Simple Baselines for Interactive Video Retrieval with Questions and Answers

Kaiqu Liang
Princeton University
kl2471@princeton.edu

Samuel Albanie
University of Cambridge
sma71@cam.ac.uk

Abstract

To date, the majority of video retrieval systems have been optimized for a "single-shot" scenario in which the user submits a query in isolation, ignoring previous interactions with the system. Recently, there has been renewed interest in interactive systems to enhance retrieval, but existing approaches are complex and deliver limited gains in performance. In this work, we revisit this topic and propose several simple yet effective baselines for interactive video retrieval via question-answering. We employ a VideoQA model to simulate user interactions and show that this enables the productive study of the interactive retrieval task without access to ground truth dialogue data. Experiments on MSR-VTT, MSVD, and AVSD show that our framework using question-based interaction significantly improves the performance of text-based video retrieval systems. Code is available at https://github.com/kevinliang888/IVR-QA-baselines.

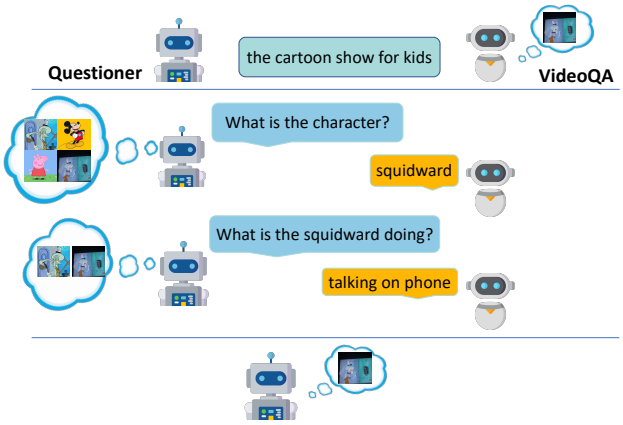

Figure 1: **Interactive video retrieval with questions and answers.** We revisit interactive video retrieval with a simple framework that employs Video Question Answering (VideoQA) to simulate user responses, sidestepping the need to collect dialogue data for development and evaluation.

1. Introduction

Given the surging popularity of video content, there is a pressing need to develop tools that enable users to search video collections accurately and efficiently. The *text-video retrieval* task aims to meet this need by ranking videos among a gallery according to how well they match a given text query. This task has seen tremendous progress in recent years, benefiting from large-scale pretraining datasets [47, 41], multimodal architectures [46, 38, 17] and robust optimisation strategies for cross-modal representation learning [45].

Much of the work to date has focused on a "single-shot" scenario in which a user submits a single query in isolation. While this may be applicable in some scenarios, there are many in which it would be useful to allow refinements to the initial ranking (particularly when the pool of valid responses to the initial query is large). To a large degree, this focus single-shot systems can be explained by practicality: obtaining the data required to develop and evaluate interactive retrieval systems is extremely challenging. User studies that engage individuals in interactive sessions are difficult to scale, while production data from product deployments is challenging to adapt to the development of new systems.

Although it has received less attention, the topic of interactive retrieval is far from new—it has been studied for more than two decades [53]. Following an initial query, various mechanisms have been offered to the user to enable them to refine results: relevance scores [53] (corresponding to highly relevant, irrelevant, etc.), relative attributes [29] (e.g. "shiner than these"), natural language descriptions [19, 57] (e.g. "unlike the given result, the one I want has fur on the back") and question-answer dialog [13, 43] (e.g. "Q: are there any people in the shot?" A: No").

Among dialogue-based approaches, the recent work of Madasu et al. [43] highlighted the potential of question-answering with free-form text questions to enhance video retrieval performance. At the heart of this approach is a question-generating system that probes the user for more details about their desired target in order to refine the retrieval ranking. In such a framework, user responses can be simulated by employing a question-answering model to synthesise plausible answers to the questions (Fig. 1). Our

work draws inspiration from ViReD [43], but differs in two key aspects. (1) The ViReD answer generator works by first captioning, then extracting the answer from the text of the caption. The disadvantage of this approach is that the answer indirectly depends on the visual content. Instead, we adopt a video question-answering model to simulate the answer, incorporating the information from the video directly with the question. (2) ViReD fine-tunes on AVSD with the human-collected dialogue while our approach does not require any fine-tuning. We show that our approach generalises well to multiple datasets, and obtains significantly better performance than ViReD.

In this work, we make the following contributions: (1) We propose a simple, yet effective framework for interactive video retrieval based on video question answering, and explore the design space of question generators within this framework. (2) We develop three simple question generators that leverage heuristics and learned models to gather discriminative information beyond objects (scenes, actions, colour etc.) in the target video. (3) Through careful experiments, we demonstrate that our system generalizes well to several video datasets and obtains significant improvement over baseline without fine-tuning. These results, which can be interpreted as leveraging an approximate interactive oracle, highlight the potential of question-based interactive retrieval frameworks to substantially enhance video search.

