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

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

In this Supplementary Material, we start by giving an additional data analysis and examples of our proposed HowToVQA69M dataset in Section A. We, then, provide additional architecture details for our VideoQA model in Section B. Next, we present additional statistics and details of the collection procedure for our manually collected iVQA evaluation benchmark in Section C. We describe additional implementation details in Section D and present experiments including cross-dataset transfer, results per answer quartile and per question type in Section E.

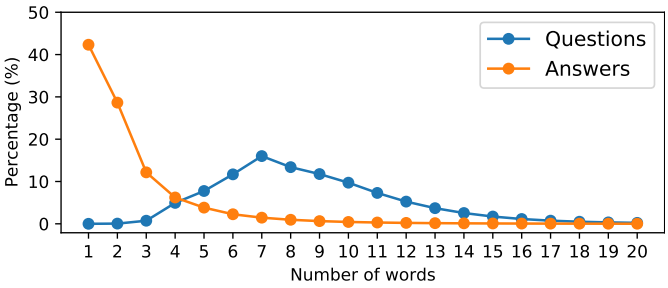
A. Analysis of HowToVQA69M dataset

Figure 1 shows the statistics of the HowToVQA69M dataset in terms of the question length, answer length and video clip duration. Overall, HowToVQA69M contains longer answers than downstream open-ended VideoQA datasets like MSRVTT-QA, MSVD-QA or ActivityNet-QA. The distribution of clip duration has a peak at around seven seconds with a long tail of longer clips. These statistics demonstrate the diversity of our HowToVQA69M dataset, both in terms of videos and answers.

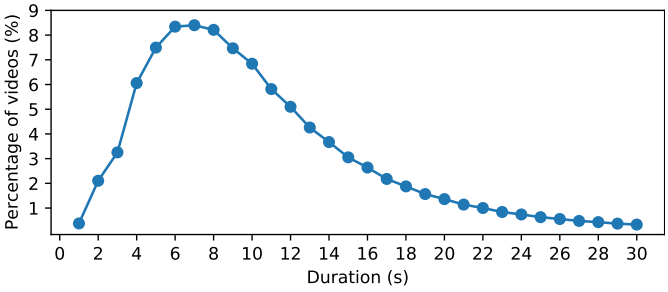
Word clouds[1] for questions and answers in HowToVQA69M are shown in Figure 2 and illustrate the diverse vocabulary in HowToVQA69M as well as the presence of speech-related words such as as *okay*, *right*, *oh*. In Figure 4 we illustrate the diversity and the noise in the automatically obtained annotations in the HowToVQA69M dataset.

We show quantitative comparisons of our question-answer generation models with [2] in the main paper (Section 6.5), and supplement it here with a qualitative comparison shown in Figure 3. We found that compared to [2] our generation method provides higher quality as well as higher diversity of question-answer pairs when applied to the uncurated sentences extracted from speech in narrated videos.

In the main paper (Section 3.2) we present a manual evaluation of the quality of the automatically generated video-question-answer triplets for our method and two other baselines. We complement this analysis here with inter-rater agreement statistics. For the 300 generated video-question-



(a) Question and answer length



(b) Clip duration

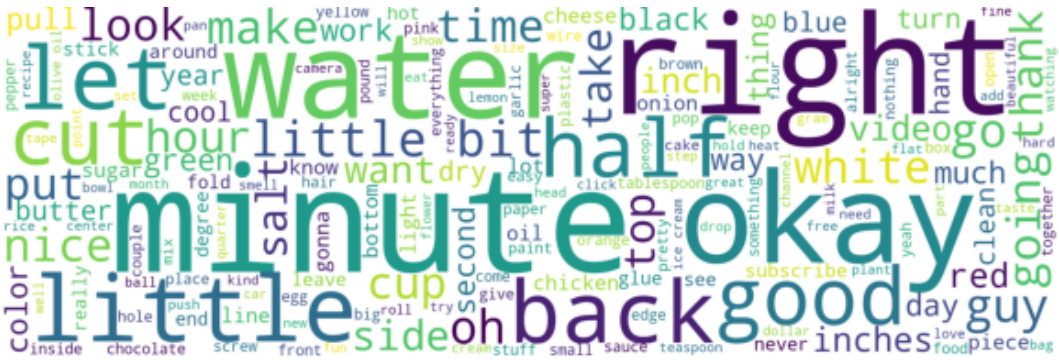
Figure 1: **Statistics of the HowToVQA69M dataset.** (a) Distribution of length of questions and answers. (b) Distribution of video clip duration in seconds.

Question Type	Total	Correct Samples (%)	QA Generation Failure (%)	QA unrelated to video (%)
Attribute	25	28	32	40
Object	17	41	24	35
Action	16	**69**	19	13
Counting	13	23	15	**62**
Place	7	0	**86**	14
People	7	0	43	57
Other	15	13	27	60

Table 1: Manual evaluation of our video-question-answer generation method on 100 randomly chosen generated examples split by question type. Results are obtained by majority voting among 5 annotators.

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

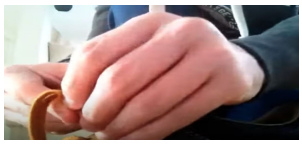
[1]To generate the word clouds, we used https://github.com/amueller/word_cloud.



(a) Answers

(b) Questions

Figure 2: Word clouds extracted from the HowToVQA69M dataset showing its diverse vocabulary and words characteristic to speech such as *okay*, *right*, or *ok*.



