Retrieving-to-Answer: Zero-Shot Video Question Answering with Frozen Large Language Models

Anonymous ICCV submission

Paper ID 6

Abstract

Video Question Answering (VideoQA) has been significantly advanced from the scaling of recent Large Language Models (LLMs). The key idea is to convert the visual information into the language feature space so that the capacity of LLMs can be fully exploited. Existing VideoQA methods typically take two paradigms: (1) learning cross-modal alignment, and (2) using an off-the-shelf captioning model to describe the visual data. However, the first design needs costly training on many extra multi-modal data, whilst the second is further limited by limited domain generalization. To address these limitations, a simple yet effective **Retrieving-to-Answer** *(R2A) framework is proposed. Given an input video, R2A first retrieves a set of semantically similar texts from a generic text corpus using a pretrained multi-modal model* (e.g., *CLIP). With both the question and the retrieved texts, a LLM* (e.g., *DeBERTa) can be directly used to yield a desired answer. Without the need for cross-modal fine-tuning, R2A allows for all the key components* (e.g., *LLM, retrieval model, and text corpus) to plug-and-play. Extensive experiments on several VideoQA benchmarks show that despite with 1.3B parameters and no fine-tuning, our R2A can outperform the 61× larger Flamingo-80B model [1] even additionally trained on nearly 2.1B multi-modal data. The code will be released.*

1. Introduction

Video Question Answering (VideoQA) aims to answer a question regarding a reference video [55]. Due to the open-end nature, manually annotating a large comprehensive dataset dedicated for VideoQA is practically impossible [25,47,48]. An appealing approach to address this challenge is *zero-shot learning* as pioneered by recent attempts [1, 34, 50, 51, 59]. Instead of training a task-specific model, they resort to learn a general-purpose multi-modal model using strong pretrained Large Language Models (LLMs) [5, 10, 15, 28, 37, 40, 52], because LLMs can accommodate rich knowledge from text data at scale.

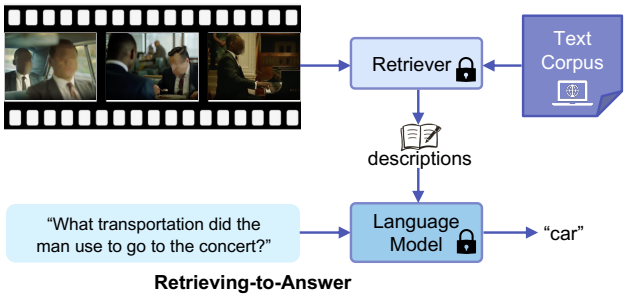

Figure 1. Overview of our **Retrieving-to-Answer** (R2A) framework for zero-shot video question answering. Given a reference video, R2A efficiently retrieves a set of semantically similar texts from an external corpus (*e.g.*, WebVid [2]). With both the retrieved texts and the question, a pretrained language model can be used directly to yield the final answer. Without additional training, R2A allows any component to plug-and-play.

To capitalize off-the-shelf LLMs for VideoQA, the key lies in how to bridge the gap between texts and videos effectively. There are two existing paradigms. The ***first*** adopts a cross-modal alignment strategy that projects visual features into soft prompts in the text embedding space. This design has two limitations: (1) High training cost due to both large-sized model and data [2, 35, 46, 49, 63]. (2) Less flexibility in component upgrading, as changing any component requires model retraining. To avoid both issues, the ***second*** paradigm instead connects the vision and text modalities by using an off-the-shelf caption model to convert the reference video into textual description [51]. However, this method relies on caption models finetuned towards the target domain making them less generalizable.

More broadly in image generation, instead of generating visual elements from scratch, leveraging relevant elements retrieved from a large image collection could facilitate the synthesis of complex scenes [18]. Connecting with the captioning-based paradigm above, we draw an analogy that rather than generating captions using a caption model, retrieving from a text corpus could be an alternative way to

obtain related text descriptions for the video. To that end, a strong receiver and a comprehensive text corpus are needed. For the former, we see the potential in recent tremendous advances in contrastive multi-modal models (*e.g.*, CLIP) showing remarkable abilities in zero-shot cross-modal retrieval [40]. The latter is generally available, *e.g.*, large diverse texts from the internet.

Under the above analysis and insight, we propose a simple yet effective ***Retrieving-to-Answer*** (R2A) framework. Instead of costly video captioning, we resort to more efficient cross-modal text retrieval from a generic text corpus (*e.g.*, uncurated web data WebVid [2]), simply leveraging a pretrained contrastive multi-modal model (*e.g.*, CLIP [40]). Given an input video, we first retrieve a pool of semantically similar texts from the corpus in the multi-modal model's feature space. This step can be considered as semantic video summarization. With the retrieved texts and the question, a pretrained large language model (*e.g.*, DeBERTa [15]) can be then directly applied to generate the answer. R2A achieves state-of-the-art performance on multiple VideoQA benchmarks and simultaneously addresses the limitations of previous paradigms: (1) Our modular design allows our R2A to accommodate readily available pretrained models for the VideoQA task without the need for fine-tuning. (2) Our R2A is able to generalize to new tasks/domains without further adaption because of both the selected powerful LLMs and multi-modal foundation model [3] (*e.g.*, CLIP [40] with remarkable generalization ability validated on novel domains and tasks [16, 66]), and the usage of a large diverse text corpus. Moreover, as the text corpus just needs to be encoded once, the whole retrieval process of R2A is fast (*e.g.*, with naive implementation, retrieving 10 captions per video only takes 3.5ms on a 10M-sized text corpus).

