IVCR-200K: A LARGE-SCALE MULTI-TURN DIALOGUE BENCHMARK FOR INTERACTIVE VIDEO CORPUS RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, significant developments have been made in both video retrieval and video moment retrieval tasks, which respectively retrieve complete videos or moments for a given text query. These advancements have greatly improved user satisfaction during the search process. However, previous work has failed to establish meaningful **"interaction"** between the retrieval system and the user, and its one-way retrieval paradigm can no longer fully meet the personalization and dynamics needs of at least 80.8% of users.

In this paper, we introduce a more realistic setting, the Interactive Video Corpus Retrieval task (IVCR) that enables multi-turn, conversational, realistic interactions between the user and the retrieval system. To facilitate research on this challenging task, we introduce IVCR-200K, a bilingual, multi-turn, conversational, abstract semantic high-quality dataset that supports video retrieval and even moment retrieval. Furthermore, we propose a comprehensive framework based on multi-modal large language models (MLLMs) to support users' several interaction modes with more explainable solutions. Our extensive experiments demonstrate the effectiveness of our dataset and framework. The datasets, codes and leaderboards are available at: `https://ivcr200k.github.io/IVCR`.

1 INTRODUCTION

With the rapid proliferation of video platforms such as YouTube and TikTok, an ever-increasing number of videos are being produced every day, underscoring the significance of the video retrieval task in the multi-modal field (Yan et al., 2023; Zhang et al., 2023; Zeng et al., 2021). Typically, users employ descriptive sentences, and the retrieval system (Xu et al., 2016; Luo et al., 2022) sorts by matching textual descriptions and visual videos, ultimately returning the user's preferred videos, as depicted in Figure 1(a). At a more granular level, as shown in Figure 1(b), researchers have proposed the video moment retrieval task (Gao et al., 2017; Zeng, 2022), which utilizes textual descriptions to retrieve a small moment within the complete video. These tasks significantly enhance user satisfaction during the search process.

However, the majority of video retrieval systems operate in a "one-way" manner, which may not fully cater to the personalized and dynamic preferences of users. This "one-way" approach inhibits user interaction with the system, resulting in every request from the user needing to be rewritten. In fact, it is a common phenomenon that users desire **"multi-turn interaction"** with systems. To delve deeper into this phenomenon, we devised a questionnaire[1] regarding user search behavior, depicted in Figure 2. A striking 80.8% of respondents expressed a preference for interactive search functionality. Similarly, within the ShareGPT[2] conversation dataset, the average interaction round between users and the chat system stands remarkably high at 7.27. Moreover, our questionnaire indicate that interactive demands exhibit intricate behavioral patterns, as illustrated in Figure 1(c): 1) Long2Short: Keep looking for clips within the long videos that have already been scanned. 2) Short2Long: Search full-length videos based on known short videos. 3) Analogous: When the user inputs "I would like to watch a movie similar to this clip", the system should be able to provide a video with similar content.

[1] Details of this questionnaire can be found in supplementary material A.
[2] https://sharegpt.com/

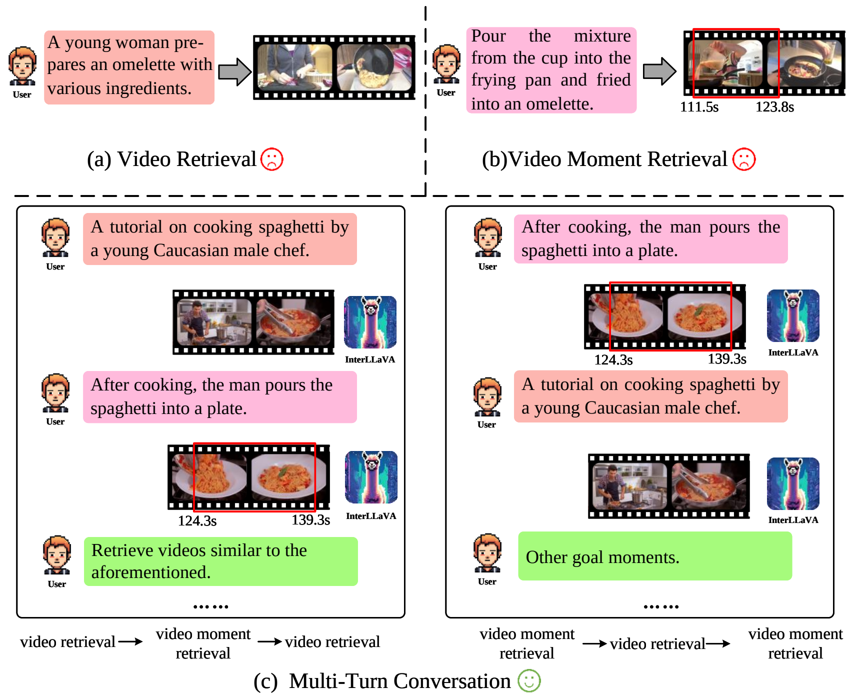
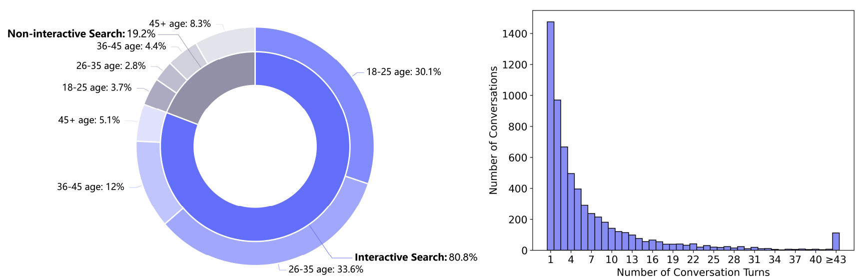


Figure 1: Visualization of the video retrieval, moment retrieval and our interactive retrieval.



