Just Ask: Learning to Answer Questions from Millions of Narrated Videos

Antoine Yang[1,2], Antoine Miech[1,2,+], Josef Sivic[3], Ivan Laptev[1,2], Cordelia Schmid[1,2]

[1]Inria Paris [2]Département d'informatique de l'ENS, CNRS, PSL Research University [3]CIIRC CTU Prague [+]Now at DeepMind

https://antoyang.github.io/just-ask.html

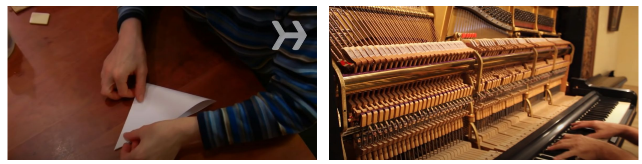
Abstract

Recent methods for visual question answering rely on large-scale annotated datasets. Manual annotation of questions and answers for videos, however, is tedious, expensive and prevents scalability. In this work, we propose to avoid manual annotation and generate a large-scale training dataset for video question answering making use of automatic cross-modal supervision. We leverage a question generation transformer trained on text data and use it to generate question-answer pairs from transcribed video narrations. Given narrated videos, we then automatically generate the HowToVQA69M dataset with 69M video-question-answer triplets. To handle the open vocabulary of diverse answers in this dataset, we propose a training procedure based on a contrastive loss between a video-question multi-modal transformer and an answer transformer. We introduce the zero-shot VideoQA task and show excellent results, in particular for rare answers. Furthermore, we demonstrate our method to significantly outperform the state of the art on MSRVTT-QA, MSVD-QA, ActivityNet-QA and How2QA. Finally, for a detailed evaluation we introduce iVQA, a new VideoQA dataset with reduced language biases and high-quality redundant manual annotations.

1. Introduction

Answering questions about videos requires a detailed understanding of the visual content and its association with the natural language. Indeed, given the large diversity of questions, methods for Video Question Answering (VideoQA) should reason about scenes, objects and human actions as well as their complex temporal interactions.

Current approaches to VideoQA rely on deep fully-supervised models trained on manually annotated datasets with question and answer pairs [22, 31, 34, 35, 40, 42, 48]. Collecting and annotating VideoQA datasets, however, is cumbersome, time consuming, expensive and therefore not scalable. As a result, current VideoQA datasets are relatively small (see Figure 2). This limitation hinders the progress in the field as state-of-the-art VideoQA models often require a large amount of training data.



Speech: Fold them in half again, to make a triangle.

Generated Question: How do you make a triangle?
Generated Answer: Fold them in half again

Speech: The sound is amazing on this piano.

Generated Question: What kind of instrument is the sound of?
Generated Answer: Piano

Figure 1: Given videos with transcribed narration, we leverage language models and cross-modal supervision to obtain large-scale VideoQA data. Above are two examples from our dataset.

In this work, we address the scale issue with a new approach for automatically generating a VideoQA dataset, see Figure 1 for examples. The idea is to leverage cross-modal supervision together with text-only tools for question generation and to automatically annotate VideoQA from a *large amount of readily-available narrated videos*. Inspired by the recent progress in language generation using transformer-based language models [10], we leverage transformers trained on a question-answering text corpus to generate a diverse set of non-scripted questions and corresponding open-vocabulary answers from text. By applying these transformers to speech transcripts of narrated videos from the large-scale HowTo100M dataset [58], we create HowToVQA69M, an open-ended VideoQA dataset with 69 million video-question-answer triplets and a diverse set of more than 16M unique answers (see Figure 3). As shown in Figure 2, our HowToVQA69M is two orders of magnitude larger compared to prior VideoQA datasets.

Given the limited diversity of existing datasets, current methods typically reduce video question answering to a classification problem, where frequent answers are assigned to unique classes. Typically, up to 5K unique possible answers are considered. Such an approach, however, does not scale to the open vocabulary of 16M different answers in our dataset. To address this problem and to enable video question answering with highly diverse questions

[3]Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University in Prague.

and answers, we introduce a training procedure based on contrastive learning between a video-question multi-modal transformer and an answer transformer that can handle free-form answers. This bypasses the need to define a discrete set of answer classes.

The goal of our work is to advance truly open-ended and generic solutions to VideoQA. To evaluate generalization, we propose a new zero-shot VideoQA task where we prohibit any manual supervision of visual data during training. Our VideoQA model, trained on HowToVQA69M, demonstrates excellent zero-shot results on multiple existing datasets, especially for rare answers. Moreover, when finetuned on target datasets, our model significantly outperforms the state of the art on MSRVTT-QA [84], MSVD-QA [84] ActivityNet-QA [91], and How2QA [46].

Initial experiments showed that existing benchmarks for open-ended VideoQA [84, 91] contain a language bias [28], i.e., their questions can often be answered without looking at the video. To better evaluate the impact of visual information in VideoQA, we introduce a new open-ended VideoQA dataset (iVQA) with manually collected questions and answers, where we exclude questions that could be answered without watching the video. Moreover, to account for multiple possible answers, iVQA contains five independently collected answers for each question.

In summary, our work proposes the following three contributions:

(i) We introduce an approach to automatically generate a large-scale VideoQA dataset, HowToVQA69M. Relying on cross-modal supervision, we use transformers trained on an existing text-only question-answering corpus and generate video-question-answer triplets from videos and transcribed narrations.

(ii) We train a VideoQA model on HowToVQA69M with contrastive learning between a multi-modal video-question transformer and an answer transformer. We show the efficiency of our model in the new zero-shot VideoQA task and outperform the state of the art in four existing VideoQA benchmarks: MSRVTT-QA, MSVD-QA, ActivityNet-QA and How2QA.

(iii) Finally, we introduce a new manually annotated open-ended VideoQA benchmark iVQA that excludes non-visual questions and contains multiple possible answers for each question.

Code, datasets and trained models are available at [1].