Our **contributions** are summarized as follows: (I) We propose a novel idea of text retrieval for zero-shot video question answering, in contrast to previous cross-modal alignment learning and video captioning strategies. (II) We introduce a simple, more efficient, yet more performing *Retrieving-to-Answer* (R2A) framework, without additional fine-tuning whilst being fully open to the selection and change of any components. (III) Extensive experiments show that our R2A achieves new state-of-the-art performance on multiple benchmarks in the zero-shot setting. In particular, with only 1.3B parameters and no additional tuning, our model outperforms the cross-modal training-based Flamingo [1] with 80B parameters.

2. Related Work

2.1. Recent advances of VideoQA

Video Question Answering (VideoQA) has gained increasing attention due to its wide applications in video search, summarization, and understanding. Involving natural language comprehension, question answering, and video processing, this task presents a number of typical multi-modal learning challenges simultaneously. Many prior methods rely on supervised model learning from labeled VideoQA datasets [8, 9, 23, 24, 30, 38, 42, 44, 45, 54]. Due to the limited size of manually labeled data, the resulting models are less capable and generalizable across domains. To mitigate this obstacle, recent methods adopt a paradigm of first pretraining on large vision-language data and then fine-tuning on the target small training set [11,22,25,58,62]. This approach still focuses on task-specific settings with limited domains involved. For more domain-generalizable VideoQA, zero-shot learning has recently shown potential, with the promising ability to scale to previously unseen samples with zero supervision [1, 58, 59, 62, 63]. For example, Reserve [62] learns to understand vision-language knowledge from web videos and the corresponding transcripts. Flamingo [1] and FrozenBiLM [59] are established using frozen pretrained models through cross-modal training. In contrast, our approach can leverage readily available pretrained models without the need for costly cross-modal training.

2.2. LLMs for Vision and Language Tasks

Large Language Models (LLMs) have demonstrated impressive generalization capability on Natural Language Processing (NLP) tasks, thanks to their rich knowledge learned from vast text data [5, 10, 15, 37, 52]. Recently, LLMs have been applied to vision and language (ViL) tasks, such as image captioning [7] and visual question answering [1, 59]. Applying LLMs for ViL tasks is challenging due to the cross-modal gap. Importantly, LLMs are expensive to run, let alone fine-tuning them on large target datasets. For instance, the GPT-3 [5] model with 175B parameters requires 350GB of GPU memory to perform inference, not to mention training. Such high resource demands become practical obstacles for model training. This motivates the development of cheaper methods to bridge vision and language.

Training-based adaption Tsimpoukelli *et al.* [50] train a vision encoder that encodes visual information into text embedding for cross-modality alignment while freezing the LLMs. Flamingo [1] incorporates new cross-attention layers into existing frozen LLMs during training. FrozenBiLM [59] achieves state-of-the-art performance on zero-shot VideoQA by adapting frozen bidirectional Language Models [10]. MAPL [33] and VisualGPT [7] also leverage cross-modal fine-tuning of large pre-trained models, with focus on VQA and image caption tasks. Commonly, these methods all require joint training of vision and language models together, which is computationally expensive due to large scales of both the model [5, 52] and dataset size [2, 35, 46, 49, 63]. For example, Flamingo [1] was