(a) (b)

Figure 2: Investigation of User Search Behavior Feedback and interaction turns in ShareGPT. Users demonstrate a pronounced inclination towards interactive search and harbor high expectations regarding interaction rounds.

Therefore, drawing from these observations, we believe that the implementation of an interactive retrieval system holds significant value (Ma & Ngo, 2022; Maeoki et al., 2020), despite the challenges of complex user behaviors. Through multi-turn interaction with users, the system can adapt to individual preferences, furnishing more personalized retrieval outcomes. However, researchers have yet to delve deeply into this practical issue, one that resonates more closely with users' perspectives.

Formally, we introduce the Interactive Video Corpus Retrieval task (IVCR) for the first time. We define the "interactive" as meeting the following requirements: **1) Multi-turn.** This multi-turn interaction will extend the connection between the user and the search system. This process includes several interaction modes, such as video retrieval-only, moment retrieval-only, video-first-then-moment, moment-first-then-video, or creating a new topic for retrieval. **2) Free dialogue.** Users perform queries in natural language (Alayrac et al., 2022; Dai et al., 2024), and the retrieval system should explain the returned results in natural language form, which is more explainable and user-friendly. Furthermore, existing multi-modal retrieval datasets mostly contain low-level descriptive descriptions (e.g., "There are three dogs on the green lawn"), which do not align with the high-level abstract semantics used by users in real scenes (e.g., "Kung Fu movie where men and women fight"). **3) Real interaction.** The pioneers create simulated environments to generate interactive data (Ma

Table 1: **Comparison of IVCR-200K and other existing video-language datasets.**

Dataset	Multi-turn	Dialogue	Real interaction	Videos	Queries	Language
MSR-VTT(Xu et al., 2016)	✗	✗	✗	7,180	200K	English
MSVD(Chen & Dolan, 2011)	✗	✗	✗	1,790	70K	English
LSMDC(Rohrbach et al., 2017)	✗	✗	✗	200	118K	English
ActivityNet(Krishna et al., 2017)	✗	✗	✗	20,000	100K	English
VATEX(Wang et al., 2019)	✗	✗	✗	41,250	825K	English, Chinese
HowTo100M(Miech et al., 2019)	✗	✗	✗	1.221M	136M	English
Charades-STA(Gao et al., 2017)	✗	✗	✗	6,670	16,128	English
DiDeMo(Anne Hendricks et al., 2017)	✗	✗	✗	10,464	41K	English
TVQA(Lei et al., 2018)	✗	✓	✗	21,793	152,545	English
AVSD(Alamri et al., 2019)	✓	✓	✗	11,816	118,160	English
IVCR-200K (Ours)	✓	✓	✓	12,516	193,434	English, Chinese

& Ngo, 2022), but we emphasize that only truly understanding users can optimize a better search experience.

Unfortunately, at present, there is no available dataset or reliable framework to support this task of interactive video corpus retrieval, as shown in Table 1. **1) Dataset.** Existing video retrieval datasets are inadequate for multi-turn interaction scenarios, such as ActivityNet (Krishna et al., 2017) and DiDeMo (Anne Hendricks et al., 2017), which are single-turn datasets. Therefore, we propose an innovative interactive retrieval dataset, IVCR-200K, which is a bilingual, multi-turn, conversational, and abstract semantic high-quality dataset designed to support video retrieval and even moment retrieval. **2) Framework.** Existing retrieval methods are clearly insufficient for this conversational scenarios. For instance, solutions like CLIP (Luo et al., 2022; Fang et al., 2021) and 2D-TAN (Zhang et al., 2020) are discriminative models that cannot perform dialogue generation. Inspired by recent advances in multi-modal large language models (Li et al., 2023a; Ren et al., 2023), we combine their multi-turn dialogue, semantic understanding, and other capabilities to support users' interaction modes with a more explainable solution, named InterLLaVA. Extensive experiments demonstrate the effectiveness of our dataset and framework. We will release the code and dataset in the hope of contributing to the advance future research on real-world retrieval field.

The main contributions are summarized as follows: i)-To the best of our knowledge, this is the first work to introduce the "interactive" video corpus retrieval task (IVCR) , which effectively aligns users' multi-turn behavior in real-world scenarios and significantly enhances user experience. ii)-We introduce a dataset and an accompanying framework. Notably, the IVCR-200K dataset is a high-quality, bilingual, multi-turn, conversational, and abstract semantic dataset designed to support video and moment retrieval. The InterLLaVA framework leverages multi-modal large language models (MLLMs) to enable multi-turn dialogue experiences between users and the retrieval system.