2. Related work

Text-to-video Retrieval A long-running theme of text-video retrieval research has been the development of effective fusion mechanisms for cross-modal learning [68, 25, 26, 67, 58]. Another theme of vision-language retrieval has sought to tackle challenges associated with hubness [52] in the embedding space [14, 5]. In the past few years, there has been increasing emphasis on the use of pre-trained models that enable retrieval systems to leverage large quantities of training data that is available for related tasks, often spanning multiple modalities [38, 46, 17, 11]. Recently, there has been a particular focus on building atop models that can leverage the most scalable forms of pretraining data (e.g. CLIP [51]), which have demonstrated compelling performance across a range of vision and language tasks. Luo *et al.* [40] proposed CLIP4CLIP which adapts CLIP for text-video retrieval through a combination of architectural changes and finetuning. Other approaches have included the use of a curriculum to learn from images and video [3], BERT-style architectures [56], and additional pretraining tasks [34]. Our approach is built upon BLIP [35], a unified framework for retrieval, question answering, and captioning. Differently from the works listed above, we focus on the interactive retrieval setting.

Visual Question Answering. The goal of visual question-answering systems is to answer questions posed in natural language that concern visual content in the form of an image or video. A broad range of work in this space has sought to combine spatial image representations and sequential question representations [2, 4, 16, 39, 61, 62, 66]. To answer questions about video content, spatio-temporal video representations that encode motion and appearance have been studied in detail [15, 18, 21, 22, 23, 24, 31, 33, 70]. There has also been work that extends these ideas to open-ended visual question answering [20, 64]. Recent work on goal-oriented visual dialog [12, 13, 32, 10, 49] uses policy-based reinforcement learning [49, 12, 13] or optimizing information gain [32] to improve the performance. However, these works require intensive training on human-annotated dialog datasets and do not trivially generalise to video. In terms of VideoQA, a few recent works have explored zero-shot approaches [64, 69], where models are only trained on automatically mined video clips with short text descriptions. Our VideoQA model is based on BLIP [35] which is trained on manually annotated image-question-answer triplets.

Interactive cross-modal retrieval A number of interactive image retrieval systems have been developed, drawing inspiration from visual dialogue [60, 53, 29, 28, 27, 50, 65]. In such systems, users provide feedback to an agent to guide it toward the desired retrieval target. Two families of feedback have been studied in detail: *relevance feedback* and *difference feedback*. For the former [60, 53], users provide a relevance score for the current retrieval results. The system then re-ranks its initial results using the feedback provided by the user. For the latter approach [29, 28, 27, 50, 65], users communicate the difference between the target image and a reference image to the system with cues such as tags or descriptions. The system then "whittles away" irrelevant images, iteratively filtering to select the desired target. However, one drawback of these systems is that they place a heavy time cost on the user who must describe what is missing in the initial search results. For example, description-based methods [19, 57] require users to provide detailed sentence-level feedback and tag-based methods require users to provide a collection of attributes [27, 28, 29].

To address these concerns, Cai *et al* [7] propose a partial query system in which potentially discriminative objects are suggested to the user with the goal of narrowing down the search results quickly. However, this system doesn't utilize other cues in the image such as object color, actions, scenes, background, etc., so it can require many interactions with the user to obtain good retrieval results. By using free-form questions to elicit information from the user, our system avoids this restriction to solely object-based queries.

Most related to our approach, the ViReD work of Madasu et al. [43] highlighted the potential of question-answering with free-form text questions to enhance video retrieval performance. However, rather than performing VideoQA directly, ViReD employs captioning followed by text-based

question-answering to answer questions about the video, bottlenecking answer quality. In addition, the ViReD question generator is complex and dependent on dialog history for training—our approach requires no fine-tuning and generalises effectively to multiple datasets.

3. Method

3.1. Interactive video retrieval with Q&A

Components. Our interactive video retrieval system contains four components: a *Text Encoder*, a *Visual Encoder*, a *Ranker*, and a *Q&A System*. The Text Encoder and Visual Encoder embed partial queries and videos into a shared textual-visual space. The Ranker computes the similarity scores between text and video features which are used to rank the videos. To accelerate inference, we adopt the approach of [36] that first selects K candidates based on the video-text feature similarity, and then reranks the selected candidates based on their pairwise Image-Text Matching (ITM) scores. Specifically, the model uses an ITM head (a linear layer) to predict the score indicating whether an image-text pair is positive (matched) or negative (unmatched) given their multimodal feature.

The Q&A System forms the central module in the framework (Fig. 3). It comprises a captioner adopted from the BLIP caption model, a question generator, and an answer generator that takes the form of a VideoQA (video question-answering) model that simulates user responses. We adopt the BLIP VideoQA model [35]. Instead of formulating VideoQA as a multi-answer classification task [9, 37], Li *et al* [35] interpret it as an answer generation task, enabling open-ended VideoQA. This open-ended VideoQA allows more freedom to generate and combine answers, lending flexibility to the system.

Interaction workflow. The complete interaction process is demonstrated in Fig. 2. The user provides an initial text query to the system. The interaction loop then begins, comprising four main steps. (1) *Generate input to the Q&A System*: The current (partial) query together with candidate retrieved videos is passed to the Q&A system. To obtain the candidate videos, we first generate the text embedding corresponding to the partial query and video embeddings for all videos. The Ranker then computes similarity scores in embedding space (via cosine similarity) to rank the videos. We select as candidate videos those who fall within the top K places of the ranking and feed them as input to the Q&A System. (2) *Question generation*: We explore several kinds of question generators (described in detail in Sec. 3.2 and Sec. 3.3). Different question generators require access to different inputs. For example, for the *Auto-text-vid* question generation approach (Sec. 3.3), we first generate captions for the candidate videos, then we generate questions based on both the captions and the partial query. (3) *User answer simulation by VideoQA model*: Given the question and the target videos, we use the VideoQA model to generate an answer. (4) *New query generation*: We generate a complete sentence by incorporating the answer and adding it to the original textual query. This completes one round of interaction.