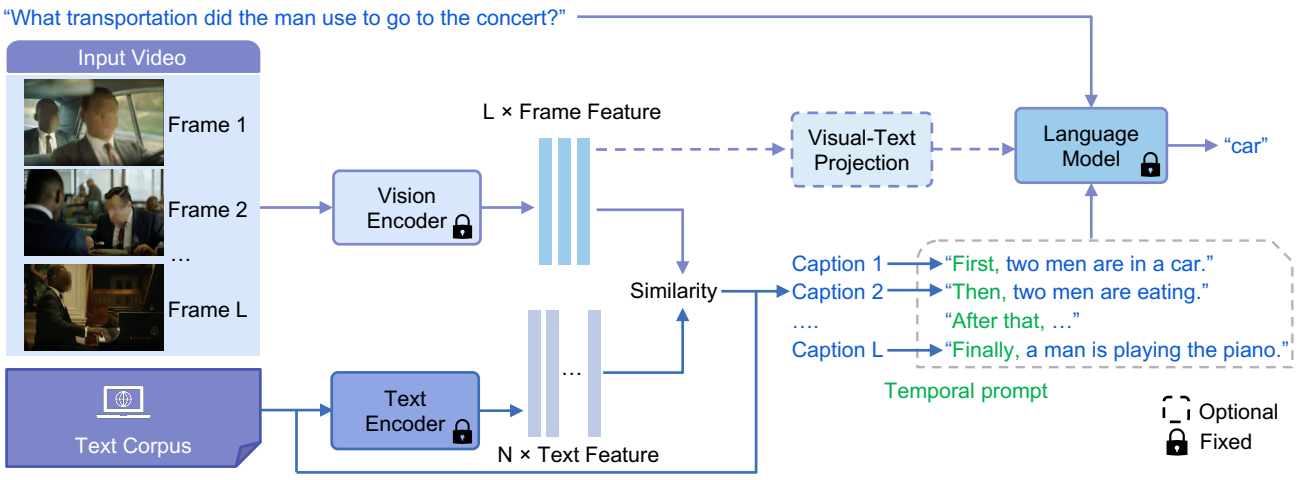

Figure 2. **Overview of our *Retrieving-to-Answer* (R2A) framework.** With the text encoder of a pretrained vision-language (ViL) model (*e.g.*, CLIP [40]), we first encode all the captions of an external text corpus (*e.g.*, WebVid [2]) into the aligned representation space *for one time*. Given a test video, **(a)** we first extract the frame features using the vision encoder of the CLIP model. **(b)** Subsequently, for each video frame, we retrieve top-k semantically similar texts from the text corpus against the corresponding frame feature vector. **(c)** We then, combined the top-k retrieved text for each frame with temporal prompts to construct a video-level textual context. **(d)** With both the question and the video-level context as input, a pretrained language model (*e.g.*, DeBERTa [15]) can be directly used to yield the final answer. **(e)** *Optionally*, a visual-text projection layer can be learned to prompt the same language model, taking the video feature as input.

trained for 500K TPU hours on billions of image/video-text pairs. Despite that approaches like FrozenBiLM [59] explored ways to mitigate this issue, the problem is still far from being solved (requiring up to hundreds of GPU hours and millions of video-text pairs for adaptation).

Language-based adaption Instead of training modality alignment modules, some recent approaches leverage language directly for connecting visual cues with LLMs [34, 51, 60, 64]. Specifically, they all use an image-caption model to generate textual descriptions of visual information. While there is no need for training-based cross-modal alignment, they additionally suffer from the limitations of the caption module used, such as limited domain knowledge and high inference cost. (which is in the same domain as the target VQA tasks) for caption generation, PnP-VQA [34] must go through multiple processing steps to generate captions for achieving good performance on visual question answering, resulting in high inference cost. In contrast, our approach retrieves text descriptions from external world knowledge using a generic multi-modal retriever [40], much more efficient than captioning whist no extra limitations. Note, our retrieval process can be implemented easily with highly optimized open-source libraries (*e.g.*, SCaNN [12] can query 2 trillion tokens in $10ms$).

2.3. Retrieval augmentation

Retrieval-augmented language models have gained attention in NLP [4, 14, 26, 27, 27, 53]. In vision-related applications, there are attempts at exploiting semantically related examples for improving model training or inference, *e.g.* visual question answering [19, 27], image captioning [41, 43, 65], recognition [17, 32], synthesis [18, 39], human parsing [31], and many more [4, 14, 26, 53]. Nagrani *et al.* [36] transfer image captions to video according to visual similarity for building video-language datasets. VideoCLIP [57] adopts retrieval-augmented training for mining hard negative samples. Our work belongs to this line of research. To the best of our knowledge, this is the first attempt that exploits the strategy of cross-modal retrieval from open-world knowledge for the VideoQA tasks.

3. Methodology

For zero-shot VideoQA, we formulate a **Retrieving-to-Answer** (R2A) framework by efficiently bridging the readily available pretrained vision and language models (*e.g.*, CLIP [40] and DeBERTa [15]). The overall architecture is depicted in Figure 2. Via retrieval-based cross-modal search in an external text corpus (*e.g.*, WebVid [2]), we transfer the information of video into text; As a result, an existing language model can be applied to VideoQA.

We first describe how to encode video (Section 3.2) and text (Section 3.1). Then, we explain in Section 3.3 how to retrieve semantically relevant textual context given a query video. We present in Section 3.5 how we answer the question with the retrieved context. Finally, we describe in Section 3.6 how to (optionally) further learn a visual-to-text layer for a further performance boost.

3.1. Text Corpus Encoding

Our R2A is charaterized with an external generic text corpus for contextual text retrieval. We use the text encoder $\texttt{E}_{\texttt{Text}}$ of the same ViL model (CLIP in this case) to text vectorization. Suppose there are N samples in the corpus, with each consisting of a word sequence T_i. For efficient retrieval, we precompute the textual features for all text samples: $v_i = \texttt{E}_{\texttt{Text}}(T_i) \in \mathbb{R}^d$, with $i = 1, ..., N$. This process takes place once. In the case of CLIP, similarly v corresponds to the feature of the `[CLS]` token.

3.2. Video Encoding

For video encoding, we use the vision encoder $\texttt{E}_{\texttt{Vis}}$ of a pretrained ViL model (*e.g.*, CLIP [40]). Suppose we have an input video $V = [f_1, f_2, ..., f_L]$, where L is the number of frames. We extract a visual feature embedding for each frame of the video using the encoder, denoted as: $z_t = \texttt{E}_{\texttt{Vis}}(f_t) \in \mathbb{R}^d$, with $t = 1, ..., L$.

In practice, we adopt CLIP's ViT-L/14 variant whilst other similar models such as ALIGN [21] can be similarly considerable. The frame feature embedding z corresponds to the output feature of the `[CLS]` token.

3.3. Video-Text Retrieval

We obtain the contextual text by video-text retrieval from the external text corpus. Specifically, we compute the similarity score at video frame level between frame feature z_t and text feature v_i:

$$d(z_t, v_i) = \frac{z_t \cdot v_i}{\|z_t\| \|v_i\|} \tag{1}$$

where $\cdot$ defines the dot product operation. This is the default choice in our main experiments.

Given all the pairwise similarity scores, we rank the corpus samples in the descending order. The contextual text for each frame is obtained as the top k matches, denoted as $r_t = [T_{t_1}, ..., T_{t_k}]$, where t_k stands the for top k-th text match for the t-th frame. The contextual text for the whole video is denoted as: $R = [r_1, ..., r_L]$.