2. Related Work

Visual Question Answering (VQA). VQA is typically tackled by classifying the image-question (or video-question) representation into a fixed vocabulary of answers. Various approaches to combine spatial image representations and sequential question representations have been proposed [7, 9, 24, 55, 83, 85, 88]. More specifically to the video domain (VideoQA), spatio-temporal video representations in terms of motion and appearance have been used in [22, 26, 31, 33, 34, 35, 40, 41, 42, 48, 84, 86, 94, 102].

Methods above are limited to pre-defined vocabularies of answers and are difficult to apply outside of specific datasets. To address this problem, Hu *et al.* [30] propose a joint embedding where image-question representations can be matched with free-form answers. Our VideoQA model follows this idea, but instead of relying on manually annotated datasets of limited scale, we train it on a large-scale VideoQA dataset that we automatically generate. In contrast to some previous works using additional video features such as subtitles [12, 36, 37, 43, 44, 46, 74, 80, 87], our video representation is exclusively based on visual information, as we focus on the visual understanding of videos.

To evaluate the generalization of VQA models, Teney and Hengel [75] define zero-shot VQA by answering previously unseen questions, which is a related but less challenging task compared to the zero-shot VQA task we propose in Section 6.2. Vatashsky and Ullman [78] address VQA using COCO image annotations [51], while our zero-shot model is trained with no manual annotations. Our proposed zero-shot VQA task is analogous to zero-shot video retrieval [57] or zero-shot action recognition [61].

Visual question generation (VQG) has been introduced in [59]. The methods in [50] and [67] propose to jointly learn VQG and VQA to improve the image VQA task. However, these works do not generate questions to obtain additional training data, but use visual data annotation for question generation as an additional loss.

VideoQA datasets. Manually collecting and annotating video-question-answer triplets is cumbersome, costly and difficult to scale. As result, current VideoQA datasets [11, 16, 17, 21, 27, 33, 38, 43, 46, 60, 69, 74, 84, 90, 91, 92, 93] are limited in size, as the largest, TGIF-QA [33], contains only 72K annotated clips (see Figure 2 for more details). To address this issue, several works have explored leveraging manually annotated video descriptions [33, 79, 84, 93, 95, 96, 97] for automatic generation of VideoQA datasets, using rule-based [29, 64] approaches.

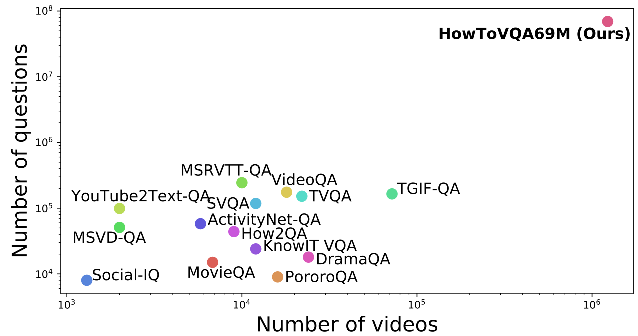


Figure 2: Comparison of our proposed large-scale HowToVQA69M dataset with existing VideoQA datasets.

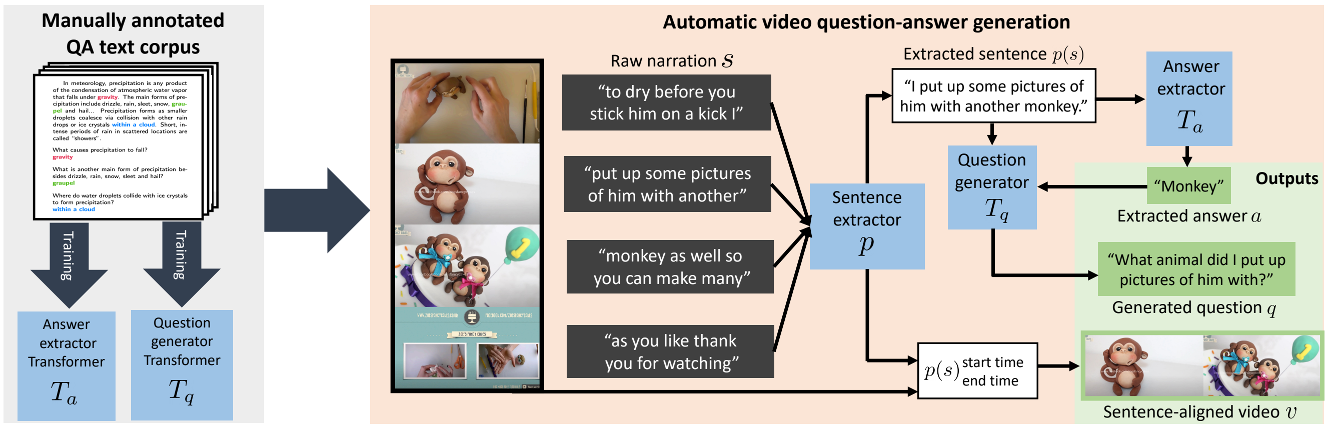


Figure 3: **Our automatic approach for large-scale generation of video-question-answer triplets from narrated (subtitled) videos.** First, at the language-only training phase (left), the transformer-based answer extractor T_a and question generator T_q are trained [62] on a manually annotated text-only question-answer corpus. Then video-question-answer triplets are automatically generated from narrated videos (right). Individual sentences are extracted from the ASR-transcribed narration using a punctuator p. Each extracted sentence is analyzed with an answer extractor T_a and a question generator T_q to produce answer a and question q. The timestamps of the narration are used to obtain a video clip v temporarily aligned to the extracted sentence to form the output video-question-answer triplet (v, q, a).

Instead, we propose to use video narrations that are available at large-scale with no manual supervision. Moreover, rule-based generation requires the manual creation of rules by experts which is expensive, and has also been recently outperformed by neural question generation [20, 89, 99] as used in our approach.