Combining responses after each round of interaction After each round of interaction, we obtain new information with the potential to improve retrieval performance. It is therefore worth considering how this information is best integrated. One approach is to treat each new piece of information as a separate query, then adopt a strategy for multi-query retrieval. In recent work, Wang *et al* [59] propose two *post hoc* inference methods, *Similarity aggregation (SA)* and *Rank aggregation (RA)*, to extend single-query to multi-query video retrieval. However, initial experiments with these approaches did not prove promising. We believe this is because, while each new piece of information may provide complementary information, it may perform poorly as a query in isolation. However, SA and RA work best with individual queries of high quality. We, therefore, adopt a different approach: we aggregate pieces of information by concatenating them into one large query. In this way, the initial query can be iteratively refined by adding more complementary information. In detail, we concatenate multiple queries using a separation token `[SEP]`. This token is derived from BERT pretraining, in which it is used to separate sentences as follows: "`[CLS]` sentence A `[SEP]` sentence B `[SEP]`", suggesting its applicability for separating content in our application setting.

3.2. Heuristic Question Generator

We now turn the design of question generators used in our system, starting with a simple heuristic question generator.

Designing questions to improve retrieval performance. The problem of asking useful questions about visual content has received attention in the field of Visual Question Generation. Mostafazadeh *et al* [48] proposed a framework that generates questions that can potentially engage a human in starting a conversation. They observe that asking a question that can be answered simply by looking at the image would be of interest to the computer vision community, but those questions are neither natural nor engaging for a person to answer. Nevertheless, for the purpose of improving retrieval performance, we focus on concrete, discriminative questions (rather than abstract and engaging questions). In particular, to develop a heuristic question generator, we target goal-driven questions [30] that focus on extracting objects ("what is the person throwing?"), attributes ("what kind of shirt is the person wearing?"), color ("what color is the frisbee?"), material ("what material is the frisbee?"), etc.

We are particularly interested in obtaining information that is highly discriminative and complementary to the original query. Ask & Confirm [7] argues that objects provide discriminative cues to distinguish a target from other candi-

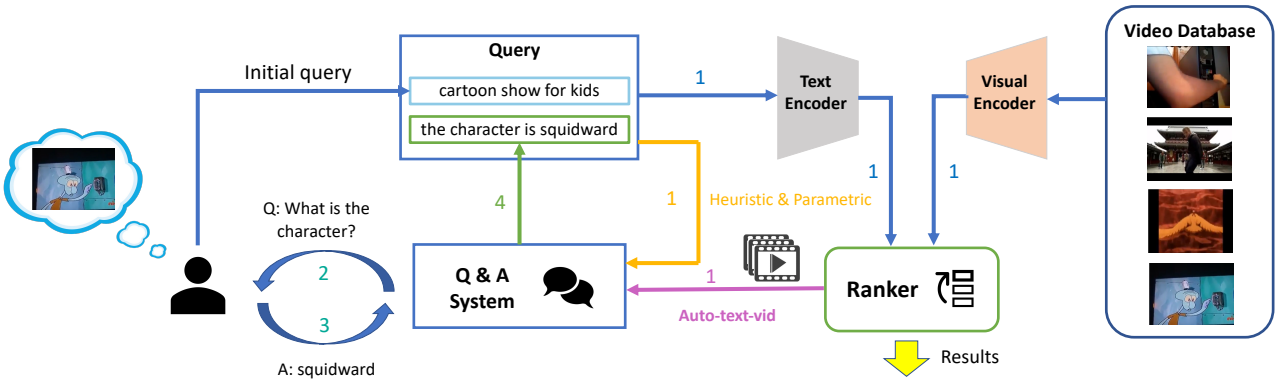

Figure 2: **Model architecture of interactive video retrieval with Q&A (Questions and Answers)**. The user provides an initial textual query to the system. The interaction loop contains 4 main steps. Step 1: Generate input to the Q&A system. For Heuristic and Auto-text question generation approaches, the input to the Q&A system are just the queries. For Auto-text-vid question generation approach, the input contains both queries and the candidate videos. Step 2: Question generation. Step 3: User answer simulation by VQA model. Step 4: New query generation. This fulfills one round of the interaction.

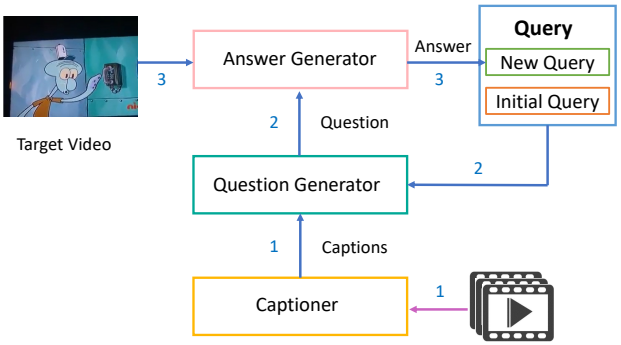

Figure 3: **Model architecture of Q&A System**. The interaction loop contains either two or three steps depending on which question generation approach is used. Step 1: Caption generation. Step 2: Question generation. Step 3: Answer generation. Auto-text-vid approach contains step 1, 2 and 3 while Heuristic and Auto-text question generation approaches contain only step 2 and 3.

dates in the context of image retrieval. Similarly, we observe the importance of the objects in the video retrieval task. However, other cues such as color, human actions, and scenes are also important, and we, therefore, aim to construct a question generator that incorporates these cues.