3.4. Temporal Aware Prompting

In order to capture temporal transitions between frames, we further process the retrieved video contextual text. Specifically, similar to [51] we add temporal aware prompts in natural language indicating the temporal order of the retrieved captions, as a result, the prompted contextual text has the following form: "`Firstly,`$\{r_1\}$`... Then,`$\{r_t\}$`... After that,... Finally,`$\{r_L\}$".

3.5. Answer Generation

To generate an answer, we exploit a pretrained language model conditioned on both the question and the contextual text we retrieve as above. As a showcase, in practice we use a BERT [10] style language model pretrained with the Masked-Language-Modeling (MLM) task. Concretely, the objective of MLM is to predict the values of all masked tokens given a word sequence under random masking. For example, given an input sequence as "`Paris [MASK] the [MASK] city of France`", via contextualizing over visible word tokens, the model can predict the two masked tokens "`is`" and "`capital`". To fit this scheme, we construct a prompt template as: "`Question:{question} Answer:[MASK] Hints:{prompted contextual text}`". We use this template to prompt the pretrained language model which could output the answer at the designated `[MASK]` position. For the example in Fig. 2, the input text for the LM will be like this: "`Question: What transportation did the man use to go to the concert? Answer: [MASK]. Hints: First, two men are in a car.... Finally, a man is playing piano.`"

3.6. R2A with Learnable Visual-Text Projection

We have discussed our standard R2A framework as above. In cases that we have access to a video-text pair training data (*e.g.*, WebVid [2], not VideoQA specific training data since we consider the zero-shot setting here) and aim to pursue further performance boost, a lightweight learnable module can be simply integrated on top (see Figure 2). We denote the fine-tuned version of R2A as **R2A-FT**. However, we note that this will discard a certain degree of flexibility as this extra training is coupled with other pretrained vision and language models.

For training cost minimization, we only learn a *visual-to-text projection layer* to map the video features to prompts, whilst freezing the language model. The video-conditioned prompts are learned to be compatible with the language model. Formally, this component can be written as:

$$p_{vid} = \{z_t \mathbf{W}_{proj}\}_{t=1}^{L} \tag{2}$$

where $\mathbf{W}_{proj}$ denotes the learnable parameters with our visual-to-text projection layer. It is randomly initialized, and the only part to be optimized. Different from existing alternatives [59], we do not include any adapters in the frozen language model.

Concretely, the model takes as input the retrieved contextual text (captions), the video-conditioned prompts, and the original caption with the video. Model training is conducted by the MLM task:

$$\texttt{log}\, p(y|p_{vid}, R, T_{vid}) = \sum_{m=1}^{M} \texttt{log}\, p(y_m|p_{vid}, R, T_{vid}) \tag{3}$$

where T_{vid} is the original caption from the training set with tokens randomly masked out, y are the values of those masked tokens and M is the number of masked tokens.

We set the mask ratio to 50%. Following the original implementation of BERT [10], we replace the masked token position with the `[MASK]` token at the probability of 80%, with a random token at 10%, and with the original token at 10%. During inference, we follow the same manner as discussed in Section 3.5, except that we further prepend the newly generated video-conditioned prompts.

4. Experiments

We first describe the VideoQA datasets used for our evaluations (Section 4.1). We then give the implementation details of our model (Section 4.2). Next, we compare our R2A with the state of the art methods (Section 4.3). Finally, we ablate the effect of different components, *e.g.*, the choice of pretrained models, the number of retrieval samples and the construction of the text corpus (Section 4.4).

4.1. Datasets and Evaluation

In this work, we focus on the challenging open-ended VideoQA datasets where the model has to generate an open-ended answer for each video-question pair. We test multiple zero-shot VideoQA benchmarks with data collected from sources of great diversity (*i.e.*, YouTube videos, Sports Videos, GIFs), including **MSRVTT-QA** [56], **MSVD-QA** [56], **ActivityNet-QA** [61], **TGIF-QA** [20] and **iVQA** [58]. We report top-1 accuracy based on exact matching between the predicted answer and ground-truth annotation following previous evaluation protocols [22].

For the external corpus, we consider text extracted from various multi-modal datasets, including two datasets scraped from web: (1) **WebVid** [2] that contains approximately 10M video-text pairs, (2) **CC3M** [46] and **CC12M** [6] that consists of 3M and 12M image-text pairs, as well as a human-annotated dataset: (3) **COCO Caption** dataset [29] with 1.5M human-generated captions describing over 330K images.

4.2. Implementation details

For video-to-text retrieval (sec. 3.3), we use the ViT-L/14 variant of CLIP [40]. For the LLMs, we adopt DeBERTa-XL [15] as our default language model. It is worth mentioning that we utilize the MLM pre-trained model checkpoint, which means it has never been trained on QA-related tasks in any modality. Unless stated otherwise, we set 500 as the maximum input length for our language models. We base our implementation on the officially released code of [59] and the unmentioned details follow their implementation.

Visual feature extraction We extract frame features using the ViT-L/14 variant of CLIP [40]. The frame preprocessing follows the official implementation of [40]. We uniformly sample 10 frames from each video and extract one feature vector for each frame by taking the output of the `[CLS]` token.