Large-scale pretraining for vision and language. Several recent methods [5, 15, 18, 32, 45, 47, 49, 53, 54, 70, 73, 98] pretrain multi-modal vision-language representations, such as transformers, using datasets with image captions, e.g., COCO [14], Conceptual Captions [68] and Visual Genome [39]. These methods are often optimized using generic objectives such as masked language losses and losses for text-image matching and image caption generation. In our work, we pretrain models using large amounts of narrated videos. In contrast to task-agnostic pretraining in the previous work, we show the benefits of task-specific pretraining for our target VideoQA task.

Learning from narrated videos. In this work, we exploit noisy correlations between videos and narrations in unlabeled instructional videos from the recent HowTo100M dataset [58]. Methods using such readily-available data have shown significant improvements on several tasks including video retrieval, action localization, action recognition and video captioning [25, 56, 57, 58, 71, 72, 100], sometimes outperforming fully-supervised baselines. Some recent works use narrated videos for VideoQA. Amrani *et al.* [6] propose a text-video pretraining approach and finetune for VideoQA. Li *et al.* [46] propose HERO, a pretraining approach restricted to multiple-choice VideoQA, for which question and answer are treated as a single text stream. Seo *et al.* [66] propose a pretraining approach based on next utterance prediction and finetune for VideoQA. Differently to these methods with task-agnostic pretraining, we propose a pretraining approach specifically dedicated for VideoQA using automatically generated question and answer pairs from narrated videos, and show in Section 6 the superiority of our approach.

3. Large-scale generation of VideoQA data

This section presents our approach to generate a large-scale VideoQA dataset from videos and transcribed narrations describing the content of the videos. Section 3.1 presents our proposed generation procedure. Section 3.2, then, describes the resulting HowToVQA69M dataset.

3.1. Generating video-question-answer triplets

We tackle the task of generating video-question-answer triplets from a large-scale instructional video dataset with transcribed spoken narration [58]. This is a challenging task because of transcription errors and lack of punctuation. We also wish to obtain highly diverse data. To address these issues, we propose to leverage powerful language models trained on text data. Our approach is illustrated in Figure 3 and details are given next.

We first present details about the generation procedure. Let s be the transcribed speech data obtained with automatic speech recognition (ASR). First, we use a recurrent neural network p, to infer punctuation in the transcribed speech data. We denote the punctuated transcript as $p(s)$. We extract video clips v temporally aligned with the inferred sentences $p(s)$ using the ASR timestamps. We found that the generation works significantly better when applied to sentences rather than the original sentence fragments from the HowTo100M dataset, see Table 1. Second, for each sentence, we apply a transformer T_a, to extract a set of potential answers: $a = T_a(p(s))$. Third, we use another transformer T_q to generate a question given each transcript sentence and

ASR: Ideally, you would want a medium spread collar... **Question:** What type of collar would you want? **Answer:** medium spread collar

ASR: Add some of your favorite sprinkles give it a mix. **Question:** What can you add to the mix? **Answer:** sprinkles

ASR: ...I'm going to show you how to unlock your ipod touch. **Question:** What will I show you? **Answer:** how to unlock your ipod touch

ASR: ...do it on the other side, and you've peeled your orange **Question:** What color did you peel on the other side? **Answer:** orange

ASR: ...I've had over a hundred emails. **Question:** How many emails have I had? **Answer:** over a hundred

Figure 4: Examples of video-questions-answer triplets generated from narrated videos in our HowToVQA69M dataset. The green color indicates relevant examples, the orange color (penultimate example) indicates a failure of the question-answer generation, and the red color (last example) indicates that the generated question-answer is unrelated to the visual content.

each extracted answer such that: $q = T_q(a, p(s))$. The output is a set of video-question-answer triplets (v, q, a).

We now explain details about the language models and their training procedures. For ASR, we follow [58] and use the readily-available ASR data provided by YouTube. For punctuation p, we use the BRNN model from [76] and the weights available at [2] trained on IWSLT2011 [23]. For T_a and T_q, we use the transformer-based T5-small and T5-base models [62], respectively. We follow [4, 13, 52] and use the weights available at [3] trained for answer span extraction and answer-aware question generation, respectively, on SQuADv1 [63]. SQuADv1 is a text-only question-answering dataset consisting of questions for which the answer is a segment of text extracted from a paragraph.

3.2. HowToVQA69M: large-scale VideoQA dataset

We have applied the previously described procedure to all 1.2M original videos from the HowTo100M dataset [58]. The result is HowToVQA69M, a dataset of 69,270,581 video clip, question and answer triplets (v, q, a). HowToVQA69M is two orders of magnitude larger than any of the currently available VideoQA datasets (see Figure 2). On average, each original video results in 43 video clips, where each clip lasts 12.1 seconds and is associated to 1.2 question-answer pairs. Questions and answers contain 8.7 and 2.4 words on average respectively. HowToVQA69M is highly diverse and contains over 16M unique answers, where over 2M unique answers appear more than once and over 300K unique answers appear more than ten times. Examples of (v, q, a) triplets from the HowToVQA69M dataset are illustrated in Figure 4.

Manual evaluation of HowToVQA69M. As shown in Figure 4, HowToVQA69M annotations are noisy, which can be attributed to: (i) errors in speech transcription, (ii) speech not describing the video content, or (iii) errors in question-answer generation. We manually evaluate the quality of 100 randomly sampled (v, q, a) triplets in HowToVQA69M by collecting 5 different annotations for each triplet to reduce variance, and report results in Table 1. Among 100 triplets generated by our method we find 30 to be correctly generated and matching well to the video content, 31 are incorrectly generated and 39 are correctly generated but unrelated to the video content. To demonstrate the influence of different components of our automatic question-answer generation procedure, we compare it with (i) a variant of our approach that does not split transcribed narrations into sentences using a punctuator, and (ii) a rule-based approach [29] for question-answer generation. Table 1 confirms the importance of punctuation and demonstrates the superior performance of our generation method compared to [29]. Inter-rater agreement statistics, and more details for the generated dataset are provided in the Supplementary Material. Further comparison with [29] is given in Section 6.5. We describe next how we use HowToVQA69M to train our VideoQA model.