2 Related Work

Video Retrieval Dataset

In recent years, with the vigorous development of the digital video new media market and continuous technological innovation, the scale of datasets related to video retrieval has rapidly expanded. For example, Xu et al.(Xu et al., 2016) constructed a video understanding dataset MSR-VTT, which contains 10K clips and 20K different text descriptions corresponding to various categories. MSVD(Chen & Dolan, 2011) is also a dataset widely used in video retrieval, which contains 1,970 videos, and each video has approximately about 40 associated sentences. Rohrbach et al.(Rohrbach et al., 2017) built the LSMDC, with 200 movies and 128,118 sentences, which is widely used in cross-model retrieval between video and text. Krishna et al.(Krishna et al., 2017) built a large-scale dataset ActivityNet Captions for dense captioning events, which contains 20k videos and a total of 100k descriptions, each with its unique start and end times. In comparison, Howto100M(Miech et al., 2019) contains more than 23k different visual tasks and 136 million video clips from 1.22M instructional web videos with narration, which is the largest video retrieval dataset. Wang et al.(Wang et al., 2019) constructed a large-scale multilingual video description dataset VATEX, which contains over 41,250 videos along with 825,000 captions in both English and Chinese. Gao et al.(Gao et al., 2017) built a dataset called Charades-STA, which augments the existing Charades (Sigurdsson et al., 2016) dataset by adding sentence temporal annotations for temporal activity localization via language. However, these datasets are mainly built to support video retrieval or video moment retrieval research rather than interactive video corpus retrieval, so they do not meet the personalized and dynamic retrieval needs of users. TVQA(Lei et al., 2018) is a large-scale video QA dataset based on six popular TV shows. It contains 152,545 QA pairs from 21,793 clips, spanning over 460 hours of video. AVSD(Alamri et al., 2019) is the only dataset for interactive video retrieval, which was created by adding dialogue data to the existing video dataset called Charades. Each video is associated with a 10-round dialogue discussing the content of the corresponding video. However, their annotations of 10-round dialogues are limited to each video, so they cannot be used for interactive video corpus retrieval.

In this paper, we built IVCR-200K dataset with 12K videos and more than 200K sentences covering 36 categories. To our best knowledge, IVCR-200K is the first and the largest video dataset for interactive video corpus retrieval. Dataset is a key step in developing deep learning based methods. We hope our dataset can inspire more efforts for the task of interactive video corpus retrieval.

Video Retrieval. Recently, numerous video datasets have been released for various video-language understanding tasks. In Table 1, we present a statistical comparison of our IVCR-200K dataset with ten video datasets used for video retrieval tasks. Video retrieval aims to retrieve relevant videos from a set of video candidates given a text query (Smeaton et al., 2006). Researchers have developed some pre-training systems (Luo et al., 2022; Fang et al., 2021; Gorti et al., 2022; Liu et al., 2022b). As an extension of video retrieval, video moment retrieval task aims to identify specific clips or moments within a video based on a given textual query (Gao et al., 2017; He et al., 2019). These studies have enhanced the service capabilities of the retrieval system. However, further development is required to meet the multi-turn interactive needs of users.

Interactive Retrieval. The concept of interactive retrieval has long been proposed in the context of combining human-machine learning techniques for multimedia content search (Thomee & Lew, 2012; Snoek et al., 2008). Currently, only a few works (Madasu et al., 2022; Maeoki et al., 2020; Ma & Ngo, 2022; Liang & Albanie, 2023) have explored this task. For example, Madasu et al.(Madasu et al., 2022) and Maeoki et al.(Maeoki et al., 2020)adopt a dialogue-based approach, utilizing a series of video-related questions and answers generated by different models as retrieval queries. Furthermore, Ma et al.(Ma & Ngo, 2022)develop a user simulation for intelligent multimedia applications to enable precise video segment search through human-computer interaction. The technical challenges in modeling multi-turn dialogue retrieval have contributed to the slow development in this direction.

Large language Models. With the breakthroughs in generative artificial intelligence, the way humans interact with machines has changed(Min et al., 2023; Zheng et al., 2024). Researchers have extended large language models to the visual perception domain, developing a series of large language models with multimodal information processing capabilities, such as Flamingo(Alayrac et al., 2022), BLIP-2(Li et al., 2023a), and LLaVA(Liu et al., 2024) for image processing, and Sora, Video LLaMA(Zhang et al., 2023), and Video Chat(Li et al., 2023b) for video understanding. Specifically, for interactive cross-modal video retrieval, future interactive video retrieval systems should function as "search assistants," engaging in genuine and coherent multi-round dialogues with users.

3 Interactive Video Corpus Retrieval Dataset

3.1 Dataset Collection and Annotation

To implement an interactive video retrieval system, we constructed a multi-turn, conversational dataset comprising 193,434 interactions sourced from 5 video repositories. This dataset encompasses functionalities such as video retrieval, video moment retrieval, and natural dialogue.

Illustrated in Figure 3, we devised a comprehensive collection pipeline:

- 1) Video source curation: Initially, we selected video datasets spanning diverse domains such as daily activities, movies, and kitchens, including selections like TVQA (Lei et al., 2018), LSMDC (Rohrbach et al., 2017),ActivityNet (Krishna et al., 2017), DiDeMo (Anne Hendricks et al., 2017), MSR-VTT (Xu et al., 2016), to ensure video source diversity. Subsequently, we filtered out select videos from these 5 original datasets. Videos featuring isolated actions or events, severe occlusion,

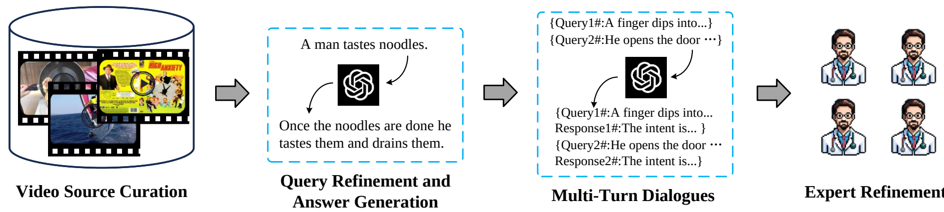


Figure 3: The pipeline of our dataset collection.

