet datasets, maintaining consistent superiority across metrics and interaction rounds. These results collectively validate the effectiveness of explicitly quantifying and minimizing uncertainty in interactive text-to-video retrieval scenarios.

4.4. Ablation Study

We perform comprehensive ablation analyses to investigate the contribution of each uncertainty-related component in UMIVR and examine the sensitivity of threshold parameters. Tab. 4 explores the individual impact of Text Ambiguity Score (TAS), Mapping Uncertainty Score (MUS), and Temporal Quality-based Frame Sampler (TQFS). Clearly, integrating each component incrementally improves retrieval performance, and their joint usage delivers the best results across metrics. Particularly, MUS and TAS demon-

(α, β)	R@1 ↑			Hit@1 ↑			Hit@10 ↑			BRI ↓
	1	3	5	1	3	5	1	3	5	5
(0.4, 0.1)	52.3	55.9	61.9	57.0	65.1	71.5	85.8	91.3	93.7	0.73
(0.5, 0.1)	51.9	57.9	63.1	56.4	66.1	72.4	86.2	92.6	94.6	0.70
(0.6, 0.1)	51.5	58.4	64.6	55.8	65.1	70.4	85.1	91.6	**94.8**	0.68
(0.4, 0.2)	52.5	60.4	63.9	57.0	67.6	73.4	86.2	91.3	93.9	0.69
(0.5, 0.2)	52.5	**61.3**	**65.0**	**57.1**	68.9	**73.9**	**86.7**	92.7	**94.8**	**0.67**
(0.6, 0.2)	51.9	61.1	64.7	56.6	68.5	73.1	85.9	**92.9**	94.5	0.68
(0.4, 0.3)	51.5	58.9	64.9	56.8	67.1	73.5	86.6	91.8	94.4	0.69
(0.5, 0.3)	**52.6**	**61.3**	64.6	**57.1**	**70.0**	73.6	86.0	92.8	94.4	0.68
(0.6, 0.3)	52.2	61.1	64.6	56.7	68.0	73.5	86.4	92.5	94.2	0.68

Table 5. Grid search for TAS threshold α and MUS threshold β.

[2] https://github.com/kevinliang888/IVR-QA-baselines

	Text-to-Video Retrieval			*Video-to-Text Retrieval*		
Methods	**R1** ↑	**R10** ↑	**MnR** ↓	**R1** ↑	**R10** ↑	**MnR** ↓
VideoCLIP [2]	30.7	71.0	18.1	30.2	70.4	23.0
[2] + **TQFS**	**31.1**(+0.4)	**72.4**(+1.4)	**16.1**(-2.0)	**31.1**(+0.9)	**72.4**(+2.0)	**21.0**(-2.0)
Xpool [15]	45.3	80.2	15.9	43.0	**83.2**	10.5
[15] + **TQFS**	**45.7**(+0.4)	**80.9**(+0.7)	**14.3**(-1.6)	**43.7**(+0.7)	82.7(-0.5)	**9.2**(-1.3)

Table 6. TQFS enhances performance as a plug-in module.

strate complementary roles, jointly addressing uncertainties in textual ambiguity and text-video mapping, while TQFS further boosts performance by reducing visual noise.

To understand parameter sensitivity, Tab. 5 presents grid search results for threshold parameters α (TAS) and β (MUS). Optimal retrieval performance emerges at $(\alpha, \beta) = (0.5, 0.2)$, consistently achieving highest Recall@1, Hit@1, Hit@10, and lowest BRI. Deviating from these optimal values results in noticeable performance drops, validating the importance of jointly tuning these thresholds. These analyses confirm that UMIVR effectively leverages explicit uncertainty modeling, and carefully calibrated thresholds significantly enhance retrieval accuracy.

4.5. Generalization

We further investigate the generalization capability of UMIVR and its components across broader scenarios and modalities. Tab. 6 demonstrates that our Temporal Quality-based Frame Sampler (TQFS) effectively serves as a plug-and-play module. When directly integrated into trained single-round TVR methods (VideoCLIP [2] and Xpool [15]), TQFS achieves consistent and meaningful improvements on MSR-VTT-1ka (e.g., gains of +0.4% to +1.4% in Recall@10 and reductions in MnR), confirming its versatility and effectiveness in enhancing existing retrieval methods without additional fine-tuning.