Punctuation	Generation method	Correct Samples	QA Generation Failure	QA unrelated to video
✓	Heilman *et al.* [29]	17	54	29
✗	Ours	23	49	28
✓	Ours	**30**	31	39

Table 1: Manual evaluation of our generation method (with and without punctuation) on a random sample of 100 examples compared with a rule-based question-answer generation of [29]. Numbers are obtained with majority voting between 5 annotators.

4. VideoQA model and training procedure

This section presents our VideoQA model in Section 4.1 and describes its training procedure in Section 4.2. Figure 5 gives an overview of the model.

4.1. VideoQA model

As illustrated in Figure 5, our VideoQA model is composed of two branches: *(i)* a video-question module f based on a transformer [77] and a mapping from the CLS token with a linear function. It takes a pair of video v and question q as input, models the multi-modal temporal interactions between v and q and then outputs an embedding vector $f(v, q) \in \mathbb{R}^d$. *(ii)* The second branch is a text encoder g that embeds an answer a as $g(a) \in \mathbb{R}^d$. We will denote our model as *VQA-T*, standing for VideoQA-Transformer. Note that using the joint (video, question) and answer embeddings allows us to deal with a large open vocabulary of answers present in our new HowToVQA69M dataset as the model can measure similarity between the in-

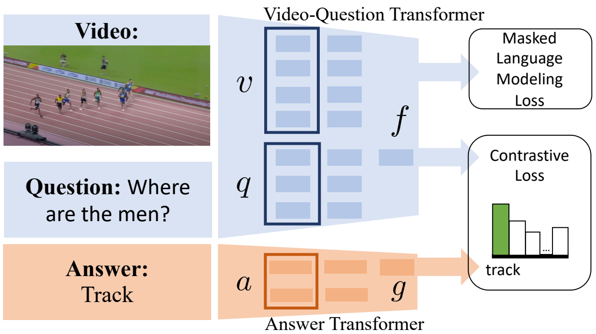


Figure 5: Overview of our VideoQA training architecture.

put video-question embedding and the embedding of any answer. This is in contrast to using a classification answer module [31, 34, 35, 40, 102] that can choose only from a fixed predefined vocabulary of answers. Our embedding can be also easily finetuned on the different downstream VideoQA datasets, which may contain new answers that have not been seen at training. In contrast, the classification answer module has to be retrained when the vocabulary of answers changes. Next, we give details of the language and video representations. Further details about the model are provided in the Supplementary Material.

Word representation. The question and answer are separately tokenized with the WordPieces embedding [81] and fed to DistilBERT [65]. DistilBERT is a light version of BERT [19] pretrained in a self-supervised fashion on English Wikipedia and the Toronto Book Corpus [101].

Video representation. We use a frozen S3D [82] pretrained on HowTo100M [58] using MIL-NCE [57]. This model is pretrained from scratch on HowTo100M only.

4.2. Training procedure

This section describes the training of our VideoQA model on the HowToVQA69M dataset and its finetuning on downstream VideoQA datasets.

Training on HowToVQA69M. We wish to make a pair of video and question (v, q) close to its correct answer a measured by the dot product of their embeddings, $f(v, q)^\top g(a)$. Conversely, the incorrect answers should be far, i.e., the dot product with their embeddings should be small. Formally, this can be done by maximizing the following contrastive objective:

$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{e^{f(v_i,q_i)^\top g(a_i)}}{e^{f(v_i,q_i)^\top g(a_i)} + \sum\limits_{(v',q',a')\sim\mathcal{N}_i} e^{f(v',q')^\top g(a')}} \right), \tag{1}$$

where (v_i, q_i, a_i) represents a triplet of generated (video clip, question, answer) from HowToVQA69M. Given a specific positive triplet (v_i, q_i, a_i), we construct the set $\mathcal{N}_i$ of negative triplets by concatenating incorrect answers a_j within the training batch to the video-question pair (v_i, q_i) as: (v_i, q_i, a_j) with $a_j \neq a_i$. In particular, if the same negative answer a_j is present multiple times in a batch, we only count it once. We found that sampling the same negative answer multiple times leads to worse results (see Section 6.6), which we believe is due to different distributions of answers in the pretraining and downstream datasets. Removing duplicate negatives helps to mitigate this difference.

Finetuning on downstream VideoQA datasets. We leverage the model pretrained on HowToVQA69M and finetune it on a downstream VideoQA dataset that typically has a smaller vocabulary of answers V (*e.g.* $|V| \sim 4000$). To this end, we adapt the training objective in (1) by constructing the negative set $\mathcal{N}_i$ from *all* incorrect answers in V. Note that in such setting (1) becomes equivalent to optimizing the standard cross-entropy objective. In the specific case of multiple-choice VideoQA, the set of negatives $\mathcal{N}_i$ is the set of incorrect answers for each sample.

Masked Language Modeling (MLM). In addition to the contrastive loss (1) we apply the masking loss [19] to question tokens during both pretraining and finetuning. We found this to have a positive regularization effect when finetuning the DistilBERT weights (see Section 6.6).

5. iVQA: new dataset for VideoQA evaluation

In this section we present our **I**nstructional **VQA** dataset (iVQA). We start from a subset of HowTo100M videos and manually annotate video clips with questions and answers. We aim to (i) provide a well-defined evaluation by including five correct answer annotations per question and (ii) avoid questions which can be answered without watching the video. The dataset is described below and more details are given in the Supplementary Material.