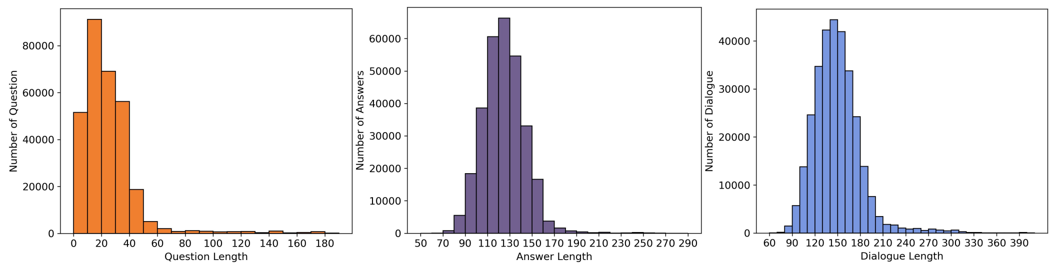


Figure 4: Distribution of question lengths, answer lengths, and dialogue lengths.

or excessively accelerated playback were excluded. Ultimately, 12,516 videos were chosen for inclusion.

- 2) Query refinement: Despite the presence of captions or descriptions with the filtered source videos, they often inadequately align with user queries in real-world scenarios. Hence, we employed GPT-4 for query refinement on captions. Specifically, we first combine the captions and user queries as input, and then use GPT-4 to generate user queries that more accurately and closely reflect the substantive content of the video.
- 3) Multi-turn dialogues: We established various dialogue dynamics, encompassing Long2Short, Short2Long, Long2Long, Short2Short, and Natural Dialogue scenarios. “Long2Short” denotes a user’s inclination to pinpoint video clips further in the current round, while “Natural Dialogue” reflects users perceiving our system as a standard chat robot. Notably, while most dialogues consist of concatenated single-round exchanges, we also gathered a limited number of multi-turn dialogues from actual users.
- 4) Interpretability: To bolster the interpretability of interactive retrieval systems, we utilized GPT-4 to craft responses, encompassing intent understanding, retrieval or localization results, and reasons.
- 5) Bilingual capability: To broaden the reach of this dataset, we employed a translation model to render the dataset into Chinese.

Notably, every output produced by GPT-4 will undergo meticulous scrutiny and refinement by human experts to guarantee the precision of knowledge. Additionally, we implemented a validation process conducted by a review team, focusing on the quality and consistency of annotations provided by different annotators. After all annotations (193,434 sentence-level queries) were completed, the reviewers further examined the annotated data. Ultimately, we acquired a multi-turn, conversational dataset comprising 200K volumes, named IVCR-200K. The entire annotation and review process took approximately five months. More details on annotation procedure is provided in the supplementary material C.

3.2 Dataset Analysis.

Property Quality. The statistical analysis of the property quality for video and textual query in the IVCR-200K dataset is shown in Figure 4 and Figure 5. In Figure 4, we present the length distribution of questions, answers, and dialogues within IVCR-200K. The average length of questions and answers in IVCR-200K is 24.5 words and 124.2 words, respectively. In contrast, the average length of questions in AVSD(Alamri et al., 2019) is 7.9 words, and the average answer length is 9.4 words. This indicates that the dialogues in our dataset are more verbose and conversational.

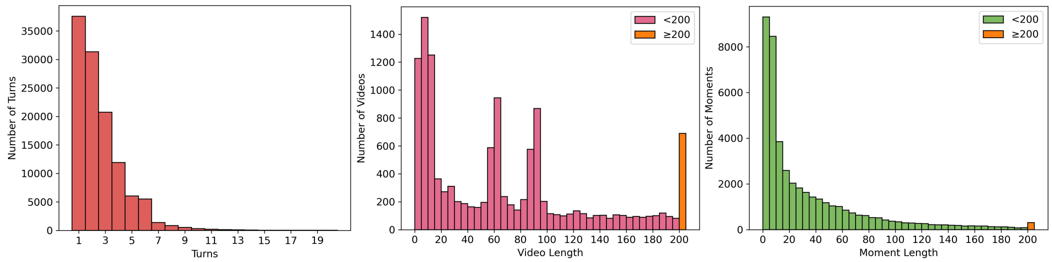


Figure 5: Distribution of turn lengths, video lengths, and moment lengths.

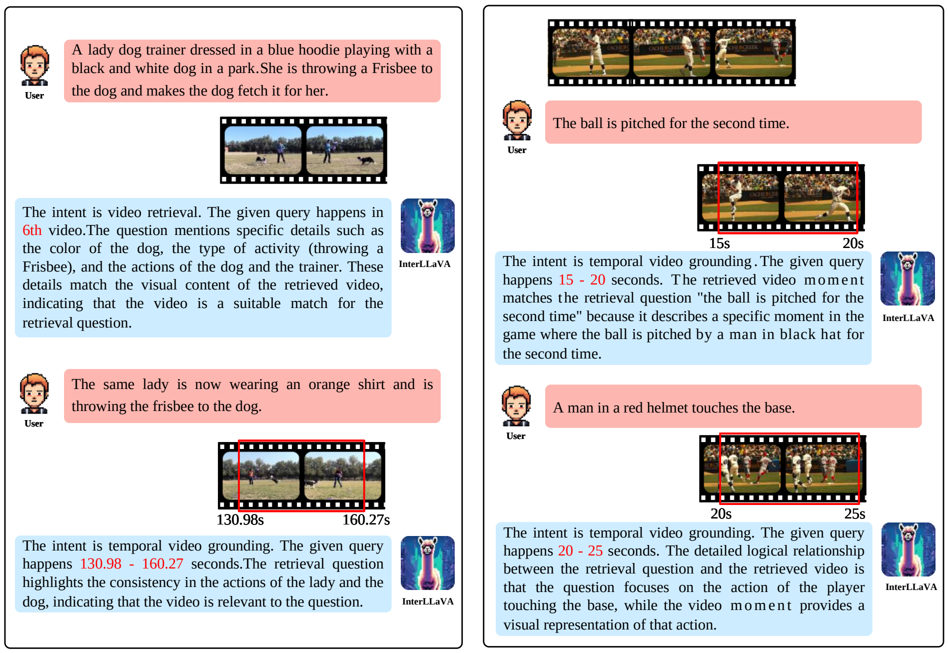


Figure 6: Examples from the IVCR-200K dataset.

