Learning to Retrieve Videos by Asking Questions

Avinash Madasu
Department of Computer Science
UNC Chapel Hill
USA
avinashm@cs.unc.edu

Junier Oliva
Department of Computer Science
UNC Chapel Hill
USA
joliva@cs.unc.edu

Gedas Bertasius
Department of Computer Science
UNC Chapel Hill
USA
gedas@cs.unc.edu

ABSTRACT

The majority of traditional text-to-video retrieval systems operate in static environments, i.e., there is no interaction between the user and the agent beyond the initial textual query provided by the user. This can be sub-optimal if the initial query has ambiguities, which would lead to many falsely retrieved videos. To overcome this limitation, we propose a novel framework for Video Retrieval using Dialog (ViReD), which enables the user to interact with an AI agent via multiple rounds of dialog, where the user refines retrieved results by answering questions generated by an AI agent. Our novel multimodal question generator learns to ask questions that maximize the subsequent video retrieval performance using (i) the video candidates retrieved during the last round of interaction with the user and (ii) the text-based dialog history documenting all previous interactions, to generate questions that incorporate both visual and linguistic cues relevant to video retrieval. Furthermore, to generate maximally informative questions, we propose an Information-Guided Supervision (IGS), which guides the question generator to ask questions that would boost subsequent video retrieval accuracy. We validate the effectiveness of our interactive ViReD framework on the AVSD dataset, showing that our interactive method performs significantly better than traditional non-interactive video retrieval systems. We also demonstrate that our proposed approach generalizes to the real-world settings that involve interactions with real humans, thus, demonstrating the robustness and generality of our framework.

CCS CONCEPTS

• **Computing methodologies** → **Visual content-based indexing and retrieval**.

KEYWORDS

interactive video retrieval, dialog generation, multi-modal learning

ACM Reference Format:
Avinash Madasu, Junier Oliva, and Gedas Bertasius. 2022. Learning to Retrieve Videos by Asking Questions. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548361

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548361

1 INTRODUCTION

The typical (static) video retrieval framework fetches a limited list of candidate videos from a large collection of videos according to a user query (e.g. 'cooking videos'). However, the specificity of this query will likely be limited, and the uncertainty among candidate videos based on the user query is typically opaque (i.e. the user might not know what additional information will yield better results).

For example, consider the scenario where you are deciding what dish to make for dinner on a Friday night. Now also suppose that you have access to an interactive AI agent who can help you with this task by retrieving the videos of relevant dishes and detailed instructions on how to make those dishes. In this particular scenario, you might start an interaction with the agent by asking it to "show some cooking videos" (Figure 1 left-most query). Any traditional video retrieval model will look for the matching cooking videos from the database and display them to the user. However, what happens if there are too many matching videos, most of which don't satisfy the user's *internal* criteria? A user-friendly video retrieval framework will not display all such videos to the user and expect them to sift through hundreds of videos to find the videos that are most relevant to them. Not only this would be overly time consuming, but it would also hurt the user experience. Instead, one way to address the uncertainty would be to ask another follow-up question of the same user: "Which cuisine do you prefer?" This would then allow the user to provide additional information clarifying some of his/her preferences (e.g., plant or meat diet, etc.) so that an AI agent can narrow down its search.

The rise of conversational AI systems, such as chat-bots and voice assistants, have made the user interaction with a digital agent relatively smooth. Inspired by this emerging technology, and a huge availability of video data, we propose ViReD , a framework for **Vi**deo **Re**trieval using **D**ialog. Injecting dialog into standard text-to-video retrieval systems has two key advantages: (i) it reduces the uncertainty associated with the initial user text queries, (ii) it enables the agent to infer user's internal preferences, thus, making the AI model more personalized to the user.

Several works [14, 19, 20, 45, 54] explored the idea of interactive mechanisms in the context of image retrieval. Prior methods in this area used relevance scores [45, 54] and attribute comparisons [19, 20] to get user feedback for retrieval. Additionally, the recent work of Cai *et al.* [7] proposed Ask-and-Confirm, a framework that allows the user to confirm if the proposed object is present or absent in the image. One downside of these prior approaches is that they typically require many interaction rounds (e.g., > 5), which increases user effort, and degrades user experience. Furthermore, these approaches significantly limit the form of the user-agent interaction, i.e., the users can only verify the presence or absence of a particular object/attribute in an image but nothing more. In

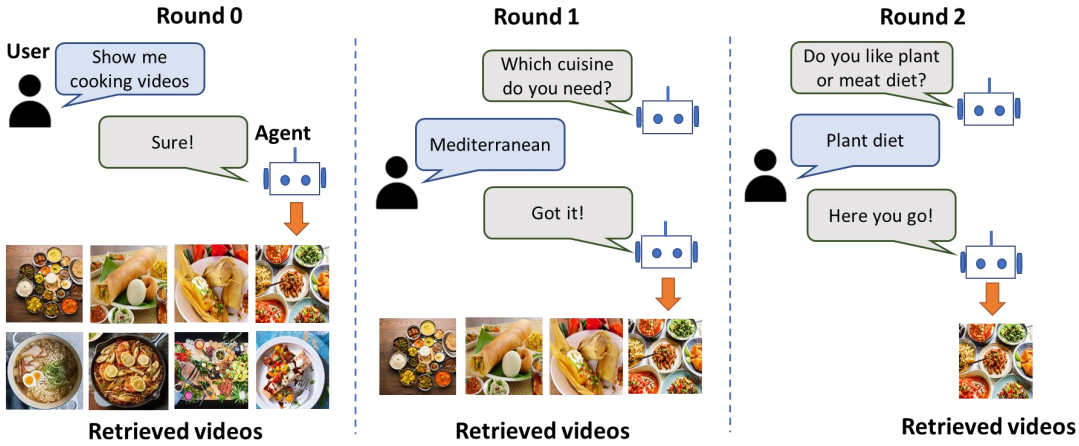

Figure 1: An illustration of our interactive dialog-based video retrieval framework. The order of conversation between the user and the agent is illustrated from left to right. The agent has access to a large video database, which is used for retrieving user-specified videos. For example, in this case, the user starts an interaction with the agent by asking it to "Show some cooking videos." The agent then searches for relevant videos in the database and returns eight candidate videos. Due to high uncertainty in the initial query, the agent then asks another follow-up question "Which cuisine do you prefer?" for which the user responds: "Mediterranean." As the number of retrieved video candidates is reduced to four, the agent asks one final question: "Do you like plant or meat diet?" The user's response (i.e., "plant diet") then helps the agent to reduce the search space to the final candidate video, which is then displayed to the user.

contrast, our ViReD framework enables the user to interact with an agent using *free-form* questions, which is a natural form of interaction for most humans. We also note that our interactive framework achieves excellent video retrieval results with a few (e.g., $2-3$) interaction rounds.