Heuristic Question Generator. Our first approach to question generation employs hand-crafted heuristics. Our strategy is to first obtain objects from the initial textual query (via pattern matching) and then ask questions about these objects. If an object is a living thing such as a person or animal, then we enquire about both the action and scene associated with it. If the object is a non-living thing such as a lamp or table, then we ask only the scene. If no object is found in the initial textual query, we ask the user to identify an object in the scene. The follow-up questions then proceed as above.

Since living things are often associated with actions, if the object is a living thing, then we employ a template of the form "who is doing what in what place". If the object is a non-living thing, the template used is "who is in what place". Once the answer is obtained, it is appended to the question, then both are appended to the current query for the next round of video retrieval.

One potential downside of this approach is the redundancy between questions and the query. For example, suppose the initial textual query is "a man is singing", then when we ask "what is the man doing?", the answer may add no new information. In practice, this is typically not the case. We find that the information of the initial query is often restricted to addressing some temporal window within the video.

However, the question provided to the user can often address a larger portion of the video. For the example above, the man may walk through the street in the following frames after singing. The answer generator could encode this information to provide a complementary cue for retrieval.

3.3. Parametric Question Generators

The heuristic question generator captures intuition about what makes for useful questions but lacks flexibility. We therefore explore question generation mechanisms that leverage a large language model that possesses the ability to ask flexible, open-ended questions. The first, *Auto-text*, asks questions based only on the text query. The second, *Auto-text-vid*, asks questions based on both the text query and the top-K ranked video candidates for the current query (here K is a hyperparameter of the system).

The foundation of both question generators is T-0++ [54], a large language model that outperforms or matches GPT-3 [6] on many NLP tasks. Since this language model has been trained on various forms of question answering and shown to be very effective at zero-shot task generalization, we directly apply this model to generate questions without finetuning.

Auto-text. To generate questions from the query, we provide the following template to the language model: "*Suppose you are given the following video descriptions Q_i, What question would you ask to help you unique identify the video?*" where Q_0 is the initial query and Q_i is the concatenated query used to retrieve videos after the i^{th} round of interaction. The new query after each round incorporates the answer to the question through concatenation. Note that in terms of the Auto-text question generation approach, the question depends both on the original query and the new descriptions generated after each round of the interaction.

Auto-text-vid. Differently from Auto-text, Auto-text-vid conditions its question generation on both the text query and the current top-k ranked retrieved videos (selected via cosine similarity for the query corresponding to the current round of interaction). This is because, intuitively, we expect the top-ranked videos selected at each round to share some common information with the target video. The use of the top-k retrieved videos as conditioning information was also considered by [43], who found it beneficial.

To proceed, we first generate captions for each of the top-ranked videos using the BLIP captioner [35]. Then we use both these captions and the original text query to generate questions. The input sentence to the language model employs the following template: *"Suppose you are given the following video descriptions: C_i. What question would you ask to help you unique identify the video described as follows: Q_i?"* where C_i are the generated captions at the i^{th} round of interaction and Q_i is the concatenated textual query used to retrieve videos after the i^{th} round of interaction. The questions generated are based on those common features, so the machine obtains more information about the target video through the interaction.

Remark. Since the top-ranked videos change at each round, there is some variation in the questions generated by *Auto-text-vid*. This may provide complementary cues, but also carries the risk of additional noise (both through mistakes in the synthetic captions and through mismatches in content between the candidates and the target).

3.4. Question Augmentation

To further increase the information extracted from the user about the target video, we also introduce two further simple techniques to augment the questions: *Ask Segment* (AS) and *Ask Object* (AO).

Ask Segment (AS) augmentation asks the user questions about *each half of the video* independently. Videos evolve over time and different objects and actions may take place in different frames. As a consequence, asking the same question about different halves of the video often yields very different answers.

Ask Object (AO) augmentation enquires about more objects in the scene. This approach is motivated by Ask & Confirm [7] in which Cai *et al* observe that objects are often highly discriminative. Rather than searching for discriminative objects as candidates through a complex RL policy [7], we simply ask the user which objects are in the scene.

Note that these strategies can be combined to generate more discriminative questions for interactive video retrieval.

4. Experiment

In this section, we first describe the datasets used in our experiments (Sec. 4.1). We then discuss implementation details (Sec. 4.2) and evaluation metrics (Sec. 4.3).

4.1. Datasets

MSR-VTT [63] is a video benchmark that is widely used for text-video retrieval. It contains 10,000 videos sourced from YouTube. Each video is annotated with 20 natural language descriptions. Following previous work, we report retrieval results on the 1K-A test set.

MSVD [8] is a dataset initially collected for translation and paraphrase evaluation. It contains 1,970 videos, spanning a duration from 1 second up to 62 seconds. The train, validation, and test splits contain 1,200, 100, and 670 videos, respectively. Each video contains approximately 40 associated sentences in English.