Additionally, Figure 5 shows the distribution of the number of turns in multi-turn dialogues. The total number of dialogue turns is 302,074, with an average of approximately 2.6 turns, which aligns with typical user retrieval behavior. Figure 5 also presents the length distribution of videos and video moment. The average length of videos is 67.26 seconds, and the average length of video moments is 34.81 seconds, with most video moments being under 60 seconds.

Diversity Quality. We conducted an analysis of our video sources, the different types of videos, and performed a frequency analysis of annotated sentences, as detailed in supplementary material C.

Visualization Quality. We also check some cases as shown in the Figure 6. More examples are available in the supplementary material D.

4 Interactive Video Corpus Retrieval Framework

4.1 Task Definition.

Let $u_{(\cdot)}$ denotes a user whose historical interactive sequence is $Q = \{q_1, q_2, q_3, q_4, ...\}$, where $q_{(\cdot)}$ represents different textual queries. Formally, the goal of this interactive video corpus retrieval task is to retrieve semantically matched videos or moments in each round i, based on historical interactive sequence $Q_{<i}$. Among them, video moment retrieval requires not only the prediction of the most suitable video v_j, but also the prediction of the optimal moment within v_j, which includes the start s

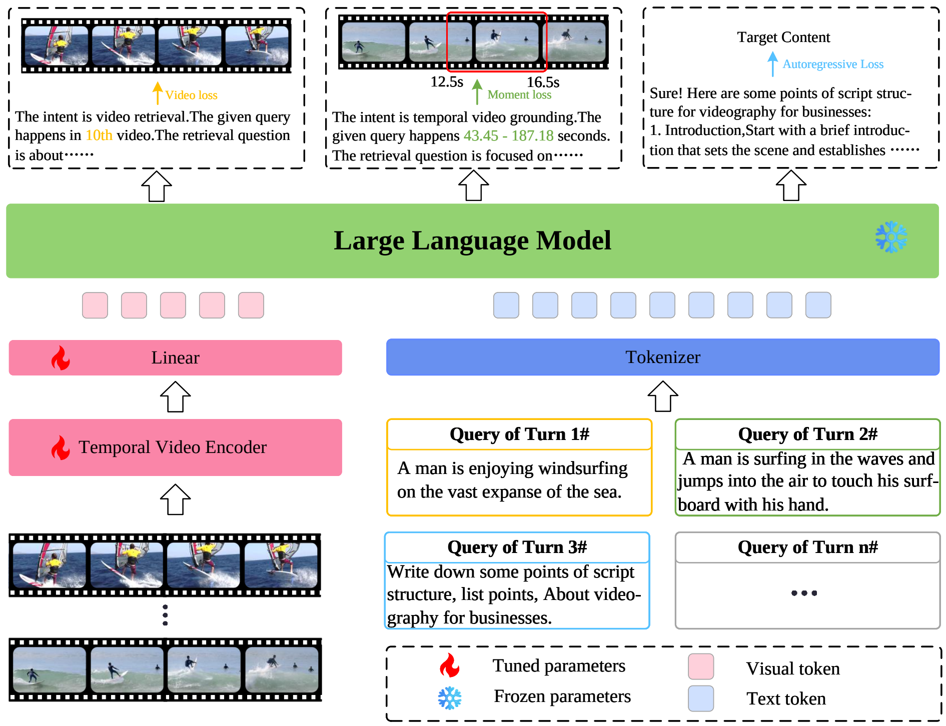


Figure 7: An overview of our framework for interactive retrieval.

and end e timestamps. In addition, the interactive video corpus retrieval task is not limited to video retrieval, but also specifically considers the identification and processing of natural dialogue intent.

4.2 Task Processing.

As illustrated in Figure 7, our InterLLaVA adapts the pretrained multi-modal large language model LLaMA-2 (7B)(Touvron et al., 2023) to tackle video retrieval, video moment retrieval, and natural dialogue in a multi-turn setting. It takes video and text query as inputs and outputs video, video moment, and natural dialogue related to textual query intent, while providing interpretable feedback. Specifically, we fine-tuned Inter-LLaVA using instruction-tuning data, which generally consists of video-instruction pairs. Here is an illustrative example, with the underlined part serving a pseudo-instruction:

> Video Retrieval:
> Question: ### Human: [User Query] <VID> <Video Start> [Video Tokens] <Video End> [Instruction]
> Answer: ### Assistant: The intent is video retrieval. The given query happens in <VID> video. [Explainable Feedback]
>
> Video Moment Retrieval:
> Question: ### Human: <Video Start> [Video Tokens] <Video End> [Timestamps] [User Query]
> Answer: ### Assistant: The intent is temporal video grounding. The given query happens in [Start Time] - [End Time] seconds. [Explainable Feedback]

During the instruction fine-tuning of InterLLaVA, text query is first performed using a pre-trained multi-modal large language model (LLaMA-2 (7B)), which is then concatenated with video and answer prompts to serve as the input for InterLLaVA. The answer prompts include retrieval intent, video/moment cues, and interpretable feedback. Later, the answer prompts are utilized as the "ground truth" of InterLLaVA's generation. In the following, we elaborate the implementations of the three tasks.