Our key technical contribution is a multimodal question generator optimized with a novel Information-Guided Supervision (IGS). Unlike text-only question generators, our question generator operates on (i) the entire textual dialog history (if any) and (ii) previously retrieved top video candidates, which allows it to incorporate relevant visual and linguistic cues into the question generation process. Furthermore, our proposed IGS training objective enables our model to generate maximally informative questions, thus, leading to higher text-to-video retrieval accuracy.

We validate our entire interactive framework ViReD on the Audio-Visual Scene Aware Dialog dataset (AVSD) [3] demonstrating that it outperforms all non-interactive methods by a substantial margin. Furthermore, compared to other strong dialog-based baselines, our approach requires fewer dialog rounds to achieve similar or even better results. We also demonstrate that our approach generalizes to the real-world scenarios involving interactions with real humans, thus, indicating its effectiveness and generality. Lastly, we thoroughly ablate different design choices of our interactive video retrieval framework.

2 RELATED WORK

2.1 Multimodal Conversational Agents

There has been a significant progress in designing multimodal conversational agents especially in the context of image-based visual dialog [4, 10, 11, 40, 41]. Das *et al.*[10] proposed the task of visual dialog, in which an agent interacts with a user to answer questions about the visual video content. There also exists prior work in co-operative image guessing between a pair of AI agents [11]. Furthermore, the recent work of Niu *et al.* [40] proposes a recursive visual attention scheme for visual dialog generation. We note that most of these prior approaches operate in closed-set environments, i.e., selecting questions/answers from a small set of candidates. In contrast, our model leverages visual and linguistic cues to generate open-ended questions optimized for video retrieval.

2.2 Video Question Answering

Following standard visual question answering (VQA) methods in images [1, 2, 36, 52], video based question answering (video QA) aims to answer questions about videos [23, 24, 55, 57]. Compared to visual question answering in images, video question answering is more challenging because it requires complex temporal reasoning. Le *et al.* [21] introduced a multi-modal transformer model for video QA to incorporate representations from different modalities. Additionally, Le *et al.* [22] proposed a bi-directional spatial temporal reasoning model to capture inter dependencies along spatial and temporal dimensions of videos. Recently, Lin *et al.* [29] introduced Vx2Text, a multi-modal transformer-based generative network for video QA. Compared to these prior methods, we aim to develop a framework for interactive dialog-based video retrieval setting.

2.3 Multimodal Video Retrieval

Most of the recent multimodal video retrieval systems are based on deep neural networks [5, 8, 9, 12, 13, 15, 38]. With the advent of transformer-based language models [25, 43, 44], several methods proposed transformer architectures for video retrieval [5, 9, 13]. However, these methods focus on static query-based video retrieval and perform poorly when the textual user queries are ambiguous. In contrast, our work proposes to use dialog as means to gain additional information for improving video retrieval results.

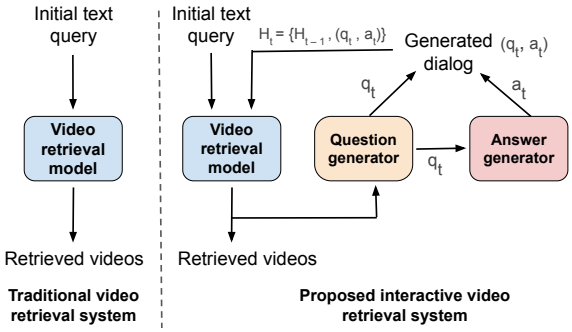

Figure 2: Comparison between the traditional (i.e., static) and our interactive dialog-based video retrieval frameworks. In the traditional set-up, the user interacts with the agent once by providing a single textual query to retrieve the desired video. In comparison, our proposed framework leverages multiple rounds of dialog with the user to improve video retrieval performance. Specifically, after the initial user query, the first round of retrieved videos are used to generate a question q_t, which the user then answers with an answer a_t. The generated dialog is added to the dialog history $H_t = \{H_{t-1}, (q_t, a_t)\}$, which is then used as additional input in the subsequent rounds of interaction.

2.4 Interactive Modeling Techniques

Ishan *et al.* [37] proposed an interactive learning framework for visual question answering. In this framework, the agent actively interacts with the oracle to get the information needed to answer visual questions. Several approaches utilized interactive mechanisms to perform image retrieval [7, 20, 35, 39, 54]. Additionally, multiple interactive video retrieval methods were proposed for the video browser showdown (VBS) benchmark [48]. These prior methods used interactive interfaces to obtain additional information related to the original search query from the user. This includes information such as attribute-like sketches [16], temporal context [31, 34], color [53], spatial position [17], and cues from the Kinect sensors [30]. In comparison, we note that our proposed dialogue-based video retrieval framework is complementary to these prior approaches. In particular, we believe that combining dialog with the above listed cues (e.g., sketches, position, color, etc.) should boost the subsequent video retrieval accuracy.

3 VIDEO RETRIEVAL USING DIALOG

In this section, we introduce VIRED , our proposed video retrieval framework using dialog. Formally, given an initial text query T specified by the user, and the previously generated dialog history H_{t-1}, our goal is to retrieve k most relevant videos $V_1, V_2, ..., V_k$.