Video-to-Text Retrieval The feature similarity calculation is identical to the original CLIP implementation. We use the naive algorithm for the nearest neighbor search (*i.e.*, calculating the similarity between all pairs and selecting the top-k for each query). More advanced nearest neighbor search algorithms (*e.g.* [13]) can be used for larger datasets. Duplicate entries are removed from the retrieved set.

Multi-modal Fine-Tuning For the multi-modal fine-tuning paradigm, we learn a linear projection for visual features as stated in Section 3.6. We use an Adam optimizer with a constant learning rate of 1.5e-5, *no* weight decay and $\beta_1, \beta_2 = 0.9, 0.95$. By default, we use a batch size of 64, limit the input sequence length to 64 for training, and train for only one epoch on each dataset. We alter each input token with a probability of 0.5.

4.3. Comparison with the State-of-the-arts

Quantitative comparison Table 1 presents the results of our method in comparison with current state-of-the-art approaches on zero-shot VideoQA. Our method outperforms approaches that were additionally trained on million to billion-scale vision-language data, except on iVQA, where our method underperforms Flamingo [1]. However, it is worth noting that Flamingo, while trained with billion-scale data, even the smallest version has significantly more parameters than our method. Furthermore, when using the same language model, our method consistently improves over VidIL [51], which utilizes a caption model to connect video with the language model. These results confirm the effectiveness of our method, as well as the informativeness of the retrieved captions for VideoQA.

Qualitative comparison Figure 3 illustrates qualitative results of zero-shot VideoQA for our Retrieving-to-Answer in comparison to FrozenBiLM [59] and the text-only baseline without access to visual information (w/o retrieval). First, we observe that for questions only baseline, the LM can predict answers based on commonsense reasoning, *e.g.* in the second example, it is very likely for a teacher to write problems on paper if we do not consider the visual input. Second, for FrozenBiLM, predictions can be misled by inaccurate visual information (*e.g.* in the first example, it predicts "paint" instead of "egg"). In contrast, our R2M can predict the correct answer based on high-quality informative context retrieved from the supportive text corpus.

Efficiency analysis As shown in Table 2, the inference time per video for our R2A is 0.11s, which can be further broken down into video encoding latency, retrieval latency, and LM inference latency. For FrozenBiLM, the inference time is 0.07s, which is composed of video encoding latency and LM inference latency. Given the exact

Method	Language model	Language #params	Vision model	Vision #params	Benchmarks MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA	iVQA
Training based Adaption									
CLIP ViT-L/14 [40]	Custom	123M	ViT-L/14	300M	2.1	7.2	1.2	3.6	9.2
Just Ask [58]	DistilBERT	66M	S3D	12M	5.6	13.5	12.3	-	13.3
Reserve [62]	Custom	-	ViT-L/16	300M	5.8	-	-	-	-
Flamingo-3B [1]	Chinchilla-like	2.6B	NFNet-F6	629M	11.0	27.5	-	-	32.7
Flamingo-9B [1]	Chinchilla-like	8.7B	NFNet-F6	629M	13.7	30.2	-	-	35.2
Flamingo-80B [1]	Chinchilla-like	80B	NFNet-F6	629M	17.4	35.6	-	-	**40.7**
FrozenBiLM [59]	DeBERTa-v2-XL	890M	ViT-L/14	300M	16.9	33.7	25.9	41.9	26.2
Language based Adaption									
VidIL* [51]	DeBERTa-v2-XL	890M	ViT-L/14	300M	16.6	31.7	-	-	-
R2A (Ours)	DeBERTa-v2-XL	890M	ViT-L/14	300M	**18.3**	**37.0**	**26.3**	**52.2**	29.3

Table 1. **Comparison with state-of-the-arts on zero-shot videoQA.** 50 retrieved sentences are used and the prompt is "`Hints:`". *indicates replacement of LM beyond original definitions by authors for fair comparisons and also due to a lack of access to the original Language Model (GPT-3).

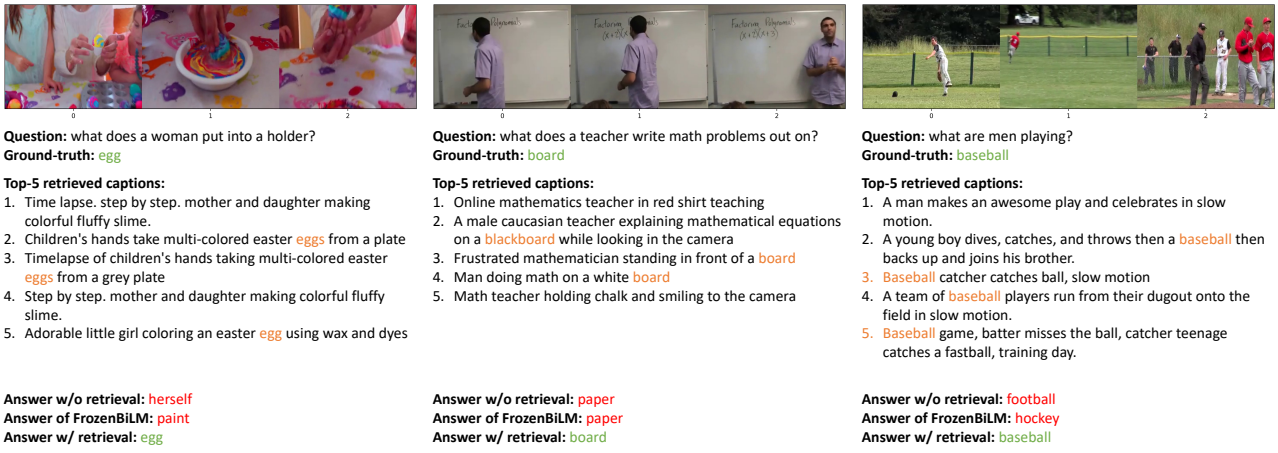

Figure 3. **Qualitative comparison of model predictions with and without retrieved captions.** For illustrative purposes, we highlight words in orange to indicate answer cues from captions. Words in green indicate correct answer predictions and in red for incorrect ones.