Data Collection. iVQA videos are obtained by randomly sampling 7-30 sec. video clips from the HowTo100M dataset [58]. We avoid overlap between datasets and make sure iVQA and HowToVQA69M have no videos in common. Each clip is manually annotated with a question and 5 answers on Amazon Mechanical Turk. We ask workers to annotate questions about objects and scenes in the video and remove videos that could not be annotated. The correctness of annotations is manually verified by the authors. Moreover, we manually reduce the language bias by excluding questions that could be answered without watching the video. To increase diversity, each question is answered by 5 different workers. The answers are restricted to 4 words and are complemented by a confidence level. Questions that receive multiple answers with low confidence are removed.

Statistical Analysis. iVQA contains 10,000 video clips with one question and five corresponding answers per clip. We split the dataset into 60%/20%/20% train/validation/test subsets. On average, questions and answers contain 7.6 and 1.1 words respectively. The average duration of video clips

Method	Pretraining Data	iVQA		MSRVTT-QA		MSVD-QA		ActivityNet-QA		How2QA
		Top-1	Top-10	Top-1	Top-10	Top-1	Top-10	Top-1	Top-10	Top-1
Random	$\emptyset$	0.09	0.9	0.02	0.2	0.05	0.5	0.05	0.5	25.0
QA-T	HowToVQA69M	4.4	23.2	2.5	6.5	4.8	15.0	11.6	45.8	38.4
VQA-T	HowTo100M	1.9	11.9	0.3	3.4	1.4	10.4	0.3	1.9	46.2
VQA-T (Ours)	HowToVQA69M	**12.2**	**43.3**	**2.9**	**8.8**	**7.5**	**22.4**	**12.2**	**46.5**	**51.1**

Table 2: Comparison with baselines for zero-shot VideoQA. Top-1 and top-10 (for open-ended datasets) accuracy are reported.

Question: What design are they making?
GT Answer: rose (4), rose flower (1)
QA-T (HowToVQA69M): pinwheel
VQA-T (HowTo100M): piping bag
VQA-T (HowToVQA69M): rose

Question: What is in the man's hand?
GT Answer: shovel (3), spade (2)
QA-T (HowToVQA69M): coin
VQA-T (HowTo100M): planting
VQA-T (HowToVQA69M): shovel

Question: What type of material is the man touching?
GT Answer: wood (5)
QA-T (HowToVQA69M): brick
VQA-T (HowTo100M): electric saw
VQA-T (HowToVQA69M): wood

Question: What is the woman decorating?
GT Answer: cake (5)
QA-T (HowToVQA69M): cupcakes
VQA-T (HowTo100M): raspberries
VQA-T (HowToVQA69M): cake

Question: What fruit is shown at the end?
GT Answer: watermelon (5)
QA-T (HowToVQA69M): pineapple
VQA-T (HowTo100M): slotted spoon
VQA-T (HowToVQA69M): watermelon

Question: What object is seen on the left, at the beginning of the video?
GT Answer: teapot (4), pot (1)
QA-T (HowToVQA69M): mirror
VQA-T (HowTo100M): espresso
VQA-T (HowToVQA69M): teapot

Figure 6: **Zero-shot VideoQA on iVQA**. The values next to the ground truth (GT) answers indicate the number of annotators that gave the answer.

is 18.6 seconds. The majority of questions have at least 2 annotators providing the same answer. Similarly to [8], this motivates us to define the following accuracy measure for a given answer a: $acc(a) = \min(\frac{\#\text{ground truth answers} = a}{2}, 1)$. This metric assigns 100% accuracy to answers confirmed by at least 2 annotators, 50% accuracy to answers confirmed by only 1 annotator and 0% otherwise. Note that this definition is specific to *multiple* ground truth answers per question.

6. Experiments

This section demonstrates the benefits of training using our generated HowToVQA69M dataset and compares our method to the state of the art. We first outline the used datasets, baseline methods and implementation details in Section 6.1. We then present results for the novel zero-shot VideoQA task in Section 6.2. The comparison to the state of the art in VideoQA and alternative training strategies is given in Section 6.3. Section 6.4 presents results for rare answers. Finally, we compare our VideoQA generation approach to previous methods in Section 6.5 and present ablation studies in Section 6.6.

6.1. Evaluation Protocol

Datasets. We use two datasets for training and five datasets for evaluation as described below. We follow previous evaluation protocols for open-ended settings [40, 91] and use a fixed vocabulary of training answers. Unless stated otherwise, we report top-1 test accuracy and use original splits for training, validation and test.

For training we use our new **HowToVQA69M** dataset introduced in Section 3.2 with 90% and 10% videos in training and validation subsets. For comparison, we also train our model using a large-scale text-video dataset, **HowTo100M** [58], that contains videos with transcribed narrations but *no video-question-answer* triplets. Test and validation videos of downstream datasets are excluded from HowTo100M and HowToVQA69M.

We evaluate results on four open-ended VideoQA downstream datasets: **MSRVTT-QA** [84], **MSVD-QA** [84], **ActivityNet-QA** [91] and our new **iVQA** dataset (see Section 5). We also evaluate on a multiple-choice VideoQA dataset **How2QA** [46] where each question is associated with one correct and three incorrect answers.

Baselines. To evaluate the contribution of the visual modality, we compare our *VQA-T* model with its language-only variant *QA-T*. *QA-T* does not use video input, i.e. we set the input v of the video-question transformer to zero (see Figure 5). To evaluate our generated dataset, we also compare *VQA-T* trained on HowToVQA69M and on HowTo100M. Since HowTo100M has no (v, q, a) triplets, we only train the f branch of *VQA-T* on HowTo100M using the standard masking and cross-modal matching losses [15, 46, 53, 72, 100]. In the zero-shot setting we evaluate *VQA-T* trained on HowTo100M by computing $f(v, [q, a])$ for concatenated pairs of questions and answers $[q, a]$. During finetuning we also initialize the g branch of *VQA-T* with parameters of the text encoding obtained from f (see further details in Supplementary Material).