Our high-level framework, which is illustrated in Figure 2, consists of three main components: (i) a multimodal question generator trained with an information-guided supervision (IGS), (ii) an answer generation oracle, which can answer any questions about a given video, thus, simulating an interaction with a human, and (iii) a video retrieval module, which takes as inputs the initial textual query and any generated dialog history and retrieves relevant videos from a large video database. We now describe each of these components in more detail.

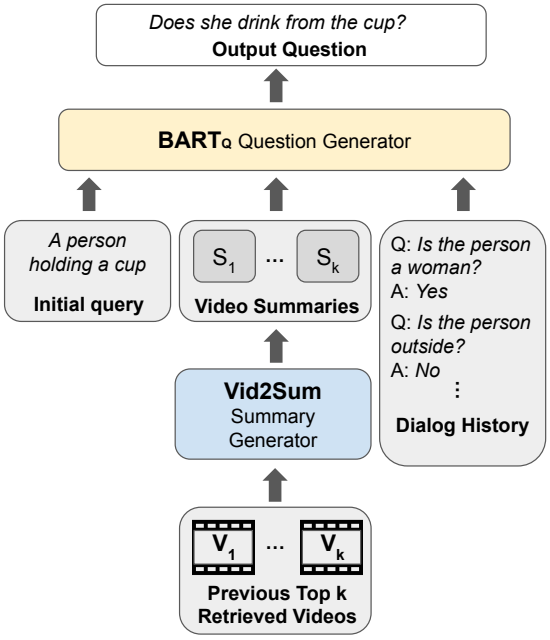

Figure 3: Illustration of the proposed question generator. It receives (i) an initial user-specified textual query, (ii) top-k retrieved candidate videos (from the previous interaction rounds), and (iii) the entire dialog history as its inputs. We then use a pretrained caption generator (Vid2Sum [51]) to map the videos into text. Afterward, all of the text-based inputs (including the predicted video captions) are fed into an autoregressive BART model for a new question generation.

3.1 Question Generator

As illustrated in Figure 3, at time t, our question generator takes as inputs (i) the initial text query T, (ii) top k retrieved videos at time $t-1$, and (iii) previously generated dialog history H_{t-1}. To eliminate the need for ad-hoc video-and-text fusion modules [26, 29], we use Vid2Sum video caption model [51] trained on the AVSD dataset to predict textual descriptions for each of the top-k previously retrieved videos. Specifically, given a video V_i, the Vid2Sum model provides a detailed textual summary of the video content, which we denote as S_i. Afterward, the predicted summaries for all k videos retrieved at timestep $t-1$, denoted as $S_1, S_2, ..., S_k$, are fed into the question generator along with the initial textual query T and previous dialog history H_{t-1}. More precisely, we concatenate the (i) initial text query, (ii) the predicted summaries and (iii) the previous generated dialog history and pass it to an autoregressive BART language model [25] for generating the next question.

$$X_q = \text{Concat}(T, S_1, S_2, .., S_k, H_{t-1}), \quad (1)$$

$$q_t = \text{BART}_q(X_q). \quad (2)$$

3.2 Answer Generation Oracle

The answer generator serves as an oracle that can answer any questions about a given video. We design our answer generator with a goal of simulating the presence of a human in an interactive dialog setting. Our goal is to use our answer generator to answer any

open-ended questions posed by our previously described question generator. This characteristic is highly appealing as it makes our framework flexible and applicable to many diverse dialog scenarios. In contrast, the majority of prior methods [7] are typically constrained to a small set of closed-set question/answer pairs, which makes it difficult to generalize them to diverse real-world dialog scenarios. In our experimental section 6.4, we also conduct a user-study evaluation demonstrating that our answer generation oracle effectively replaces a human answering the questions.

Our answer generator takes a video V_i and a question q_t as its inputs. Similar to before, we first use a pretrained video caption model Vid2Sum [51] to output a detailed textual summary S_i, which we then use as part of the inputs to the answer generator. Afterward, the generated summary S_i and the question q_t are concatenated and passed to a separate BART answer generation model to generate an answer a_t about the video V_i.

$$X_a = \text{Concat}(S_i, q_t), \tag{3}$$

$$a_t = \text{BART}_a(X_a). \tag{4}$$

Note that the BART models used for question and answer generation have the same architecture but that their weights are different (i.e., they are trained for two different tasks).

With these individual components in place, we can now generate t rounds of dialog using our previously defined question and answer generators. The whole dialogue history generated over t rounds can then be written as:

$$H_t = (\{q_1, a_1\}, \{q_2, a_2\}, ..., \{q_t, a_t\}). \tag{5}$$

The generated dialog history H_t is then used as an additional input to the video retrieval framework, which we describe below.

3.3 Text-to-Video Retrieval Model

Our video retrieval model (VRM) takes an initial textual query T and previous dialog history H_t and returns a probability distribution $p \in \mathbb{R}^N$ that encodes the (normalized) similarity between each video $V^{(i)}$ in the database of N videos and the concatenated text query $[T, H_t]$. Formally, we can write this operation as:

$$p = \text{VRM}(T, H_t), \tag{6}$$

where p_i encodes a probability that the i^{th} video $V^{(i)}$ is the correct video associated with the concatenated textual query $[T, H_t]$.

Our video retrieval model consists of two main components: (i) a visual encoder $F(V; \theta_v)$ with learnable parameters θ_v and (ii) a textual encoder $G(T, H_t; \theta_t)$ with learnable parameters θ_t. During training, we assume access to a manually labeled video retrieval dataset $\mathcal{X} = \{(V^{(1)}, T^{(1)}, H_t^{(1)}), \ldots, (V^{(N)}, T^{(N)}, H_t^{(N)})\}$, where $T^{(i)}$ and $H_t^{(i)}$ depict textual queries and dialog histories associated with a video $V^{(i)}$ respectively. As our visual encoder, we use a video transformer encoder [6] that computes a visual representation $f^{(i)} = F(V^{(i)}; \theta_v)$ where $f^{(i)} \in \mathbb{R}^d$. As our textual encoder, we use DistilBERT [47] to compute a textual representation $g^{(i)} = G(T^{(i)}, H_t^{(i)}; \theta_t)$ where $g^{(i)} \in \mathbb{R}^d$. We can jointly train the visual and textual encoders end-to-end by minimizing the sum of video-to-text and text-to-video matching losses as is done in [5]:

$$\mathcal{L}_{\text{v2t}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(f^{(i)} \cdot g^{(i)})}{\sum_{j=1}^{B} \exp(f^{(i)} \cdot g^{(j)})}, \tag{7}$$

$$\mathcal{L}_{\text{t2v}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(g^{(i)} \cdot f^{(i)})}{\sum_{j=1}^{B} \exp(g^{(i)} \cdot f^{(j)})}. \tag{8}$$

Here, B is the batch size, and $f^{(i)}, g^{(j)}$ are the embeddings of the i^{th} video and j^{th} text embeddings (corresponding to the j^{th} video) respectively. Note that we use the matching text-video pairs in a given batch as positive and all the other pairs as negative samples.