Video Retrieval. For this task, we propose combining a fast two-tower video model with a multi-modal large language model through a re-ranking mechanism. Specifically, in the first phase, the video retrieval model predicts the top-10 video sequence V_j based on videos and text queries. In the second phase, these top-10 video sequences and the text queries are input into a multi-modal large language model for re-ranking, outputting the most relevant video v_j. This approach retrieves the most relevant videos efficiently, reduces the memory and computational burden on the language model, and excludes irrelevant content. Notice that the first phase adopts offline video sequence extraction, while the second phase is trained end-to-end with the other tasks.

Video Moment Retrieval. For this task, we employ a traditional two-stage retrieval method, utilizing a fast two-tower model for video retrieval and a multi-modal large language model for precise moment localization. Specifically, we implement a two-phase approach. In the first phase, the video retrieval model directly output the top-1 video v_j. In the second phase, the textual query and the top-1 video are input into a multi-modal large language model to generate reasonable and coherent response and video moment. To enhance the feature fusion in the time dimension, we adopt a sliding video Q-Former and initialize it from the Video-LLaMA(Zhang et al., 2023) checkpoint. Moreover, we perform instruction tuning on our IVCR-200K dataset, which contains timestamp-related and natural dialogue data, to further strengthen InterLLaVA's timestamp localization and natural dialogue capabilities.

Training and Inference. In training, we implement a two-phase approach. In the first phase, we train a video retrieval model based on the video and text features encoded by BLIP-2(Li et al., 2023a), utilizing X-Pool(Gorti et al., 2022) as the base model. The video retrieval model acts as a plug-in for the multi-modal large language model, retrieving the top-10 video sequences or the top-1 video. In the second phase, we fine-tunes the InterLLaVA with instruction data to achieve instruction following. To better tailor LLaMA for video tasks, we leverage the LoRA(Hu et al., 2021) technique for efficient parameter fine-tuning. To adapt to our IVCR task, we designed a new loss function for training InterLLaVA. For training the large model, we employ a language model loss to generate the target answer R_a with a length of L_t. This loss is based on the probability of predicting each word in the answer sequence given the context, such as video tokens F_v and the query tokens F_q. It is formulated as

$$\begin{aligned} \mathcal{L}_M &= -\log P_\Theta(R_a|F_v, F_q) \\ &= -\sum_{i=1}^{L_t} \log P_\Theta(r_i|R_{a,<i}, F_v, F_q), \end{aligned} \tag{1}$$

where Θ represents the trainable parameters, and $R_{a,<i}$ refers to the answer tokens preceding the current prediction token r_i.

Since our goal is to enhance the large language model's ability for video re-ranking, a direct idea is to directly optimize the predicted video index with the ground truth video index. Let v_p be the predicted video index, and v_g denotes the ground truth video index. The cross-entropy loss function is computed as

$$\mathcal{L}_V = -\sum_{i=1}^{N} v_{g,i} \log(v_{p,i}), \tag{2}$$

where N is the total number of video indices, $v_{g,i}$ is the ground truth probability distribution(with 1 for the correct index and 0 for others), and $v_{p,i}$ is the predicted probability for the i-th video index.

Similarly, let c_p be the predicted video moment interval, and c_g denotes the ground truth video moment interval. we force the model to align each predicted moment candidate with the ground truth moment. Our model is trained by a Intersection over Union (IoU) loss(Yu et al., 2016) as

$$\mathcal{L}_C = 1 - \mathrm{IoU}(c_p, c_g). \tag{3}$$

The overall loss function for training the InterLLaVA is the sum of these three losses, formulated by

$$\mathcal{L} = \mathcal{L}_M + \alpha \cdot \mathcal{L}_V + \beta \cdot \mathcal{L}_C, \tag{4}$$

where $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ are trade-off parameters that balance the three loss terms.

In inference, we input the textual query into InterLLaVA. Subsequently, InterLLaVA then outputs intent analysis, video prediction or video moment prediction, as well as explainability feedback.

Table 2: Overall performance comparison of baselines. The "–" indicates not applicable, while bold represents optimal performance.

Types	Methods	R@1 ↑	R@10 ↑	R@1 IoU=0.5 ↑	R@1 IoU=0.7 ↑	BLEU-4 ↑	GPT-4 Score ↑
Video Retrieval	CLIP4Clip(Luo et al., 2022)	25.9	59.9	–	–	–	–
	X-Pool (Gorti et al., 2022)	25.3	61.1	–	–	–	–
	TS2-Net (Liu et al., 2022b)	49.1	80.1	–	–	–	–
	T-MASS (Wang et al., 2024)	30.2	74.5	–	–	–	–
	BLIP-2 (Li et al., 2023a)	53.5	88.6	–	–	–	–
Moment Retrieval	2D-TANZhang et al. (2020)	–	–	49.87	35.21	–	–
	UMT (Liu et al., 2022a)	–	–	13.45	7.31	–	–
	MMN (Wang et al., 2022b)	–	–	43.23	32.36	–	–
	MomentDiff (Li et al., 2024a)	–	–	11.59	3.4	–	–
	CG-DETR (Moon et al., 2023)	–	–	48.3	28.77	–	–
	GroundingGPT (Li et al., 2024b)	–	–	12.82	4.65	0.0018	0.68
	VTimeLLM (Huang et al., 2024)	–	–	17.95	7.76	0.0035	0.74
	TimeChat (Ren et al., 2023)	–	–	21.24	9.80	0.0	0.64
Interactive Video Retrieval	**InterLLaVA (Ours)**	**58.61**	–	**12.83**	**7.54**	**0.42**	**0.76**

5 Experiments

5.1 Experimental Settings.

Datasets Splits. Our datasets are split into 3 non-overlapping subsets, where 0.8, 0.1 and 0.1 are used for training, validation and testing. Specifically, our training set consists of 11,618 videos and 91,809 textual queries, while the test set includes 449 videos and 2,589 textual queries. The validation set also contains 449 videos and 2,608 textual queries.