Implementation details. For the HowToVQA69M training, we use the Adam optimizer and mini-batches with 4096

Pretraining data	iVQA	MSRVTT QA	MSVD QA	ActivityNet QA	How2QA
$\emptyset$	23.0	39.6	41.2	36.8	80.8
HowTo100M	28.1	40.4	43.5	38.1	81.9
HowToVQA69M	**35.4**	**41.5**	**46.3**	**38.9**	**84.4**

Table 3: Benefits of pretraining our *VQA-T* model on our new HowToVQA69M dataset (last row) compared to no pretraining (first row) or pretraining on HowTo100M (second row). In each case our *VQA-T* model was then finetuned on the downstream VideoQA datasets. Top-1 accuracy is reported.

video clips sampled from 128 random videos. The optimization over 10 epochs lasts 2 days on 8 Tesla V100 GPUs. Further details are included in the Supplementary Material.

6.2. Zero-shot VideoQA

In this section, we address the *zero-shot VideoQA* task where we prohibit any manual supervision of visual data during training. We explore this setup to evaluate the generalization of *VQA-T* trained on HowToVQA69M to unseen downstream datasets. For consistency, we use the vocabulary of answers from downstream datasets during testing (see Section 6.1).

Zero-shot results are presented in Table 2. We first observe that the use of visual cues by *VQA-T* outperforms *QA-T* when both models are trained on HowToVQA69M. This demonstrates the importance of the cross-modality in HowToVQA69M despite the VideoQA annotation being exclusively generated from text-only methods. Since HowToVQA69M has been generated using no manual annotation of visual data, our approach is scalable and can lead to further improvements by increasing the dataset size, as we discuss in Section 6.6.

Training on HowToVQA69M significantly outperforms the training on HowTo100M and the random baseline. This confirms the advantage of our HowToVQA69M dataset for the VideoQA task over other generic text-video datasets that do not contain video-question-answer triplets. We emphasize that our training does not use any information about target VideoQA datasets. Qualitative results for zero-shot VideoQA are presented for our approach and compared with baselines in Figure 6. We observe that *QA-T* (trained on HowToVQA69M) provides plausible but video-unrelated answers to the questions. Moreover, *VQA-T* (trained on HowTo100M) is able to associate visual content with related answers, but fails to have a complex multimodal understanding. Our *VQA-T* model trained on HowToVQA69M, on the other hand, correctly understands questions and uses information in the video to provide correct answers, confirming results in Table 2.

6.3. Benefits of HowToVQA69M pretraining

This section evaluates the effect of *VQA-T* pretraining in combination with finetuning on target datasets. As shown in Table 3, pretraining on HowToVQA69M provides consistent and significant improvements for all datasets when compared to pretraining on HowTo100M and no pretraining. In particular, we observe the largest improvement for our new iVQA dataset which comes from the same domain as HowToVQA69M. Hence, the automatic generation of training data for other domains using our method can lead to further improvements on other datasets.

We compare our pretrained model to the state-of-the-art in VideoQA in Tables 4-5. Notably, *VQA-T* pretrained on HowToVQA69M outperforms previous methods on all tested datasets. In particular, our method improves over the recent CoMVT approach [66] that has been pretrained on HowTo100M. These strong results show the importance of our proposed HowToVQA69M dataset.

Method	Pretraining data	MSRVTT-QA	MSVD-QA
E-SA [84]		29.3	27.6
ST-TP [33]		30.9	31.3
AMU [84]		32.5	32.0
Co-mem [26]		32.0	31.7
HME [22]		33.0	33.7
LAGCN [31]		—	34.3
HGA [35]		35.5	34.7
QueST [34]		34.6	36.1
HCRN [40]		35.6	36.1
ClipBERT [42]	COCO [14] + Visual Genome [39]	37.4	—
SSML [6]	HowTo100M	35.1	35.1
CoMVT [66]	HowTo100M	39.5	42.6
VQA-T	$\emptyset$	39.6	41.2
VQA-T	HowToVQA69M	**41.5**	**46.3**

Table 4: Comparison with state of the art on MSRVTT-QA and MSVD-QA (top-1 accuracy).

	Pretraining data	ActivityNet QA	How2QA
E-SA [91]		31.8	—
MAR-VQA [102]		34.6	—
HERO [46]	HowTo100M + TV Dataset	—	74.1
CoMVT [66]	HowTo100M	38.8	82.3
VQA-T	$\emptyset$	36.8	80.8
VQA-T	HowToVQA69M	**38.9**	**84.4**

Table 5: Comparison with state of the art on ActivityNet-QA and the public val set of How2QA (top-1 accuracy).

Pretraining data	Finetuning	Q1	Q2	Q3	Q4
$\emptyset$	✓	38.4	16.7	5.9	2.6
HowTo100M	✓	46.7	22.0	8.6	3.6
HowToVQA69M	✗	9.0	8.0	9.5	7.7
	✓	**47.9**	**28.1**	**15.6**	**8.5**

Table 6: Results of our *VQA-T* model with different training strategies, on subsets of iVQA corresponding to four quartiles with Q1 and Q4 corresponding to samples with most frequent and least frequent answers, respectively.

Generation Method	Zero-shot			Finetune		
	iVQA	ActivityNet QA	How2QA	iVQA	ActivityNet QA	How2QA
[29]	7.4	1.1	41.7	31.4	38.5	83.0
Ours	**12.2**	**12.2**	**51.1**	**35.4**	**38.9**	**84.4**

Table 7: Comparison of our question-answer generation approach with Heilman *et al*. [29], evaluated by downstream performance of the model trained on the generated VideoQA data.

6.4. Results for rare answers

Training on downstream VideoQA datasets typically leads to particularly large improvements for questions with most frequent answers. As shown in Table 6, our approach brings significant improvements both for common and rare answers compared to models trained from scratch or pretrained on HowTo100M. Interestingly, for the most rare answers in iVQA (Q3 and Q4) our model without finetuning (zero-shot mode) outperforms finetuned models that have not been pretrained on HowToVQA69M. We make similar observations for rare answers in other datasets and report corresponding results in the Supplementary Material. We conclude that VideoQA specific pretraining on additional large-scale, diverse data helps improve generalization of VideoQA models.