During inference, given an initial user query T and the previous dialog history H_t, we extract a textual embedding $g = G(T, H_t; \theta_t)$ using our trained textual encoder where $g \in \mathbb{R}^{1\times d}$. Additionally, we also extract visual embeddings $f^{(i)} = F(V^{(i)}; \theta_v)$ for every video $V^{(i)}$ where $i = 1 \ldots N$. We then stack the resulting visual embeddings $[f^{(1)}; \ldots; f^{(N)}]$ into a single feature matrix $Y \in \mathbb{R}^{N\times d}$. Afterward, the video retrieval probability distribution $p \in \mathbb{R}^{1\times N}$ is computed as a normalized dot product between a single textual embedding g and all the visual embeddings Y. This can be written as: $p = \text{Softmax}(gY^\top)$. For simplicity, throughout the remainder of the draft, we denote this whole operation as $p = \text{VRM}(T, H_t)$.

4 INFORMATION-GUIDED SUPERVISION FOR QUESTION GENERATION

Our goal in the above-described question generation step is to generate questions that will maximize the subsequent video retrieval performance. To do so, the generator must be able to comprehend: (i) the information it has already obtained, e.g. through the history of dialogue and initial query; (ii) its current belief and ambiguity over potential videos that should be retrieved to the user, e.g. through the current top candidate videos; and (iii) the potential information gain of posing new questions, e.g. the anticipated increase in performance that may be had by posing certain questions.

Although providing currently known information and belief over videos is straightforward via the dialogue history and top-k candidate videos, respectively, comprehending (and planning for) future informative questions is difficult. A major challenge stems from the free-form nature of questions that may be posed. There is a large multitude of valid next questions to pose. Thus, explicitly labeling the potential information gain of all valid next questions does not scale. One may define the task of posing informative queries as a Markov decision process (MDP), where the current state contains known information, actions include possible queries to make, and rewards are based on the number of queries that were made versus the accuracy of the resulting predictions [27, 50]. Previous interactive image retrieval [7, 39] approaches have used similar MDPs optimized through reinforcement learning to train policies that may select next questions from a limited finite list. However, these reinforcement learning (RL) approaches suffer when the action space is large (as is the case with open-ended question generation) and when rewards are sparse (as is the case with accuracy after final prediction) [27]. Thus, we propose an alternative approach, information-guided question generation supervision (IGS), that bypasses a difficult RL problem, by explicitly defining informative targets for the question generated based on a post-hoc search.

Suppose that for each video $V^{(i)}$, $i \in \{1, \ldots, N\}$, we also have m distinct human-generated questions/answers relevant to the video $D^{(i)} = \{D_1^{(i)}, \ldots, D_m^{(i)}\}$. Typically, such data is collected independently to any particular video retrieval system; e.g. in the AVSD [3] dataset, users ask (and answer) multiple questions about the content of a given video (without any particular goals in mind). However, these human-generated questions can serve as potential targets for our question generator. With IGS, we propose to filter through $D^{(i)}$ according to the retrospective performance as follows. During training, we collect targets for the question generator at each round of dialogue separately. Let $T^{(i)}$, be an initial textual query corresponding to ground truth video $V^{(i)}$. Then, also let $S_{t,1}^{(i)}, \ldots, S_{t,k}^{(i)}$ be our predicted text summaries of top-k retrieved candidate videos after the t^{th} rounds of dialogue, $H_t^{(i)}$ (note that $H_0^{(i)} = \emptyset$). We try appending question/answers (q, a) in $D^{(i)}$ not in $H_t^{(i)}$ to $H_t^{(i)}$ and see remaining question would most improve retrieval performance. That is, we collect

$$(q_{t+1}^{*(i)}, a_{t+1}^{*(i)}) = \underset{(q,a) \in (D^{(i)} \setminus H_t^{(i)})}{\text{argmax}} \left[\text{VRM}(T^{(i)}, H_t^{(i)} \cup \{(q, a)\}) \right]_i, \tag{9}$$

where VRM is our previously described video retrieval model (see Sec. 3.3). Note that here, $\left[\text{VRM}(T^{(i)}, H_t^{(i)} \cup \{(q, a)\})\right]_i = p_i$, which depicts our previously defined retrieval probability between the ground truth video $V^{(i)}$, and the concatenated text query $T^{(i)}, H_t^{(i)} \cup \{(q, a)\}$. Each of the retrospective best questions are then set up as a target for the question generator at the $t + 1^{\text{th}}$ round

$$\mathcal{D}_{t+1} = \left\{ \left(\left(T^{(i)}, S_{t,1}^{(i)}, \ldots, S_{t,k}^{(i)}, H_t^{(i)} \right), q_{t+1}^{*(i)} \right) \right\}_{i=1}^{N}, \tag{10}$$

where $\left(T^{(i)}, S_{t,1}^{(i)}, \ldots, S_{t,k}^{(i)}, H_t^{(i)} \right)$ are the respective initial query, our predicted text summaries of top-k previous retrievals, and dialogue history that are inputs to the question generator, BART_q. The target question/answers are appended to the histories $H_{t+1}^{(i)} = H_t^{(i)} \cup \{(q_{t+1}^{*(i)}, a_{t+1}^{*(i)})\}$, and the next round of target questions $\mathcal{D}_{t+2}$ is similarly collected. Please note that $\mathcal{D}_{t+1}$ depends on $\mathcal{D}_t$ since we consider appending questions to previous histories. That is, at each round we look for informative questions based on the histories seen at that round. Jointly, the dataset $\mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots \cup \mathcal{D}_M$ serve as a *supervised* dataset to directly train the question generator, BART_q, to generate informative questions.