Furthermore, leveraging the multimodal capabilities of VideoLLaVA and LanguageBind, we extend UMIVR to interactive text-to-image retrieval. We evaluate UMIVR on the Visual Dialog (VisDial) [44] dataset, which includes 8,000 test images and corresponding captions from human-annotated dialogues based on MS-COCO [32]. As shown in Tab. 7, UMIVR consistently outperforms competitive interactive retrieval methods (ChatIR [23], PlugIR [22], and ivrHeuristic) across interaction rounds. Specifically, UMIVR achieves noticeable improvements in Recall@10 and Hit@10, highlighting its robustness and generalizability beyond video retrieval tasks.

	Round	0	2	4	6	8	10
R@10	ChatIR[23]	71.5	74.9	76	78.2	78.8	79.5
	PlugIR[22]	71.1	75.5	76.1	76.1	75.1	74.3
	ivrHeuristic	**73.2**	78.7	76.7	78.7	77.8	80.3
	UMIVR(ours)	**73.2**	**83.0**	**83.6**	**84.5**	**85.1**	**85.0**
Hit@10	ChatIR[23]	71.5	78.9	82.0	84.4	85.6	86.4
	PlugIR[22]	71.1	83.1	87.6	89.4	90.7	91.5
	ivrHeuristic	**73.2**	79.5	84.1	85.7	86.2	86.8
	UMIVR(ours)	**73.2**	**87.0**	**89.1**	**90.4**	**91.3**	**91.8**

Table 7. UMIVR achieves competitive improvements in interactive text-to-image retrieval on VisDial dataset.

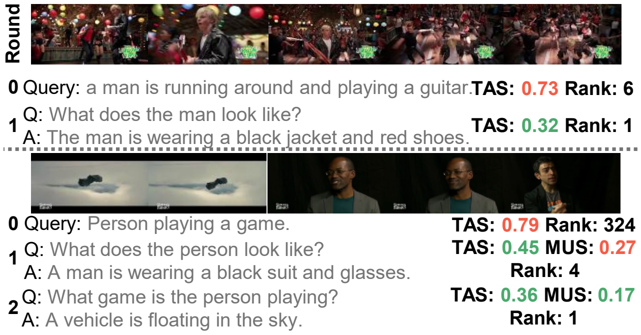

Round
0 Query: a man is running around and playing a guitar.
TAS: 0.73 Rank: 6
1 Q: What does the man look like?
A: The man is wearing a black jacket and red shoes.
TAS: 0.32 Rank: 1
0 Query: Person playing a game.
TAS: 0.79 Rank: 324
1 Q: What does the person look like?
A: A man is wearing a black suit and glasses.
TAS: 0.45 MUS: 0.27
Rank: 4
2 Q: What game is the person playing?
A: A vehicle is floating in the sky.
TAS: 0.36 MUS: 0.17
Rank: 1

Figure 5. **Case study of UMIVR's interactive retrieval process.** The examples illustrate how uncertainty-aware question generation progressively refines ambiguous queries, reducing the Text Ambiguity Score (TAS) and Mapping Uncertainty Score (MUS) while improving retrieval rank.

4.6. Case Study

Fig. 5 illustrates two qualitative examples demonstrating UMIVR's interactive retrieval process. Initially, queries contain significant ambiguity, resulting in high TAS and low rankings. Through uncertainty-minimizing clarifying questions, UMIVR progressively refines the queries by explicitly addressing textual and mapping uncertainties. This refinement effectively reduces TAS and MUS, substantially improving retrieval ranks after each interaction round. These examples highlight UMIVR's effectiveness in adaptively resolving uncertainty to achieve precise retrieval.