6.5. Comparison of VideoQA generation methods

In this section, we compare our question-answer generation approach to Heilman *et al*. [29], that was notably used in [84, 93, 95, 96, 97] to generate VideoQA data from video descriptions. We run the method of [29] on sentences extracted from HowTo100M, apply our pretraining method on the generated data and show results in Table 7. Note that we do not choose MSRVTT-QA and MSVD-QA as downstream datasets for this comparison because their evaluation sets were automatically generated using Heilman *et al*. [29]. We find that our generation method leads to significantly better performance both in zero-shot and finetuning settings. We also provide a qualitative comparison in the Supplementary Material, further demonstrating the benefit of our transformer-based question-answer generation approach compared to previous methods. We also show the benefit of our generated HowToVQA69M dataset by comparing our results to cross-dataset transfer using existing VideoQA datasets in the Supplementary Material.

6.6. Ablation studies

Pretraining losses. As shown in Table 8, removing duplicate negative answers in our contrastive loss, as discussed in Section 4.2, is beneficial notably in the zero-shot setting. Moreover, adding the MLM loss at pretraining improves the downstream results for both zero-shot and finetuning when used in combination with our contrastive learning strategy. These results motivate our proposed pretraining approach.

MLM	Sampling without answer repetition	Zero-shot		Finetune	
		iVQA	MSVD-QA	iVQA	MSVD-QA
✗	✗	11.1	6.1	34.7	45.6
✗	✓	12.1	7.0	34.3	45.0
✓	✗	10.9	6.4	34.3	45.1
✓	✓	**12.2**	**7.5**	**35.4**	**46.3**

Table 8: Effect of MLM loss and our negative sampling strategy on HowToVQA69M training.

Pretraining data size	Zero-shot		Finetune	
	iVQA	MSVD-QA	iVQA	MSVD-QA
0%	—	—	23.0	41.2
1%	4.5	3.6	24.2	42.8
10%	9.1	6.2	29.2	44.4
20%	9.5	6.8	31.3	44.8
50%	11.3	7.3	32.8	45.5
100%	**12.2**	**7.5**	**35.4**	**46.3**

Table 9: Effect of the training size of HowToVQA69M.

Importance of scale. Results of our method after pretraining on different fractions of HowToVQA69M are shown in Table 9. We construct these subsets such that larger subsets include the smaller ones. These results suggest that the scale is an important factor and that we can expect further improvements with additional pretraining data, both in the zero-shot and finetuning settings.

7. Conclusion

We propose a novel and scalable approach for training VideoQA models without manually annotated visual data. We automatically generate HowToVQA69M – a large-scale VideoQA training dataset generated from narrated videos with readily-available speech transcripts, significantly exceeding existing datasets by size and diversity. We demonstrate several benefits of pretraining on HowToVQA69M. We are the first to demonstrate zero-shot VideoQA results without the use of any manually annotated images or videos. Furthermore, finetuning our HowToVQA69M pretrained model on downstream tasks outperforms the state of the art on MSRVTT-QA, MSVD-QA, ActivityNet-QA and How2QA. We further validate our approach on a new iVQA benchmark we manually collect.

Acknowledgements. This work was granted access to the HPC resources of IDRIS under the allocation 2020-101267 made by GENCI. The work was funded by a Google gift, the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), the Louis Vuitton ENS Chair on Artificial Intelligence, the European Regional Development Fund under project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15 003/0000468) and A. Miech's Google PhD fellowship. We thank P.-L. Guhur and M. Tapaswi for advice on using Amazon Mechanical Turk, E. Berthier, Q. Le Lidec and E. Chane-Sane for the manual evaluation of generated VideoQA data, and I. Rocco for proofreading.

References

[1] Just Ask project webpage. https://antoyang.github.io/just-ask.html. 2

[2] Punctuator. https://github.com/ottokart/punctuator2, 2017. 4

[3] Question generation using transformers. https://github.com/patil-suraj/question_generation, 2020. 4

[4] Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. Synthetic QA corpora generation with roundtrip consistency. In *ACL*, 2019. 4

[5] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *IJCNLP*, 2019. 3

[6] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *AAAI*, 2021. 3, 7

[7] 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 *CVPR*, 2018. 2

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015. 6

[9] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal tucker fusion for visual question answering. In *CVPR*, 2017. 2

[10] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 1

[11] Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. LifeQA: A real-life dataset for video question answering. In *LREC*, 2020. 2

[12] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iPerceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. In *WACV*, 2021. 2

[13] Ying-Hong Chan and Yao-Chung Fan. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, 2019. 4

[14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3, 7

[15] 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, 6

[16] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Seungchan Lee, Minsu Lee, and Byoung-Tak Zhang. DramaQA: Character-centered video story understanding with hierarchical qa. In *AAAI*, 2021. 2

[17] Anthony Colas, Seokhwan Kim, Franck Dernoncourt, Siddhesh Gupte, Daisy Zhe Wang, and Doo Soon Kim. TutorialVQA: Question answering dataset for tutorial videos. In *LREC*, 2020. 2

[18] Karan Desai and Justin Johnson. VirTex: Learning visual representations from textual annotations. In *CVPR*, 2021. 3

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 5

[20] Xinya Du, Junru Shao, and Claire Cardie. Learning to ask: Neural question generation for reading comprehension. In *ACL*, 2017. 3

[21] Chenyou Fan. EgoVQA - an egocentric video question answering benchmark dataset. In *ICCV Workshops*, 2019. 2

[22] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *CVPR*, 2019. 1, 2, 7

[23] Marcello Federico, Sebastian Stüker, Luisa Bentivogli, Michael Paul, Mauro Cettolo, Teresa Herrmann, Jan Niehues, and Giovanni Moretti. The IWSLT 2011 evaluation campaign on automatic talk translation. In *LREC*, 2012. 4