5 EXPERIMENTS

5.1 Dataset

We test our model on the audio-visual scene aware dialog dataset (AVSD) [3], which contains ground truth dialog data for every video in the dataset. Specifically, each video in the AVSD dataset has 10 rounds of human-generated questions and answers describing various details related to the video content (e.g., objects, actions, scenes, people, etc.). Thus, we believe that this dataset is well suited to our setting. In total, the AVSD dataset consists of 7, 985 training, 863 validation, and 1, 000 testing videos [35]. Throughout our experiments, we use standard training, validation and test splits.

5.2 Implementation Details

5.2.1 Question Generator. We train our question generator using the BART large architecture. We set the maximum sentence length to 120. During generation, we use the beam search of size 10. The question generator is trained for 5 epochs with a batch size of 32.

5.2.2 Answer Generator. We also use the BART large architecture to train our answer generator. Note that the question and answer generators use the same architecture but are trained with two different objectives, thus, resulting in two distinct models. The maximum sentence length for answer generation is set to 135. During generation, we use the beam search of size 8. The model is trained with a batch size of 32 for 2 epochs.

5.2.3 Video Retrieval Model. We use Frozen-in-Time (FiT) [5] codebase to implement our video retrieval model. Specifically, we fine-tune the provided pretrained model on the AVSD dataset for 20 epochs with a batch size of 16. Early stopping is applied if the validation loss doesn't improve for 10 epochs. We use AdamW [32] optimizer with a learning rate of $3e^{-5}$.

5.2.4 Vid2Sum Captioning Model. We fine-tune the video captioning model on the training set of AVSD for 5 epochs. We use the same hyper parameters as in [51]. During inference, we use our trained Vid2Sum model to generate textual summaries for each input video. The generated summary has a maximum length of 25.

5.3 Evaluation Metrics

We measure the video retrieval performance using standard Recall@k ($k = 1, 5, 10$), and MedianR, MeanR evaluation metrics. Recall@k calculates the percentage of test data for which the ground-truth video is found in the retrieved k videos (the higher the better). Additionally, the MeanR and MedianR metrics depict the mean and the median rank of the retrieved ground truth videos respectively (the lower the better). Unless noted otherwise, all models are averaged over 3 runs.

5.4 Video Retrieval Baselines

LSTM [35]. Maeok *et al.* [35] proposed an LSTM-based model that processes human-generated ground truth dialog for video retrieval. Unlike this prior approach, our interactive ViReD approach does not require ground truth dialog data during inference. Instead, during each round of interaction, our method generates novel open-ended questions that maximize video retrieval accuracy.

Frozen-in-Time [5]. We fine-tune the Frozen-in-Time (FiT) model to retrieve the correct video using the initial textual query T as its input (without using dialog).

Frozen-in-Time w/ Ground Truth Human Dialog. We fine-tune the Frozen-in-Time model using the textual query and the full 10 rounds of human-generated ground truth dialog history. Unlike our ViReD approach, which uses our previously introduced question and answer generators to generate dialog, this Frozen-in-Time w/ Dialog baseline uses 10 rounds of manually annotated human dialog history during inference. In this setting, we concatenate 10 rounds of ground truth dialog with the initial text query, and use the concatenated text for video retrieval.

Table 1: Comparison to prior video retrieval models. In the "Pretrain Data" column, we list external datasets used for pretraining. The "Dialog" and "Dialog Rounds" columns depict whether the dialog is used as additional input, and if so how many rounds of it. Based on these results, we observe that our VIRED approach outperforms all baselines, including a strong Frozen-in-Time baseline augmented with 10 rounds of human-generated ground truth dialog.

Method	Pretrain Data	Dialog	Dialog Rounds (↓)	R@1 (↑)	R@5 (↑)	R@10 (↑)	MedianR (↓)	MeanR (↓)
LSTM [35]	ImageNet [46]	✓	10	4.2	13.5	22.1	–	119
FiT [5]	WebVid2M [5]	✗	–	5.6	18.4	27.5	25	95.4
FiT w/ Human dialog [35]	WebVid2M [5]	✓	10	10.8	28.9	40	18	58.7
VIRED	WebVid2M [5]	✓	**3**	12	30.5	42.1	17	69.1
VIRED	CLIP [42]	✓	**3**	**24.9**	**49.0**	**60.8**	**6.0**	**30.3**

Table 2: Comparison to the state-of-the-art on the video question answering task on the AVSD dataset. Our answer generator outperforms most prior methods and achieves comparable performance to the state-of-the-art Vx2Text [29] system.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
MA-VDS [18]	0.256	0.161	0.109	0.078	0.113	0.277
QUALIFIER [56]	0.276	0.177	0.121	0.086	0.119	0.294
Simple [49]	0.279	0.183	0.13	0.095	0.122	0.303
RLM [28]	0.289	0.198	0.145	0.11	0.14	0.337
VX2TEXT [29]	**0.311**	**0.217**	**0.16**	**0.123**	0.148	0.35
Ours	0.308	0.215	0.158	0.121	**0.149**	**0.351**

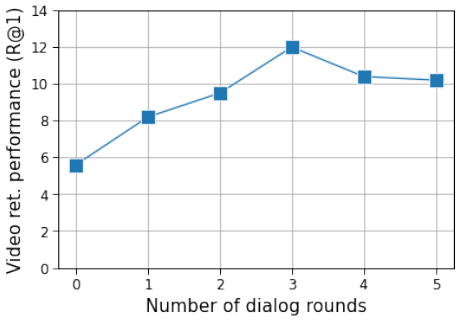

Figure 4: We study the video retrieval performance (R@1) as a function of the number of dialog rounds. Based on these results, we observe that the video retrieval accuracy consistently improves as we consider additional rounds of dialog. We also note that the performance of our interactive framework saturates after 3 rounds of dialog.