5. Conclusion

In this paper, we introduced UMIVR, an uncertainty-aware interactive framework for text-to-video retrieval that systematically quantifies and minimizes three fundamental uncertainties—text ambiguity, mapping uncertainty, and frame uncertainty. By proposing principled, training-free uncertainty metrics, UMIVR adaptively generates clarifying questions, iteratively refining user intent and significantly enhancing retrieval accuracy. Extensive experiments on benchmarks including MSR-VTT-1k demonstrated that UMIVR surpasses prior interactive and non-interactive methods, highlighting its effectiveness and generalizability. Our work thus establishes a robust uncertainty-minimizing foundation for interactive multimodal retrieval, opening promising directions for future research in uncertainty-aware interactive learning across vision-language tasks.

Acknowledgements

This work is supported by Australian Research Council (ARC) Discovery Project DP230101753, and CSIRO's Science Leader Project R-91559.

References

[1] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019. 5

[2] AskVideos. Askvideos-videoclip: Language-grounded video embeddings. GitHub, 2024. 8

[3] Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, and Xing Sun. Ask&confirm: Active detail enriching for cross-modal retrieval with partial query. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1815–1824, 2021. 3

[4] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, Portland, OR, 2011. 5

[5] Zhi Chen, Yadan Luo, Ruihong Qiu, Sen Wang, Zi Huang, Jingjing Li, and Zheng Zhang. Semantics disentangling for generalized zero-shot learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[6] Zhi Chen, Zecheng Zhao, Jingcai Guo, Jingjing Li, and Zi Huang. Svip: Semantically contextualized visual patches for zero-shot learning. In *ICCV2025*, 2025. 3

[7] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999. 4

[8] Ioana Croitoru, Simion-Vlad Bogolin, Yang Liu, Samuel Albanie, Marius Leordeanu, Hailin Jin, and Andrew Zisserman. Teachtext: Crossmodal generalized distillation for text-video retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11563–11573, 2021. 2

[9] Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979, 2017. 3

[10] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 5

[11] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. UATVR: uncertainty-adaptive text-video retrieval. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 13677–13687. IEEE, 2023. 2, 3

[12] Han Fang, Zhifei Yang, Xianghao Zang, Chao Ban, and Hao Sun. Mask to reconstruct: Cooperative semantics completion for video-text retrieval. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 2

[13] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nat.*, 630(8017):625–630, 2024. 3

[14] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part IV*, pages 214–229. Springer, 2020. 1, 2

[15] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4996–5005, 2022. 8

[16] Shubham Gupta, Peter W. J. Staar, and Christian de Sainte Marie. Clustering items from adaptively collected inconsistent feedback. In *International Conference on Artificial Intelligence and Statistics*, 2024. 2

[17] Donghoon Han, Eunhwan Park, Gisang Lee, Adam Lee, and Nojun Kwak. MERLIN: multimodal embedding refinement via llm-based iterative navigation for text-video retrieval-rerank pipeline. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*, pages 547–562. Association for Computational Linguistics, 2024. 2, 3, 7

[18] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110(3):457–506, 2021. 2

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[20] Jie Jiang, Shaobo Min, Weijie Kong, Hongfa Wang, Zhifeng Li, and Wei Liu. Tencent text-video retrieval: Hierarchical cross-modal interactions with multi-level representations. *IEEE Access*, 2022. 6, 7

[21] 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. 3

[22] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 791–809. Association for Computational Linguistics, 2024. 3, 4, 5, 6, 7, 8

[23] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. In *Neural Information Processing Systems*, 2023. 3, 8

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoy-

anov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020. 2

[25] Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. A survey on deep active learning: Recent advances and new frontiers. *CoRR*, abs/2405.00334, 2024. 2

[26] Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. *ArXiv*, abs/2309.17093, 2023. 2

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, pages 12888–12900. PMLR, 2022. 1, 2, 4

[28] Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, and Yongdong Zhang. Progressive spatio-temporal prototype matching for text-video retrieval. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4077–4087. IEEE, 2023. 6

[29] 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. 2, 5, 7

[30] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5971–5984. Association for Computational Linguistics, 2024. 4

[31] Chengzhi Lin, Ancong Wu, Junwei Liang, Jun Zhang, Wenhang Ge, Wei-Shi Zheng, and Chunhua Shen. Text-adaptive multiple visual prototype matching for video-text retrieval. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. 2