[24] 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. In *EMNLP*, 2016. 2

[25] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 3

[26] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *CVPR*, 2018. 2, 7

[27] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. KnowIT VQA: Answering knowledge-based questions about videos. In *AAAI*, 2020. 2

[28] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: : Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 2

[29] Michael Heilman and Noah A Smith. Good question! Statistical ranking for question generation. In *ACL*, 2010. 2, 4, 8

[30] Hexiang Hu, Wei-Lun Chao, and Fei Sha. Learning answer embeddings for visual question answering. In *CVPR*, 2018. 2

[31] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *AAAI*, 2020. 1, 2, 5, 7

[32] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 3

[33] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 2, 7

[34] 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 *AAAI*, 2020. 1, 2, 5, 7

[35] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *AAAI*, 2020. 1, 2, 5, 7

[36] Hyounghun Kim, Zineng Tang, and Mohit Bansal. Dense-caption matching and frame-selection gating for temporal localization in VideoQA. In *ACL*, 2020. 2

[37] Junyeong Kim, Minuk Ma, Trung Pham, Kyungsu Kim, and Chang D Yoo. Modality shifting attention network for multi-modal video question answering. In *CVPR*, 2020. 2

[38] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. In *IJCAI*, 2017. 2

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016. 3, 7

[40] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *CVPR*, 2020. 1, 2, 5, 6, 7

[41] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Neural reasoning, fast and slow, for video question answering. In *IJCNN*, 2020. 2

[42] 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 *CVPR*, 2021. 1, 2, 7

[43] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, 2018. 2

[44] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. TVQA+: Spatio-temporal grounding for video question answering. In *ACL*, 2020. 2

[45] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 3

[46] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*, 2020. 2, 3, 6, 7

[47] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[48] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond RNNs: Positional self-attention with co-attention for video question answering. In *AAAI*, 2019. 1, 2

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

[50] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou. Visual question generation as dual task of visual question answering. In *CVPR*, 2018. 2

[51] 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 *ECCV*, 2014. 2

[52] Luis Enrico Lopez, Diane Kathryn Cruz, Jan Christian Blaise Cruz, and Charibeth Cheng. Transformer-based end-to-end question generation. *arXiv preprint arXiv:2005.01107*, 2020. 4

[53] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 3, 6

[54] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 3

[55] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016. 2

[56] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. UniViLM: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3

[57] 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 *CVPR*, 2020. 2, 3, 5

[58] 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 *ICCV*, 2019. 1, 3, 4, 5, 6

[59] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. In *ACL*, 2016. 2

[60] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. MarioQA: Answering questions by watching gameplay videos. In *CVPR*, 2017. 2

[61] 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. *arXiv preprint arXiv:2103.00020*, 2021. 2

[62] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 3, 4

[63] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine

comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 4

[64] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NeurIPS*, 2015. 2

[65] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 5

[66] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *CVPR*, 2021. 3, 7

[67] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-consistency for robust visual question answering. In *CVPR*, 2019. 2

[68] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018. 3

[69] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *ACM international conference on Multimedia*, 2018. 2

[70] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 3

[71] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019. 3

[72] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *ICCV*, 2019. 3, 6

[73] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 3

[74] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding stories in movies through question-answering. In *CVPR*, 2016. 2

[75] Damien Teney and Anton van den Hengel. Zero-shot visual question answering. *arXiv preprint arXiv:1611.05546*, 2016. 2

[76] Ottokar Tilk and Tanel Alumäe. Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Interspeech 2016*, 2016. 4

[77] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[78] Ben-Zion Vatashsky and Shimon Ullman. VQA with no questions-answers training. In *CVPR*, 2020. 2

[79] Weining Wang, Yan Huang, and Liang Wang. Long video question answering: A matching-guided attention model. *Pattern Recognition*, 2020. 2

[80] Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. On modality bias in the TVQA dataset. *arXiv preprint arXiv:2012.10210*, 2020. 2

[81] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 5

[82] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 5

[83] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016. 2

[84] 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 *ACM international conference on Multimedia*, 2017. 2, 6, 7, 8

[85] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *ECCV*, 2016. 2

[86] Hongyang Xue, Wenqing Chu, Zhou Zhao, and Deng Cai. A better way to attend: Attention with trees for video question answering. *IEEE Transactions on Image Processing*, 2018. 2

[87] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. BERT representations for video question answering. In *WACV*, 2020. 2

[88] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 2

[89] Kaichun Yao, Libo Zhang, Tiejian Luo, Lili Tao, and Yanjun Wu. Teaching machines to ask questions. In *IJCAI*, 2018. 3

[90] Yunan Ye, Zhou Zhao, Yimeng Li, Long Chen, Jun Xiao, and Yueting Zhuang. Video question answering via attribute-augmented attention network learning. In *ACM SIGIR*, 2017. 2

[91] 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 *AAAI*, 2019. 2, 6, 7

[92] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-IQ: A question answering benchmark for artificial social intelligence. In *CVPR*, 2019. 2

[93] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *AAAI*, 2017. 2, 8

[94] 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)*, 2019. 2

[95] Zhou Zhao, Shuwen Xiao, Zehan Song, Chujie Lu, Jun Xiao, and Yueting Zhuang. Open-ended video question an-

swering via multi-modal conditional adversarial networks. *IEEE Transactions on Image Processing*, 2020. 2, 8

[96] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, Yueting Zhuang, Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *IJCAI*, 2017. 2, 8

[97] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *IJCAI*, 2018. 2, 8

[98] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, 2020. 3

[99] Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. Neural question generation from text: A preliminary study. In *National CCF Conference on Natural Language Processing and Chinese Computing*, 2017. 3

[100] Linchao Zhu and Yi Yang. ActBERT: Learning global-local video-text representations. In *CVPR*, 2020. 3, 6

[101] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015. 5

[102] Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. Multichannel attention refinement for video question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2020. 2, 5, 7