6 RESULTS AND DISCUSSION

6.1 Quantitative Video Retrieval Results

In Table 1, we compare our method with the previously described video retrieval baselines. We summarize our key results below.

6.1.1 The Importance of Pretraining. The results in Table 1 indicate that large-scale pretraining provides significant boost in video retrieval performance. Specifically, we note that the original Frozen-in-Time (FiT) baseline pretrained on large-scale WebVid2M [5] outperforms the previous state-of-the-art LSTM approach [35] by 1.4% according to R@1 even without using any dialog data. Furthermore, we also note that using CLIP4Clip [33] backbone with CLIP [42] pretraining leads to even better results, thus, indicating the importance of large-scale language-based pretraining.

6.1.2 Dialog Effectiveness. Next, we demonstrate that dialog is a highly effective cue for the video retrieval task. Specifically, we first show that the FiT baseline augmented with 10 rounds of human-generated ground truth dialog performs 5.2% better in R@1 than the same FiT baseline that does not use dialog (Table 1). This is a significant improvement that highlights the importance of additional information provided by dialog.

6.1.3 The Number of Dialog Rounds. Next, we observe that despite using only 3 rounds of dialog VIRED outperforms the strong FiT w/ Human Dialog baseline, which uses 10 rounds of ground truth dialog. It is worth noting that these 10 rounds of dialog were generated in a retrieval-agnostic manner (i.e., without any particular goal in mind), which may explain this result. Nevertheless, this result indicates that a few questions (e.g., 3) generated by our model are as informative as 10 task-agnostic human generated questions.

In Figure 4, we also plot the R@1 video retrieval accuracy as a function of the number of the dialog rounds. We observe that the performance of our system consistently improves as we use more dialog rounds. Furthermore, we note that the performance saturates after 3 rounds of interactions.

6.2 Video Question Answering Results

6.2.1 Comparison to the State-of-the-Art. As discussed above, we use our answer generator to simulate human presence in an interactive dialog setting. To validate the effectiveness of our answer generator, we evaluate its performance on the video question answering task on AVSD using the same setup as in Simple [49], and Vx2Text [29]. We present these results in Table 2 where we compare our answer generation method with the existing video question answering baselines. Our results indicate that our answer generation model significantly outperforms many previous methods, including MA-VDS [18], QUALIFIER [56], Simple [49] and RLM [28]. Furthermore, we note that our answer generator achieves similar

Table 3: To validate the effectiveness of our interactive framework in the real-world setting, we replace our automatic answer generator oracle with several human subjects. Specifically, we randomly select 50 videos from the AVSD dataset, and ask 3 subjects to provide answers to our generated questions. We then use those dialogs to measure video retrieval performance as before. Our results suggest that our framework generalizes effectively to the settings involving real human subjects. The video retrieval is performed only on the subset of 50 selected videos.

Method	R@1 (↑)	R@5 (↑)	R@10 (↑)	MedianR (↓)	MeanR (↓)
Answer Generator	43.8	79.2	91.7	2.0	3.6
Human Subject #1	45.2	81	92.7	2.0	3.5
Human Subject #2	**45.8**	**81.2**	**93**	**2.0**	**3.4**
Human Subject #3	45.6	**81.2**	92.9	**2.0**	**3.4**

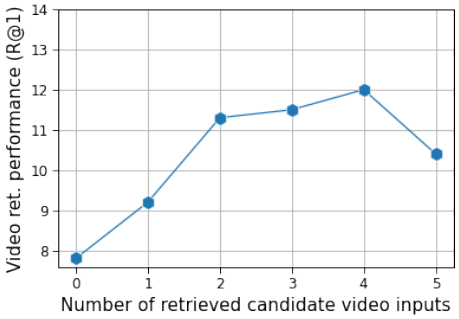

Figure 5: We investigate the video retrieval performance as a function of the number of retrieved candidate video inputs that are fed into the question generator. These results indicate that the video retrieval performance is the best when we use 4 retrieved videos as inputs to our question generator.

performance as the recent Vx2Text model [29]. These results indicate that our answer generator is comparable or even better than the state-of-the-art video-based question answering approaches.

6.2.2 Replacing Our Answer Generator with a Human Subject. To validate whether our interactive framework generalizes to the real-world setting, we conduct a human study where we replace our proposed answer generator with several human subjects. To do this, we randomly select 50 test videos from AVSD, and ask 3 human subjects to answer questions produced by our question generator. We then use the answers of each subject along with the generated questions as input to the video retrieval model (similar to our previously described setup). In Table 3, we report these results for each of 3 human subjects. These results suggest that our interactive framework works reliably even with real human subjects. Furthermore, we note that compared to the variant that uses an automatic answer generator, the variant with a human in the loop performs only slightly better, thus, indicating the robustness of our automatic answer generation framework. Note that in this case, the video retrieval is performed on the subset of 50 selected videos.

6.3 Ablation Studies

Next, we ablate various design choices of our model. Specifically, (i) we validate the effectiveness of our proposed Information-guided Supervision (IGS), (ii) the importance of using retrieved candidate videos for question generation, and (iii) how video retrieval performance changes as we vary the number of candidate video inputs to the question generator.

6.3.1 Effectiveness of IGS. To show the effectiveness of IGS, we compare the performance of our interactive video retrieval framework when using (i) IGS as a training objective to the question generator vs. (ii) using a video retrieval-agnostic objective. Specifically, we note the AVSD dataset has 10 pairs of questions and answers associated with each video. For the retrieval-agnostic baseline, we use the original order of the questions (i.e., as they appear in the dataset) to construct a supervisory signal for the question generator. In other words, we train our BART question generator to ask questions in the same order as the original AVSD human annotators did. In contrast, for our IGS-based objective, we order the questions such that they would maximize the subsequent video retrieval accuracy at each round of questions/answers. We report that IGS outperforms the retrieval-agnostic baseline by **4.3**% in R1 video retrieval accuracy, which is a significant boost. Thus, this result validates the effectiveness of our proposed IGS technique.