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 8

[33] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 279. BMVA Press, 2019. 1, 2

[34] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, pages 319–335. Springer, 2022. 2

[35] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of CLIP for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2, 4, 6

[36] Chenyang Lyu, Manh-Duy Nguyen, Van-Tu Ninh, Liting Zhou, Cathal Gurrin, and Jennifer Foster. Dialogue-to-video retrieval. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part II*, pages 493–501. Springer, 2023. 5, 6

[37] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP: end-to-end multi-grained contrastive learning for video-text retrieval. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 638–647. ACM, 2022. 2

[38] Avinash Madasu, Junier Oliva, and Gedas Bertasius. Learning to retrieve videos by asking questions. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 356–365. ACM, 2022. 2, 5, 6

[39] Avinash Madasu, Junier B. Oliva, and Gedas Bertasius. Learning to retrieve videos by asking questions. *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 3

[40] María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997. 3

[41] 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

[42] OpenAI. ChatGPT. `https://openai.com/blog/chatgpt/`, 2023. 3, 4

[43] Mandela Patrick, Po-Yao Huang, Yuki Markus Asano, Florian Metze, Alexander G. Hauptmann, João F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[44] Badri N. Patro, Shivansh Patel, and Vinay P. Namboodiri. Granular multimodal attention networks for visual dialog. *ArXiv*, abs/1910.05728, 2019. 8

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

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[47] Hossein A. Rahmani, Xi Wang, Mohammad Aliannejadi, Mohammadmehdi Naghiaei, and Emine Yilmaz. Clarifying

the path to user satisfaction: An investigation into clarification usefulness. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, pages 1266–1277. Association for Computational Linguistics, 2024. 2

[48] Yong Rui, Thomas S. Huang, Michael Ortega-Binderberger, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 8:644–655, 1998. 3

[49] Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, et al. Multitask prompted training enables zero-shot task generalization, 2021. 2, 4

[50] Yuqing Song, Shizhe Chen, and Qin Jin. Towards diverse paragraph captioning for untrimmed videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11245–11254. Computer Vision Foundation / IEEE, 2021. 2

[51] Kaibin Tian, Yanhua Cheng, Yi Liu, Xinglin Hou, Quan Chen, and Han Li. Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. In *AAAI Conference on Artificial Intelligence*, 2024. 5, 6

[52] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2804–2815. IEEE, 2023. 2, 6

[53] Hong Wu, Hanqing Lu, and Songde Ma. Willhunter: Interactive image retrieval with multilevel relevance measurement. In *International Conference on Pattern Recognition*, 2004. 3

[54] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society, 2016. 5

[55] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Rui Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language alignment. In *International Conference on Learning Representations*, 2022. 2, 4, 5, 6

[56] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *European Conference on Computer Vision*, 2018. 5

[57] Bingqing Zhang, Sen Wang, Yifan Liu, Brano Kusy, Xue Li, and Jiajun Liu. Object detection difficulty: Suppressing over-aggregation for faster and better video object detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1768–1778, 2023. 4

[58] Bingqing Zhang, Zhuo Cao, Heming Du, Xin Yu, Xue Li, Jiajun Liu, and Sen Wang. Tokenbinder: Text-video retrieval with one-to-many alignment paradigm. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 4957–4967, 2025. 2

[59] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 970–981. ACM, 2022. 2

[60] Zecheng Zhao, Zhi Chen, Zi Huang, Shazia Sadiq, and Tong Chen. Continual text-to-video retrieval with frame fusion and task-aware routing. In *SIGIR 2025*, 2025. 2

[61] Zecheng Zhao, Selena Song, Tong Chen, Zhi Chen, Shazia Sadiq, and Yadan Luo. Are synthetic videos useful? a benchmark for retrieval-centric evaluation of synthetic videos. *arXiv preprint arXiv:2507.02316*, 2025. 2

[62] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 4

[63] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-guided feature aggregation for video object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 408–417. IEEE Computer Society, 2017. 4