AVSD (Audio-Visual Scene aware Dialog dataset) [1] is a collection of videos associated with ground truth dialog data. Each video dialog consists of 10 rounds of human-generated questions and answers describing various details about the video content. The AVSD dataset consists of 7985 training samples and 1863 validation samples. We follow prior work [43] and split the validation samples into a new validation split of 863 samples and 1000 test samples. Note that our approach does not make use of the dialog data, since we simulate the user directly. We use only the caption as the initial textual query to retrieve the video.

4.2. Implementation Details

For the visual encoder, we employ a vision transformer with ViT-B/16 which contains 12 layers and 12 heads. This model has a hidden dimensionality of 768, an MLP size of 3072, and an input patch size of 16×16. The text transformer used for text encoder is $\mathrm{BERT}_{\mathrm{BASE}}$ with 12 layers, 12 self-attention heads, and a hidden size of 768. The text transformer used for the image-grounded text encoder is the same as the text encoder except that it uses additional cross-attention layers to model vision-language interactions.

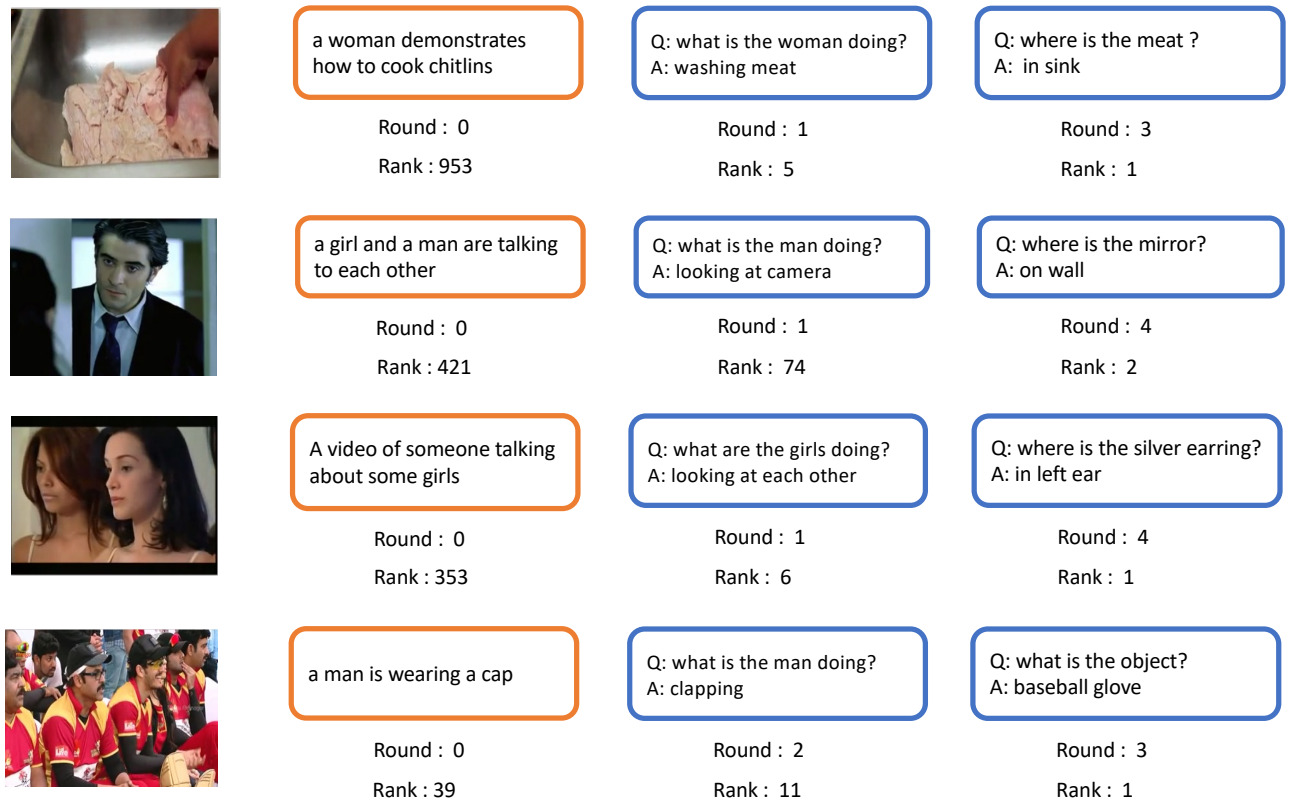

Figure 4: **Qualitative results of the dialogue generated by our Q & A System on MSR-VTT.** The questions are generated by the Heuristic Question Generator. The initial queries of the target videos are shown in orange boxes. The questions and answers generated through each interaction are shown in blue boxes. Both the initial retrieval rank and the rank after the interaction are demonstrated. In each case, we observe a boost in performance. Note, however, that the answers are not always correct (e.g. in the fourth row "A: baseball glove" when the sport does not appear to be baseball). However, since the answer is "nearly correct", it still provides a useful cue to the retrieval model.

The architecture of the image-grounded text decoder is similar to the image-grounded text encoder but it replaces the bi-directional self-attention layers with causal self-attention layers. All videos are resized to 384×384 as input.