6.3.2 The Importance of Using Retrieved Videos for Question Generation. We also verify the importance of using retrieved candidate videos for the question generation process. Specifically, we compare the video retrieval performance (i) when the question generator uses top-k retrieved videos as part of its inputs and (ii) when it does not. Our results suggest that using top-k retrieved videos for question generation produces a substantial boost of **3.9**% in R1 video retrieval performance. This indicates the usefulness of the additional video input modality to our question generator. We use $k = 4$ in this experiment.

6.3.3 Ablating the Number of Video Inputs for Question Generation. Next, in Figure 5, we study the video retrieval performance as a function of the number of retrieved video inputs fed to the question generator. These results indicate that the performance gradually increases with every additional video candidate input and reaches the peak when using $k = 4$ retrieved videos. We also observe that the performance slightly drops if we set k larger than 4. We hypothesize that this happens because the input sequence length to the BART question generator becomes too long, potentially causing overfitting or other optimization-related issues.

6.3.4 Ablating the Choice of a Language Model. We also investigate the effect of using different language models for our question and answer generators. Specifically, we experiment with BART-Base, T5-Small, T5-Large, and BART-Large models. The results are shown in Table 4. From the results, it is evident that BART-large outperforms all the other models according to all evaluation metrics.

Table 4: We investigate the effectiveness of our question and answer generators with several different language models: BART-Base, T5-Small and T5-Large, and BART-Large. We evaluate our question generator with respect to the video retrieval task using using standard video retrieval metrics (the left part of the table). The results for the answer generator are evaluated using standard language generation metrics (see the right part of the table). Based on these results, we observe that BART-Large outperforms all the other variants according to all metrics.

Language Model	Question Generation for Video Retrieval					Answer Generation					
	R@1	R@5	R@10	MedianR	MeanR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
BART-Base	8.4	26.2	37.4	22	87.5	0.292	0.197	0.141	0.107	0.132	0.321
T5-Small	8.9	27.1	38.5	20	82.1	0.296	0.201	0.145	0.110	0.139	0.328
T5-Large	11.2	29.4	41.3	17.5	72.3	0.303	0.207	0.150	0.114	0.142	0.336
BART-Large	**12**	**30.5**	**42.1**	**17**	**69.1**	**0.308**	**0.215**	**0.158**	**0.121**	**0.149**	**0.351**

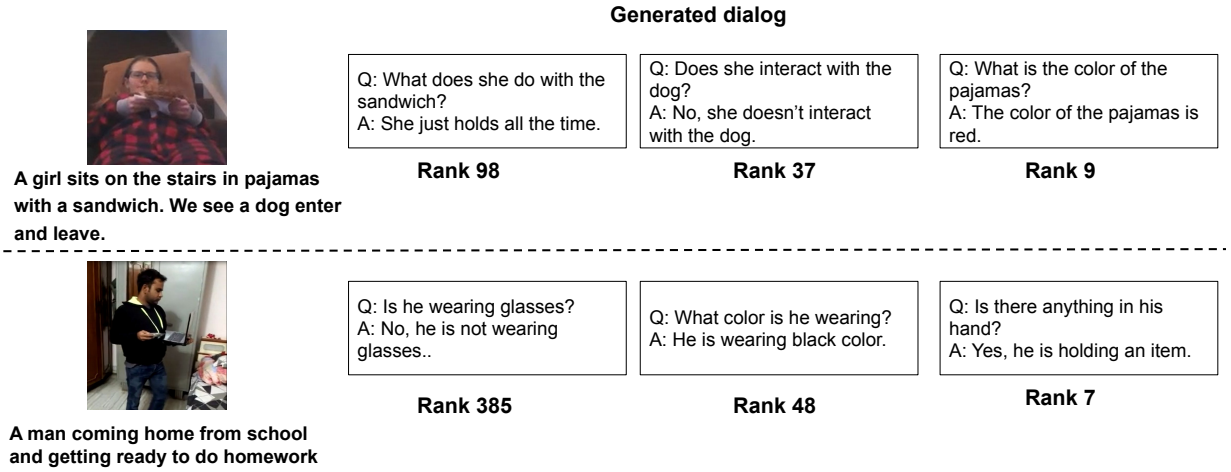

Figure 6: Qualitative results of our interactive video retrieval framework. On the left we illustrate the keyframe of the ground truth video V_{gt} (i.e., the video that the user wants to retrieve) and the initial textual query for that video. From left to right, we visualize the three rounds of our generated dialog history (using our question generator and the answer generator oracle). Furthermore, under each dialog box, we also illustrate the rank of the ground truth video V_{gt} among all videos in the database (i.e., the lower the better). Based on these results, we observe that each dialog round significantly improves video retrieval results (as indicated by the lower rank of the ground truth video). These results indicate the usefulness of dialog cues.

6.4 Qualitative Results

In Figure 6, we also illustrate some of our qualitative interactive video retrieval results. On the left we show the keyframe of the ground truth video V_{gt} (i.e., the video that the user wants to retrieve) and the initial textual query for that video. From left to right, we illustrate the three rounds of questions and answers produced by our question and answer generators. Additionally, under each question/answer box, we also visualize the rank of the video V_{gt} among all videos in the database (i.e., the lower the better, where rank of 1 implies that the correct video was retrieved).

Our results indicate a few interesting trends. First, we observe that each dialog round boosts video retrieval performance, which is indicated by the lower rank of the ground truth video. Second, we note that our question generator learns to ask question that produce new pieces of information (i.e., information not mentioned in the initial textual query). Lastly, we observe that the questions asked by our model focus on diverse concepts including gender, presence of certain objects, human actions, clothes colors, etc. This highlights the generality of our open-ended question generator.

7 CONCLUSIONS

In this work, we introduced VIRED , an interactive framework for video retrieval using dialog. We demonstrated that dialog provides valuable cues for video retrieval, thus, leading to significantly better performance compared to the non-interactive baselines. Furthermore, we also showed that our proposed (i) multimodal question generator, and (ii) information-guided supervision techniques provide significant improvements to our model's performance.