Evaluation Metrics. We employ two types of metrics to assess our framework. For single-turn evaluation, we utilize R@1 and R@10 to gauge video retrieval proficiency, where 1/10 denotes the top-ranked videos. R@1 IoU={0.5, 0.7} is employed to assess video moment retrieval capability, with IoU=0.5 indicating that the IoU socre between the localized moment and the ground truth must exceed 0.5. Metrics such as BLEU-4 and GPT-4 Score are deployed to evaluate text generation. We classify GPT-4 scores into four categories: highly relevant (1), moderately relevant (0.6), somewhat relevant (0.4), and irrelevant (0). Moreover, we conduct multi-turn performance based on the aforementioned metrics, and any error between between rounds will affect subsequent scores.

Baselines. We selected the following five state-of-the-art models as benchmarks for video retrieval, all based on the prevailing pre-trained model CLIP(Radford et al., 2021). Additionally, to comprehensively evaluate the performance of video moment retrieval, we selected five methods as benchmarks. Furthermore, we chose three models based on multi-modal large language models as additional benchmarks for comparison. Please refer to the supplementary materials to obtain the detailed introduction of our baseline.

Implementation Details. We employ a pre-trained ViT-G/14 from EVA-CLIP(Sun et al., 2023) and the sliding video Q-Former(Ren et al., 2023) as the image encoder, with LLaMA-2 (7B)(Touvron et al., 2023) as the language model backbone. We train our InterLLaVA using the AdamW optimizer with an initial learning rate of 3e-5 and weight decay of 1e-6 in training phases 1 and 2. Fine-tuning was performed on IVCR-200K for 5 epochs with a batch size of 32. As depicted in Figure 7, the parameters of ViT and LLM remained frozen, while those of the image Q-Former, video Q-Former, and linear layer were tuned. For video retrieval, 12 frames are used, while for moment retrieval, 96 frames are used. All experiments were conducted on 4 Nvidia 4090 GPUs. In addition, the trade-off parameter α an β in Eq. (4) are set to 0.01.

5.2 Overall Performance Comparison

To evaluate the challenges presented by the IVCR-200K dataset, we conducted a comprehensive study on models for different tasks and our benchmark model. In Table 2, we compared our InterLLaVA with other state-of-the-art methods in video retrieval and video moment methods. Please refer to the supplementary materials to obtain the detailed introduction of our baseline. The detailed introductions to our baselines are provided in supplementary material E.

Table 3: The performance of different pre-retrieval modules.

Models	R@1 ↑	R@1 IoU=0.5 ↑	R@1 IoU=0.7 ↑
CLIP4Clip	58.84	10.84	6.59
X-Pool	58.61	11.18	6.15
T-MASS	57.59	11.88	6.33
BLIP-2	57.91	12.83	7.54

Table 4: Multi-Turn analysis of our framework.

	R@1 ↑	R@1 IoU=0.5 ↑	R@1 IoU=0.7 ↑
Turn 1#	41.58	6.56	5.01
Turn 2#	15.54	9.30	5.34
Turn 3#	10.60	9.30	5.48
Turn 4#	6.25	12.41	8.62

Overall Observations. 1) Notice that the IVCR task presents significant challenges in the field of video retrieval. While existing traditional models have achieved notable success in single tasks such as video retrieval and video moment retrieval, they fall short compared to our InterLLaVA in terms of considering the importance of flexibly adjusting retrieval strategies based on retrieval intent. This limitation restricts the flexibility and adaptability of video retrieval to some extent. 2) For video moment retrieval, compared to multimodal large language-based methods (e.g., TimeChat(Ren et al., 2023)), traditional methods (e.g., 2D-TAN(Zhang et al., 2020)) achieve superior performance in moment localization. Their advantage lies in the ability to perceive richer contextual information. 3) Moreover, the CLIP-based and BLIP-2-based models, TS2-Net(Liu et al., 2022b) and BLIP-2(Li et al., 2023a), have demonstrated excellent performance on video retrieval task. This proves their ability to more effectively align key textual and video information.

5.3 Robustness Analysis

In this section, we will delve into our framework from two perspectives: retrieval module, and multi-turn analysis. We will examine the retrieval module's functionality within the framework, and evaluate the performance of multi-turn dialogue.

Retrieval Module. We validate the effectiveness of interactive retrieval modeling by substituting different video retrieval models in Table 3. Our observations are as follows: 1) Upon comparing Tables 2 and 3, it becomes apparent that, for the video retrieval task, CLIP-based models (e.g., X-Pool) demonstrate significantly greater performance improvements (25.3 ⇒ 59.85) compared to the BLIP-2(Li et al., 2023a) model. 2) In contrast, for the video moment retrieval task, CLIP-based models exhibit slightly diminished performance, suggesting that InterLLaVA's video localization capabilities are influenced by the underlying video retrieval model. Overall, these observations empirically validate the effectiveness of video retrieval models and large language models in modeling interactive retrieval.