In our interactive system, we adopt the BLIP retrieval model, the VQA model, and the captioner. The BLIP retrieval model is composed of an image encoder, a text encoder, and an image-grounded text encoder. The BLIP VQA model is composed of an image encoder, an image-grounded text encoder, and an image-grounded text decoder. The BLIP captioner contains an image encoder and an image-grounded text decoder. To perform zero-shot transfer video retrieval, we uniformly sample 8 frames per video and concatenate the frame features into a single sequence. To generate questions, we adopt T0++ [54], a powerful encoder-decoder language model trained on a collection of English-prompted datasets. This model attains strong zero-shot task generalization on English natural language prompts (outperforming GPT-3 on many tasks while being 16x smaller).

The Auto-text-vid question generation approach depends on the current top-k ranked retrieved videos. We experimented with different k (3, 5, 10, 15) and observed that when k = 5, we obtained the best performance. This may be because setting k to be too large produces too much noise in the generated captions while setting k to be too small does not provide an opportunity to extract all potentially discriminative information from related videos.

Baselines. In addition to the BLIP baseline, we also experimented with the interactive approach in [43] for fair comparisons. As public code is not available at the time of writing, we simulate their answer generation by first captioning the target video (by BLIP captioner) and then using the large language model (T0++) to generate the answer based on the caption and question. On the AVSD dataset, we also experimented with the approach in [44, 43] using 10 rounds of manually annotated human dialogue history for each video. However, simply concatenating all 10 rounds of ground truth dialog with the initial text query got a slightly

worse performance than the baseline. Instead, we concatenated all the answers with the initial text query for retrieval.

4.3. Evaluation

Metrics. We adopt evaluation metrics that are widely used in the standard single-query retrieval setting and report recall@K (R@K, higher the better) and median rank (MdR, lower the better) in our experiments. R@K calculates the percentage of test data for which the ground-truth video is found in the retrieved K videos. MdR metric captures the median rank of the retrieved ground truth videos.

Time cost. We next assess the time cost of producing answers during the interaction. For this, we sample 50 videos from the MSR-VTT dataset. We aim to compare our answer generation approach using the VideoQA model with the approach in ViReD [43] (We call this answer generation approach *CAP+LM*). In addition, we also conduct a user study on answer generation. We invite 4 users to answer the auto-generated question (Auto-text) on the target videos and record the time they spent on each question. (Note that each user went through all 50 videos before the interaction.)

5. Results

Method	ITA	R1 ↑	R5 ↑	R10 ↑	MdR ↓
FiT [3]	×	31.9	61.1	72.6	3
CLIP4Clip [40]	×	41.5	69.4	79.3	2
CenterCLIP [71]	×	48.4	73.8	82.0	2
X-CLIP [42]	×	49.3	75.8	84.8	2
BLIP [35]	×	43.3	65.6	74.7	2
BLIP + ViReD [43]	√	47.0	72.3	81.4	2
BLIP + Auto-text w/o AO	√	47.6	72.8	81.8	2
BLIP + Auto-text-vid w/o AO	√	49.0	72.9	80.0	2
BLIP + Auto-text	√	62.2	84.5	90.7	1
BLIP + Auto-text-vid	√	57.6	82.2	88.7	1
BLIP + Heuristic	√	**67.9**	**89.5**	**94.9**	**1**

Table 1: Comparison to state-of-the-art methods for text-to-video retrieval on the 1k test split of MSRVTT dataset. ITA indicates whether it is an interactive retrieval system. Recall metrics are reported in percent.

In this section, we discuss experimental results. On MSR-VTT (Tab. 1), without AO (asking objects), we observe that both the Auto-text and Auto-text-vid question generation approaches outperform the BLIP baseline. We observe that Auto-text-vid, which makes use of the top-ranked retrieved videos at each round, performs better than Auto-text. With AO, there is a significant performance boost for both approaches. For instance, we observe that the Auto-text Q&A System outperforms the baseline by +18.9% in recall@1.

Method	ITA	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
FiT [3]	×	33.7	64.7	76.3	3
CLIP4Clip [40]	×	46.2	76.1	84.6	2
X-CLIP [42]	×	50.4	80.6	-	-
CenterCLIP [71]	×	50.6	80.3	88.4	1
BLIP [35]	×	44.2	73.7	80.9	2
BLIP + ViReD [43]	√	51.0	78.1	87.2	1
BLIP + Auto-text	√	68.7	91.2	95.8	1
BLIP + Auto-text-vid	√	66.1	88.8	93.7	1
BLIP + Heuristic	√	**74.2**	**93.4**	**97.9**	**1**

Table 2: Comparison to state-of-the-art methods for text-to-video retrieval on the test split of the MSVD dataset. ITA indicates whether it is an interactive retrieval system.

Method	ITA	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
FiT [3]	×	5.6	18.4	27.5	25
LSTM [44]	√	4.2	13.5	22.1	-
FiT + Human dialog [43]	√	10.8	28.9	40.0	18
ViReD [43]	√	24.9	49.0	60.8	6
BLIP [35]	×	27.3	50.0	61.2	5.5
BLIP + Human dialog [43]	√	27.4	50.4	62.7	5
BLIP + ViReD [43]	√	28.4	54.3	66.8	4
BLIP + Auto-text	√	36.0	61.6	71.6	3
BLIP + Auto-text-vid	√	36.5	61.0	71.1	3
BLIP + Heuristic	√	**39.5**	**69.1**	**77.8**	**2**

Table 3: Comparison to state-of-the-art methods for text-to-video retrieval on the test split of AVSD dataset. ITA indicates whether it is an interactive retrieval system.