In summary, our method is (i) conceptually simple, (ii) it achieves state-of-the-art results on the interactive video retrieval task on the AVSD dataset, and (iii) it can generalize to the real-world settings involving human subjects. In the future, we will extend our framework to other video-and-language tasks such as interactive video question answering and interactive temporal moment localization.

ACKNOWLEDGMENTS

This research was partly funded by NSF grant IIS2133595 and by NIH grant 1R01AA02687901A1.

REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356* (2016).

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4971–4980.

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

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1728–1738.

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[7] Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, and Xing Sun. 2021. Ask&Confirm: Active Detail Enriching for Cross-Modal Retrieval With Partial Query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1835–1844.

[8] Huizhong Chen, Matthew Cooper, Dhiraj Joshi, and Bernd Girod. 2014. Multi-modal language models for lecture video retrieval. In *Proceedings of the 22nd ACM international conference on Multimedia*. 1081–1084.

[9] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. 2021. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11583–11593.

[10] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 326–335.

[11] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*. 2951–2960.

[12] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. 2021. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3354–3363.

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

[14] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathan Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, et al. 1995. Query by image and video content: The QBIC system. *computer* 28, 9 (1995), 23–32.

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

[16] Silvan Heller, Rahel Arnold, Ralph Gasser, Viktor Gsteiger, Mahnaz Parian-Scherb, Luca Rossetto, Loris Sauter, Florian Spiess, and Heiko Schuldt. 2022. Multi-modal interactive video retrieval with temporal queries. In *International Conference on Multimedia Modeling*. Springer, 493–498.

[17] Khanh Ho, Vu Xuan Dinh, Hong-Quang Nguyen, Khiem Le, Khang Dinh Tran, Tien Do, Tien-Dung Mai, Thanh Duc Ngo, and Duy-Dinh Le. 2022. UIT at VBS 2022: An Unified and Interactive Video Retrieval System with Temporal Search. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II* (Phu Quoc, Vietnam). Springer-Verlag, Berlin, Heidelberg, 556–561. https://doi.org/10.1007/978-3-030-98355-0_54

[18] Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, et al. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2352–2356.

[19] Adriana Kovashka and Kristen Grauman. 2013. Attribute pivots for guiding relevance feedback in image search. In *Proceedings of the IEEE International Conference on Computer Vision*. 297–304.

[20] Adriana Kovashka, Devi Parikh, and Kristen Grauman. 2012. Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2973–2980.

[21] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166* (2019).

[22] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. 2020. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. *arXiv preprint arXiv:2010.10095* (2020).

[23] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696* (2018).

[24] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574* (2019).

[25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[26] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200* (2020).

[27] Yang Li and Junier Oliva. 2021. Active feature acquisition with generative surrogate models. In *International Conference on Machine Learning*. PMLR, 6450–6459.

[28] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Bridging Text and Video: A Universal Multimodal Transformer for Audio-Visual Scene-Aware Dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 2476–2483.

[29] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. 2021. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7005–7015.

[30] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. 2012. Actions Speak Louder than Words: Searching Human Action Video Based on Body Movement. In *Proceedings of the 20th ACM International Conference on Multimedia* (Nara, Japan) *(MM '12)*. Association for Computing Machinery, New York, NY, USA, 1261–1262. https://doi.org/10.1145/2393347.2396432

[31] Jakub Lokoč, František Mejzlík, Tomáš Souček, Patrik Dokoupil, and Ladislav Peška. 2022. Video Search with Context-Aware Ranker and Relevance Feedback. In *International Conference on Multimedia Modeling*. Springer, 505–510.

[32] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[33] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860* (2021).

[34] Zhixin Ma, Jiaxin Wu, Zhijian Hou, and Chong-Wah Ngo. 2022. Reinforcement Learning-Based Interactive Video Search. In *International Conference on Multimedia Modeling*. Springer, 549–555.

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

[36] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 947–952.

[37] Ishan Misra, Ross Girshick, Rob Fergus, Martial Hebert, Abhinav Gupta, and Laurens Van Der Maaten. 2018. Learning by asking questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 11–20.

[38] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 19–27.

[39] Nils Murrugarra-Llerena and Adriana Kovashka. 2021. Image retrieval with mixed initiative and multimodal feedback. *Computer Vision and Image Understanding* 207 (2021), 103204.

[40] Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6679–6688.

[41] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10860–10869.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[43] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).

[45] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. 1998. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on circuits and systems for video technology* 8, 5 (1998), 644–655.

[46] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

[47] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108 (2019).

[48] Klaus Schoeffmann. 2014. A User-Centric Media Retrieval Competition: The Video Browser Showdown 2012-2014. *IEEE MultiMedia* 21, 4 (2014), 8–13. https://doi.org/10.1109/MMUL.2014.56

[49] Idan Schwartz, Alexander G Schwing, and Tamir Hazan. 2019. A simple baseline for audio-visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12548–12558.

[50] Hajin Shim, Sung Ju Hwang, and Eunho Yang. 2018. Joint active feature acquisition and classification with variable-size set encoding. In *Advances in neural information processing systems*. 1368–1378.

[51] Yuqing Song, Shizhe Chen, and Qin Jin. 2021. Towards Diverse Paragraph Captioning for Untrimmed Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11245–11254.

[52] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4223–4232.

[53] Minh-Triet Tran, Nhat Hoang-Xuan, Hoang-Phuc Trang-Trung, Thanh-Cong Le, Mai-Khiem Tran, Minh-Quan Le, Tu-Khiem Le, Van-Tu Ninh, and Cathal Gurrin. 2022. V-FIRST: A Flexible Interactive Retrieval System for Video at VBS 2022. In *International Conference on Multimedia Modeling*. Springer, 562–568.

[54] Hong Wu, Hanqing Lu, and Songde Ma. 2004. Willhunter: interactive image retrieval with multilevel relevance. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 2. IEEE, 1009–1012.

[55] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. 2020. Bert representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1556–1565.

[56] Muchao Ye, Quanzeng You, and Fenglong Ma. 2022. QUALIFIER: Question-Guided Self-Attentive Multimodal Fusion Network for Audio Visual Scene-Aware Dialog. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 248–256.

[57] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*.