Multi-Turn Analysis. To evaluate the effectiveness of the model, we compared its performance across different turns of dialogue. As shown in Table 4, as the number of retrieval turns increases, the performance of video retrieval slightly decreases, whereas the performance of video moment retrieval improves.This finding highlights the significant role of context learning in enhancing video localization ability during multi-turn retrieval. It also suggests that video retrieval itself is relatively less influenced by multi-turn context understanding.

6 Conclusions

In this paper, we propose a more realistic task to establish an "interaction" between the retrieval system and the user, which involves multi-turn, conversational, and realistic interactions. To facilitate research on this challenging task, we introduce a dataset and framework designed to serve this novel purpose. Notably, our IVCR-200K dataset is a high-quality, bilingual, multi-turn, conversational, and abstract semantic dataset that supports both video and moment retrieval. Our framework is based on MLLMs, which provide more explainable solutions to support users' interaction modes. Our extensive experiments demonstrate the effectiveness of our dataset and framework.

Moving forward, we plan to expand the scope of this research by increasing the size of the dataset and model parameters. Additionally, we will endeavor to develop more sophisticated model architectures to enhance the model's capabilities, considering the challenges posed by interactive retrieval.

REFERENCES

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*, pp. 7558–7567, 2019.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5803–5812, 2017.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding. *arXiv:2102.05095*, 2(3):4, 2021.

Da Cao, Yawen Zeng, Meng Liu, Xiangnan He, Meng Wang, and Zheng Qin. Strong: Spatio-temporal reinforcement learning for cross-modal video moment localization. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 4162–4170, 2020.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

David 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: Human Language Technologies*, pp. 190–200, 2011.

Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10638–10647, 2020.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018.

Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9346–9355, 2019.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv:2106.11097*, 2021.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference Computer Vision*, pp. 214–229, 2020.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5267–5275, 2017.

Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16167–16176, 2022.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5006–5015, 2022.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

Dongliang He, Xiang Zhao, Jizhou Huang, Fu Li, Xiao Liu, and Shilei Wen. Read, watch, and move: Reinforcement learning for temporally grounding natural language descriptions in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8393–8400, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv:2106.09685*, 2021.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *Empirical Methods in Natural Language Processing*, pp. 1369–1379, 2018.

Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4953–4963, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742, 2023a.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv:2305.06355*, 2023b.

Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. Momentdiff: Generative video moment retrieval from random to real. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*, volume 36, 2024a.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. Groundinggpt: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 6657–6678, 2024b.

Kaiqu Liang and Samuel Albanie. Simple baselines for interactive video retrieval with questions and answers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11091–11101, 2023.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv:2311.10122*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *British Machine Vision Conference*, pp. 279, 2019.

Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042–3051, 2022a.

Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European Conference on Computer Vision*, pp. 319–335, 2022b.

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.

Zhixin Ma and Chong Wah Ngo. Interactive video corpus moment retrieval using reinforcement learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 296–306, 2022.

Avinash Madasu, Junier Oliva, and Gedas Bertasius. Learning to retrieve videos by asking questions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 356–365, 2022.

Sho Maeoki, Kohei Uehara, and Tatsuya Harada. Interactive video retrieval with dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 952–953, 2020.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2630–2640, 2019.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on international conference on multimedia retrieval*, pp. 19–27, 2018.

WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided query-dependency calibration in video representation learning for temporal grounding. *arXiv:2311.08835*, 2023.

Keyu Pan and Yawen Zeng. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv:2307.16180*, 2023.

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*, pp. 8748–8763, 2021.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. *arXiv:2312.02051*, 2023.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123:94–120, 2017.

Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference Computer Vision*, pp. 510–526, 2016.

Alan F Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and trecvid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pp. 321–330, 2006.

Cees GM Snoek, Marcel Worring, Ork de Rooij, Koen EA van de Sande, Rong Yan, and Alexander G Hauptmann. Videolympics: real-time evaluation of multimedia retrieval systems. *IEEE MultiMedia*, 15(1):86–91, 2008.

Xue Song, Jingjing Chen, Zuxuan Wu, and Yu-Gang Jiang. Spatial-temporal graphs for cross-modal text2video retrieval. *IEEE Transactions on Multimedia*, 2021.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv:2303.15389*, 2023.

Bart Thomee and Michael S Lew. Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval*, 1:71–86, 2012.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–10, 2024.

Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3313–3322, 2022a.

Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5079–5088, 2021.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4581–4591, 2019.

Zhenzhi Wang, Limin Wang, Tao Wu, Tianhao Li, and Gangshan Wu. Negative sample matters: A renaissance of metric learning for temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2613–2623, 2022b.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5288–5296, 2016.

Rui Yan, Mike Zheng Shou, Yixiao Ge, Jinpeng Wang, Xudong Lin, Guanyu Cai, and Jinhui Tang. Video-text pre-training with learned regions for retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 3100–3108, 2023.

Jiahui Yu, Yuning Jiang, Zhangyang Wang, Zhimin Cao, and Thomas Huang. Unitbox: An advanced object detection network. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 516–520, 2016.

Yawen Zeng. Point prompt tuning for temporally language grounding. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2003–2007, 2022.

Yawen Zeng, Da Cao, Xiaochi Wei, Meng Liu, Zhou Zhao, and Zheng Qin. Multi-modal relational graph for cross-modal video moment retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2215–2224, 2021.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv:2306.02858*, 2023.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 12870–12877, 2020.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.