	VideoQA	CAP+LM	User
Avg time (s)	0.1	8.0	5.7

Table 4: Comparison of the average time cost of different answer and question generation approaches. We show the results of the answer generation approach of VQA (ours) and CAP+LM (simulation of ViReD [43] answer generator) and the user study.

Interestingly, we observe that Auto-text outperforms Auto-text-vid when used in combination with AO. This suggests that simply asking the user directly for the objects that they are interested in (rather than inferring them from closely ranked videos) offers higher signals.

We also observe that the Heuristic question generator yields dramatic gains: the recall@1 is improved by 24.6% (from 43.3% to 67.9%), 30% (from 44.2% to 74.2%) and

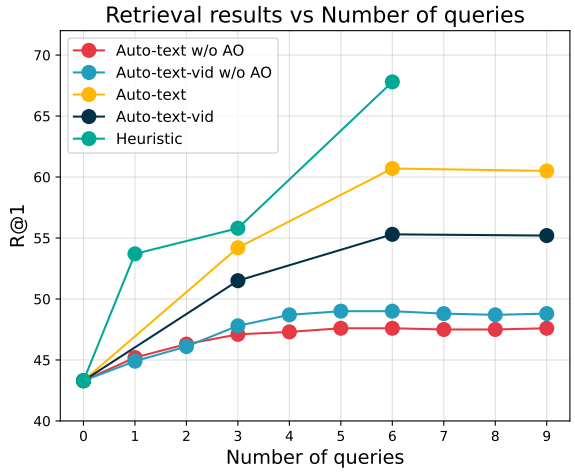

Figure 5: **Retrieval result vs Number of queries on MSRVTT.** Performance of all different approaches improves as the number of rounds increases. Heuristic question generation approach achieves the most significant improvement.

12.2% (from 27.3% to 39.5%) on MSR-VTT, MSVD and AVSD respectively. Using Auto-text question generator, the BLIP baseline recall@1 is improved by 18.9% (from 43.3% to 62.2%), 24.5% (from 44.2% to 68.7%), and 8.7% (from 27.3% to 36.0%) on MSR-VTT, MSVD, and AVSD respectively. This indicates that generated questions elicit useful information from the (simulated) user that substantially improves retrieval performance. Compared to prior interactive approaches (ViReD, Human dialog) on AVSD, our approach does not use dialog training data for finetuning but nevertheless outperforms existing work. Interestingly, the heuristic approach performs best over all three different datasets.

As seen in Fig. 4, a few rounds of interaction are sufficient to locate the large video. We demonstrate the cases where the initial query is incomplete, hindering performance. However, using the question-answering system, the search results quickly improve and we obtain significantly better retrieval performance. The generated questions are of high quality, aiming to confirm discriminative information from the video. In addition, the answers generated by VideoQA model are reasonable in the majority of cases (though not always perfect). They successfully capture different aspects of the information in the video (such as action, location, and object). Another useful property of the generated answers is that they are short phrases, simulating concise user answers.

Number of queries. Additional rounds of interaction lead to additional new queries. We observe in Fig. 5 that the retrieval performance increases as the number of queries increases. This effect is clearest for Heuristic question generation approach. The performance of both Auto-text and Auto-text-vid increases faster than their variants without AO. This further confirms that asking users about objects in the videos is an effective approach to improve retrieval performance at each round of the interaction. We observe that without AO, recall@1 of Auto-text-vid increases faster than that of Auto-text while the opposite result is observed with AO. This agrees with our argument "simply asking the user directly for the objects that they are interested in (rather than inferring them from closely ranked videos) offers higher signal". For all the approaches, Performance stabilizes as the number of queries grows beyond 6. Note that the maximum number of queries for heuristic question generation is fixed to 6 and using 6 rounds works best.

Computational cost. As shown in Tab. 4, Our approach using the VideoQA model is significantly more efficient than CAP+LM. This is as expected because two steps (captioning and question-answering) are involved in CAP+LM approach while VideoQA directly incorporates the video and question information to generate answers. In our naive implementation, CAP+LM is even slower than the user response (though this could be optimised).

6. Discussion

Limitations. (1) T0++ [55] is good at question-answering tasks but might not be good at generating questions. Therefore fine-tuning the model for the question generation task could be helpful. (2) We adopt the BLIP model for both retrieval and VideoQA. In principle, there is the potential for "leakage" in which the VideoQA model encodes information that is specific to BLIP, but irrelevant to the video. In practice, different parameter checkpoints are used for different VideoQA and retrieval components, so we believe that such leakage is unlikely to occur in practice.

Broader Impact. Improved video retrieval has a wide range of societally useful applications, spanning domains such as education, entertainment, and security. However, as with many technologies, it is subject to dual use. For instance, it could be used to facilitate unlawful surveillance.

7. Conclusion

In this work, we proposed several simple baselines for interactive video retrieval based on the generation of questions and answers. Through experiments on standard benchmarks, we demonstrated the effectiveness of generated questions in combination with a VideoQA model to simulate user responses through the interaction process. In future work, we hope to explore the use of question-answering systems to simulate users in other retrieval tasks.