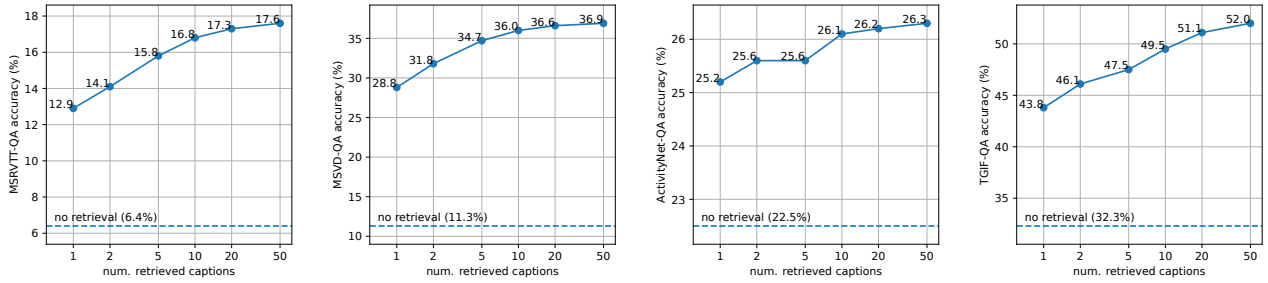

Figure 4. **Effects of number of retrieved captions on zero-shot VideoQA.** WebVid-10M is used as the retrieval set and "`Subtitles:`" is used as the prompt in all experiments.

same video input setup, the video encoding time is identical for our method and FrozenBiLM. Our LM inference latency is slightly larger than FrozenBiLM's because our input sequence (including the retrieved contextual text) is longer. It is worth noting that, the time consumed by retrieval is only 3.5ms (2,800 queries per second on a single A100 GPU from 10M samples) which is almost negligible compared to the total inference time. In addition to the inference time, we also report the pre-inference preparation time which is cross-modal training time for Flamingo and FrozenBiLM. In the case of R2A, it is the time needed to extract the features of the retrieval set which is fractional compared to the time required for cross-modal training. Overall, our results demonstrate that R2A achieves competitive

Method	Pre-inference computation	Inference time per video
Flamingo-80B [1]	500Kh (TPUv4)	-
FrozenBiLM [59]	160h (V100)	0.07s
R2A (Ours)	1.4h (V100)	0.11s

Table 2. **Efficiency comparison.** We compare efficiency with two previous methods. *Pre-inference* includes the cross-modal training cost for training based methods and retrieval set feature extraction cost for ours. We report efficiency using 10 retrieved sentences, at which *R2A* outperforms FrozenBiLM on all tasks.

Ret. modality	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
Video-Video	15.7	32.6	25.0	45.2
Video-Text	**18.3**	**37.0**	**26.3**	**52.2**

Table 3. **Video-Text vs. Video-Video retrieval.** 50 retrieved sentences are used and the prompt is "`Hints:`".

ret.?	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
Language model: BERT-Base				
✗	1.4	2.9	16.0	10.4
✓	3.0 (+1.6)	7.5 (+4.6)	15.4 (-0.6)	17.4 (+7.0)
Language model: BERT-Large				
✗	2.5	3.9	15.6	13.9
✓	7.0 (+4.5)	15.0 (+11.1)	19.6 (+4.0)	26.7 (+12.8)
Language model: RoBERTa-Large				
✗	3.0	5.5	18.5	13.2
✓	13.6 (+10.6)	24.3 (+18.8)	18.8 (+0.3)	32.0 (+18.8)
Language model: DeBERTa-v2-xlarge (default)				
✗	6.4	11.3	22.5	32.3
✓	16.8 (+10.4)	36.0 (+24.7)	26.1 (+3.6)	49.5 (+17.2)

Table 4. **Performance of the training-free setting with alternative language models.** WebVid-10M is used as the retrieval set and 10 sentences are retrieved for each sample. The prompt is "`Subtitles:`". "ret.?" denotes whether to use retrieval or not.

performance while maintaining efficient inference time and requiring minimal pre-inference preparation.

4.4. Ablation Studies

Retrieval dataset construction Table 5 presents the results of our method equipped with different text corpora for context retrieval. To investigate the impact of dataset size, we conduct experiments with various sample sizes. As shown in the second section of Table 5, the model's performance consistently improves with the increase in dataset size. Even a small dataset of 10k samples can bring a noticeable improvement over models without retrieved captions. We observe that the performance gain tends to be smaller when using CC-3M and CC-12M as the retrieval set, which we attribute to the dataset quality issues mentioned in [33]. Specifically, we find some texts in the

Retrieval database	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
No retrieval dataset (baseline)				
N/A	6.4	11.3	22.5	32.3
Down-sampling the retrieval dataset				
WV-10k	15.6	30.3	25.0	44.2
WV-100k	16.2	34.1	25.1	47.8
WV-1M	16.3	35.6	25.8	48.5
WV-10M	16.8	**36.0**	**26.1**	49.5
Using the Conceptual Captions (CC) datasets				
CC-3M	14.2	30.2	25.0	47.1
WV-10M + CC-3M	15.9	34.7	26.0	49.5
CC-12M	10.9	23.6	23.4	43.2
WV-10M + CC-12M	13.8	31.6	25.6	46.4
Using human-annotated captioning datasets				
COCO	16.4	31.2	23.9	45.3
COCO + WV-10M	**16.9**	**36.0**	26.0	**49.7**

Table 5. **Effects of retrieval database construction.** WV stands for the WebVid dataset [2]. We use 10 retrieved sentences and the prompt "`Subtitles:`" in all experiments. Note that we only use the textual part of each of the dataset.

Retrieval database	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
N/A	6.4	11.3	22.5	32.3
CC-12M (Random)	7.7	10.0	19.6	30.1
CC-12M (Ours)	10.9	23.6	23.4	43.2
WV-10M (Random)	9.5	18.5	22.3	31.4
WV-10M (Ours)	16.8	36.0	26.1	49.5

Table 6. **Retrieval vs. random sample.** `Subtitles:` is used as prompt and 10 sentences are used in all experiments.

CC datasets unreadable, which can negatively affect the model's performance. Our experiments on COCO Captions reaffirm this hypothesis, as we observe higher performance gains for high-quality captions annotated by humans, even with much smaller dataset sizes. Unless otherwise stated, we use the full WebVid-10M as the default retrieval dataset.

Impact of the number of retrieved captions We plot the top-1 accuracy against the number of retrieved captions, as shown in Figure 4. The accuracy on all four datasets consistently increases as more retrieved captions are used, demonstrating the robustness of our method. Notably, we observe a substantial performance boost even when using only one retrieved caption, indicating the importance of the retrieval operation in assisting VideoQA. Furthermore, we find that the accuracy continues to improve with an increasing number of retrieved captions. These findings suggest that our method can effectively leverage external sources of information for VideoQA task.

Impact of the quality of retrieved captions To demonstrate the effectiveness of the retrieval in providing relevant information for VideoQA, we compare our results with those obtained by randomly sampling text to feed into the LM. As shown in Table 6, our R2A consistently outperform random sampling on all datasets and retrieval sources by a

Prompt	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
`Subtitles:`	17.3	36.6	26.2	51.1
`Captions:`	17.2	36.8	26.2	50.9
`Hints:`	**18.0**	36.8	26.3	**51.2**
`Contexts:`	17.6	**36.9**	**26.4**	51.1

Table 7. **Effects of using different prompts.** We use WebVid-10M as the retrieval set and 20 retrieved sentences in all experiments.

large margin. Surprisingly, we observe that randomly sampled text show improved performance over the question-only baseline on MSRVTT and MSVD. We hypothesize that this is due to the prior distributions of some retrieval datasets, which may serve as task-specific prompts [66] for certain target datasets.

Video-Text vs. Video-Video Retrieval Apart from using only text as the external corpus, we experiment with retrieving video-text pairs: For a video-text dataset, we find the samples with the highest video-video similarity and take the corresponding text as the retrieved captions. As shown in Table 3, the video-video retrieval is significantly worse on all four datasets. We conjecture that video-video similarity is more vulnerable to data noise (*i.e.*, some video-text pairs themselves may not be well aligned, especially for those web-crawled data), which consequently makes the retrieved text not well correlated with the query video. Moreover, retrieving directly from the text corpus is also advantageous in term of storage (*i.e.*, no need to store the images or videos) and flexibility (*i.e.*, able to use both visual-text and text-only datasets).

Impact of pretrained language models We also experiment with alternative pretrained language models and report our results in Table 4. Except for one case (BERT-base on ActivityNet-QA), we observe that the retrieved contexts significantly improve performance, sometimes even doubling the accuracy on all LMs. Notably, our models benefit more from stronger language models, as we observe larger gains with the increase in language model size.

Effects of prompts To investigate the impact of prompt design on our method, we conduct experiments using a few hand-crafted prompts. We replace the words before the retrieved captions to probe the language model. The results are presented in Table 7. We find that there is no significant variation among the choices attempted, but some words such as `Hints:`" and `Contexts:`" perform slightly better on all datasets than the others, such as `Subtitles:`" and `Captions:`". Further optimization of the prompting words may potentially improve the performance.

Learning Visual-Text Projection for R2A In Table 8, we analyze the impact of multi-modal training on R2A. For efficiency purposes, we train R2A on various subsets of WebVid [2] for one epoch. Our findings reveal that while fine-tuning with relatively larger cross-modal data (*e.g.*, 500K), there are some improvements on MSRVTT and TGIF. However, fine-tuning with smaller cross-modal data (*e.g.*, 10k or 50k) can hamper the model's performance in the original setup. The reason behind this is that an improperly trained visual feature projection may introduce noisy tokens into the self-attention layers, leading to a negative impact on the language model and ultimately, on the performance.

Size	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
0	18.3	**37.0**	**26.3**	52.2
10k	12.8	26.8	22.8	32.7
50k	17.5	34.3	25.6	42.8
200k	16.7	33.7	26.0	43.1
500k	**19.7**	36.8	25.8	**52.5**

Table 8. **Learning Visual-Text Projection for R2A.** In each experiment, we downsample the WebVid-10M dataset to the given size, and fix the number of training *epochs* to one.