Acknowledgements. SA would like to acknowledge Z. Novak and N. Novak in enabling his contribution, and support from an Isaac Newton grant and an EPSRC HPC grant. KL would like to thank Bill Byrne, Olga Russakovsky, Yu Wu, Vishaal Udandarao, Allison Chen, Xindi Wu, and Zhiwei Deng for helpful discussions and feedback.

References

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

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 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*, pages 1728–1738, 2021. 2, 7

[4] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017. 2

[5] Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5205, 2022. 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5

[7] Guanyu Cai, Jun Zhang, Xinyang Jiang, Yifei Gong, Lianghua He, Fufu Yu, Pai Peng, Xiaowei Guo, Feiyue Huang, and Xing Sun. Ask&confirm: active detail enriching for cross-modal retrieval with partial query. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1835–1844, 2021. 2, 3, 5

[8] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. 5

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 3

[10] Michael Cogswell, Jiasen Lu, Rishabh Jain, Stefan Lee, Devi Parikh, and Dhruv Batra. Dialog without dialog data: Learning visual dialog agents from vqa data. *Advances in Neural Information Processing Systems*, 33:19988–19999, 2020. 2

[11] Ioana Croitoru, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie, and Yang Liu. Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11583–11593, 2021. 2

[12] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 2

[13] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2951–2960, 2017. 1, 2

[14] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014. 2

[15] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1999–2007, 2019. 2

[16] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 2

[17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 1, 2

[18] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018. 2

[19] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based interactive image retrieval. *Advances in neural information processing systems*, 31, 2018. 1, 2

[20] Hexiang Hu, Wei-Lun Chao, and Fei Sha. Learning answer embeddings for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5428–5436, 2018. 2

[21] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11021–11028, 2020. 2

[22] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 2

[23] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11101–11108, 2020. 2

[24] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings*

of the AAAI Conference on Artificial Intelligence, volume 34, pages 11109–11116, 2020. 2

[25] Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf. Temporal tessellation: A unified approach for video analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 94–104, 2017. 2

[26] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[27] Adriana Kovashka and Kristen Grauman. Attribute pivots for guiding relevance feedback in image search. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 297–304, 2013. 2

[28] Adriana Kovashka and Kristen Grauman. Attributes for image retrieval. In *Visual Attributes*, pages 89–117. Springer, 2017. 2

[29] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980. IEEE, 2012. 1, 2

[30] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Information maximizing visual question generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2008–2018, 2019. 3

[31] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 2

[32] Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. Answerer in questioner's mind: Information theoretic approach to goal-oriented visual dialog. *Advances in neural information processing systems*, 31, 2018. 2

[33] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 2

[34] 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*, pages 4953–4963, 2022. 2

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 3, 5, 7

[36] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[37] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 3

[38] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019. 1, 2

[39] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016. 2

[40] 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. *arXiv preprint arXiv:2104.08860*, 2021. 2, 7

[41] 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. 1

[42] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 7

[43] Avinash Madasu, Junier Oliva, and Gedas Bertasius. Learning to retrieve videos by asking questions. *arXiv preprint arXiv:2205.05739*, 2022. 1, 2, 5, 6, 7

[44] 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*, pages 952–953, 2020. 6, 7

[45] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 1

[46] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 1, 2

[47] 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*, pages 2630–2640, 2019. 1

[48] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*, 2016. 3

[49] Aishwarya Padmakumar and Raymond J Mooney. Dialog policy learning for joint clarification and active learning queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13604–13612, 2021. 2

[50] Devi Parikh and Kristen Grauman. Relative attributes. In *2011 International Conference on Computer Vision*, pages 503–510. IEEE, 2011. 2

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[52] Milos Radovanovic, Alexandros Nanopoulos, and Mirjana Ivanovic. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(sept):2487–2531, 2010. 2

[53] Yong Rui, Thomas S Huang, Michael Ortega, and Sharad Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on circuits and systems for video technology*, 8(5):644–655, 1998. 1, 2

[54] Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*, 2022. 5, 6

[55] Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. 8

[56] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 2

[57] Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, and Vicente Ordonez. Drill-down: Interactive retrieval of complex scenes using natural language queries. *Advances in neural information processing systems*, 32, 2019. 1, 2

[58] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 2

[59] Zeyu Wang, Yu Wu, Karthik Narasimhan, and Olga Russakovsky. Multi-query video retrieval. *arXiv preprint arXiv:2201.03639*, 2022. 3

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

[61] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406. PMLR, 2016. 2

[62] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European conference on computer vision*, pages 451–466. Springer, 2016. 2

[63] 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*, pages 5288–5296, 2016. 5

[64] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 2

[65] Xinru Yang, Haozhi Qi, Mingyang Li, and Alexander Hauptmann. From a glance to" gotcha": Interactive facial image retrieval with progressive relevance feedback. *arXiv preprint arXiv:2007.15683*, 2020. 2

[66] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 2

[67] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 2

[68] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173, 2017. 2

[69] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 2

[70] Zheng-Jun Zha, Jiawei Liu, Tianhao Yang, and Yongdong Zhang. Spatiotemporal-textual co-attention network for video question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 15(2s):1–18, 2019. 2

[71] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. *arXiv preprint arXiv:2205.00823*, 2022. 7