5. Conclusions

We propose Retrieving-to-Answer, a framework for zero-shot VideoQA without task-specific training, utilizing off-the-shelf pre-trained multi-modal contrastive model. It transfers the zero-shot ability of a pre-trained LLM to a multi-modal setting without the need for explicitly learning video-language alignment. Specifically, we summarize the video modality with text via fast cross-modal retrieval in an external text corpus. Then, we probe a pre-trained language model with both the retrieved text and the question to predict the answer. Our design is highly flexible allowing easy component updates with no extra training. Experiments show that our R2A can achieve new state-of-the-art performance on multiple benchmarks.

6. Limitations

The proposed Retrieving-to-Answer (R2A) approach is a promising direction for achieving zero-shot video question answering (VideoQA). Our attempt constitutes an important proof of concept in exploiting multimodal retrieval to enhance the generalization and robustness of current VideoQA frameworks. However, a major limitation of R2A is the extent to which the quality of the retrieved captions depends heavily on the performance of the retrieval model, as well as the diversity of the text corpus. Despite the remarkable abilities of CLIP in open-domain zero-shot cross-modal retrieval, it may still struggle to handle certain types of videos or text. Ideally, a sufficiently large dataset should encompass all topics and content of interest. In reality, however, there are still many cases where we cannot find the desired answers among existing data. Nevertheless, we posit that R2A constitutes a promising starting point and a baseline for future research on retrieval-based VideoQA.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *ArXiv preprint*, abs/2204.14198, 2022. 1, 2, 5, 6, 7

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 1, 2, 3, 4, 5, 7, 8

[3] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*, 2021. 3

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 2

[6] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 5

[7] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022. 2

[8] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1166–1174, 2021. 2

[9] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. Object-centric representation learning for video question answering. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. 1, 2, 4, 5

[11] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 2

[12] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, pages 3887–3896. PMLR, 2020. 3

[13] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, pages 3887–3896. PMLR, 2020. 5

[14] K Guu, K Lee, Z Tung, P Pasupat, and MW Chang. Realm: Retrieval-augmented language model pre-training. arxiv 2020. *arXiv preprint arXiv:2002.08909*. 3

[15] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020. 1, 2, 3, 5

[16] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *arXiv preprint arXiv:2302.11154*, 2023. 2

[17] Ahmet Iscen, Thomas Bird, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. A memory transformer network for incremental learning. *arXiv preprint arXiv:2210.04485*, 2022. 3

[18] Phillip Isola and Ce Liu. Scene collaging: Analysis and synthesis of natural images with semantic layers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3048–3055, 2013. 1, 3

[19] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020. 3

[20] 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. 5

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 4

[22] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu

Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv e-prints*, pages arXiv–2203, 2022. 2, 5

[23] Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13171–13179, 2021. 2

[24] 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

[25] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 1, 2

[26] Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems*, 33:18470–18481, 2020. 3

[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 3

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ArXiv preprint*, abs/2201.12086, 2022. 1

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[30] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7001–7011. IEEE, 2021. 2

[31] Si Liu, Xiaodan Liang, Luoqi Liu, Xiaohui Shen, Jianchao Yang, Changsheng Xu, Liang Lin, Xiaochun Cao, and Shuicheng Yan. Matching-cnn meets knn: Quasi-parametric human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1419–1427, 2015. 3

[32] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6959–6969, 2022. 3

[33] Oscar Mañas, Pau Rodriguez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. Mapl: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. *arXiv preprint arXiv:2210.07179*, 2022. 2, 7

[34] Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. Plug-and-play vqa: Zero-shot vqa by conjoining large pretrained models with zero training. *arXiv e-prints*, pages arXiv–2210, 2022. 1, 3

[35] 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 *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE, 2019. 1, 2

[36] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. *arXiv preprint arXiv:2204.00679*, 2022. 3

[37] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155, 2022. 1, 2

[38] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535, 2021. 2

[39] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018. 3

[40] 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. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 4, 5, 6

[41] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: Lightweight image captioning prompted with retrieval augmentation. *arXiv preprint arXiv:2209.15323*, 2022. 3

[42] Arka Sadhu, Kan Chen, and Ram Nevatia. Video question answering with phrases via semantic roles. *arXiv preprint arXiv:2104.03762*, 2021. 2

[43] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. In *International Conference on Content-based Multimedia Indexing*, pages 1–7, 2022. 3

[44] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 2

[45] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 2

[46] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1, 2, 5

[47] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 1

[48] 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. 1

[49] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 1, 2

[50] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 1, 2

[51] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *arXiv preprint arXiv:2205.10747*, 2022. 1, 3, 4, 5, 6

[52] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *ArXiv preprint*, abs/2109.01652, 2021. 1, 2

[53] Yuhuai Wu, Markus N Rabe, DeLesley Hutchins, and Christian Szegedy. Memorizing transformers. *arXiv preprint arXiv:2203.08913*, 2022. 3

[54] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021. 2

[55] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653, 2017. 1

[56] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 5

[57] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[58] 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, 5, 6

[59] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022. 1, 2, 3, 4, 5, 6, 7

[60] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089, 2022. 3

[61] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 5

[62] 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. *ArXiv preprint*, abs/2201.02639, 2022. 2, 6

[63] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2

[64] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598*, 2022. 3

[65] Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9837–9846, 2021. 3

[66] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 8