ASR: And then just squeeze it through like that.
Question (Heilman et al): What do then just squeeze through like that?
Answer (Heilman et al): it
Question (ours): How do you do it?
Answer (ours): squeeze it through

ASR: It is a staple in a lot of asian kitchens.
Question (Heilman et al): What is it?
Answer (Heilman et al): a staple in a lot of asian kitchens
Question (ours): In what type of kitchens is it a staple?
Answer (ours): asian kitchens

ASR: And you want it over a very low heat.
Question (Heilman et al): What do you want it over?
Answer (Heilman et al): over a very low heat
Question (ours): What kind of heat do you want it to be over?
Answer (ours): low heat

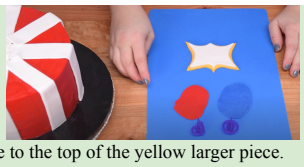
ASR: This is classic premium chicken, grilled sandwich.
Question (Heilman et al): What is classic premium chicken, grilled sandwich?
Answer (Heilman et al): this
Question (ours): What type of sandwich is this?
Answer (ours): classic premium chicken, grilled sandwich

ASR: But why do that when you can enjoy the plant for about three months, it'll, keep producing because the leaves grow from the center
Question (Heilman et al): What leaves?
Answer (Heilman et al): the
Question (ours): What part of the plant grows from the center?
Answer (ours): leaves

ASR: Next add half a cup of powdered milk and a little shake a quarter teaspoon of salt, which I know, sounds really weird.
Question (Heilman et al): What do I know the quarter teaspoon of?
Answer (Heilman et al): of salt
Question (ours): What is a quarter teaspoon of?
Answer (ours): salt

Figure 3: Qualitative examples of video-question-answer triplets generated with our trained language models compared to Heilman *et al*. [2], illustrating the higher quality and diversity of triplets obtained with our generation method.


ASR: Then you release the right and you take out the tube pretty simple.
Question: What do you take out?
Answer: The tube

ASR: So I transferred my smaller piece to the top of the yellow larger piece.
Question: What color was the larger piece?
Answer: Yellow

ASR: You can just lift them right up like that that there we go, and these are ready to cut.
Question: What do you do to get them ready to cut?
Answer: Lift them right up

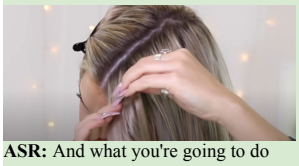

ASR: ...and we're gonna create slipknot by pulling this side of the rubber band through the center of this side.
Question: How do we create slipknot?
Answer: Pull that through

ASR: And the last thing that goes on top would be the spinach.
Question: What is the last thing that goes on top?
Answer: Spinach

ASR: So I've got nine blobs of dough here a little bit sticky.
Question: How many blobs of dough are there?
Answer: Nine

ASR: So you bring it to a point and we'll, just cut it off at the bottom.
Question: What do we do at the bottom?
Answer: Cut it off

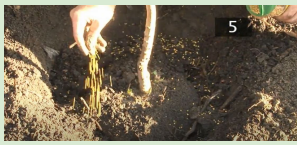
ASR: And then voila, perfect chocolate mousse.
Question: What kind of mousse is perfect?
Answer: Chocolate



ASR: And what you're going to do is take the first section underneath and pull that nice and tight.
Question: What are you going to do with the first section underneath?
Answer: Pull that nice and tight

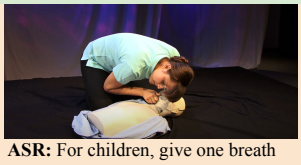
ASR: Hi I'm long lamb and today, I'm going to teach you how to prep a pineapple...
Question: What will I teach you today?
Answer: How to prep a pineapple

ASR: The violin is covered in varnish..
Question: What is the violin covered in?
Answer: Varnish



ASR: The soil can be mixed with compost or slow release fertilizer to help nourish your tree...
Question: What can be mixed with the soil to help nourish your tree?
Answer: Compost or slow release fertilizer

ASR: ...thai airbus, 340 - 600 arrived from bangkok ...
Question: What is the average size of an airbus from bangkok?
Answer: 340 - 600



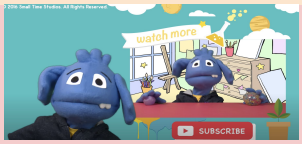
ASR: For children, give one breath every 3 to 5 seconds.
Question: How long does it take for a child to take a breath?
Answer: 3 to 5 seconds

ASR: Just let them do whatever they want and it'll still look pretty cool.
Question: What's the best way to make it look cool?
Answer: Let them do whatever they wants

ASR: The onions are chopped pretty much the same size.
Question: What are chopped pretty much the same size as the other vegetables?
Answer: The onions

ASR: I I you know, I I think this mod is really really awesome.
Question: I think this mod is what?
Answer: Really really awesome

ASR: Let me explain to you guys.
Question: What do I say to you guys?
Answer: Let me explain to you guys



ASR: You can't miss this..
Question: What can't you do?
Answer: Miss

ASR: And I will put it in a 400 degree oven for 15 minutes.
Question: How many minutes will peppers be in the 400 degree oven?
Answer: 15

Figure 4: Additional examples of videos, questions and answers from our automatically generated HowToVQA69M dataset. These examples illustrate the large data diversity in HowToVQA69M. The green color indicates relevant examples, the orange color (penultimate row) indicates a failure of the question-answer generation, and the red color (last row) indicates that the generated question-answer is unrelated to the visual content.

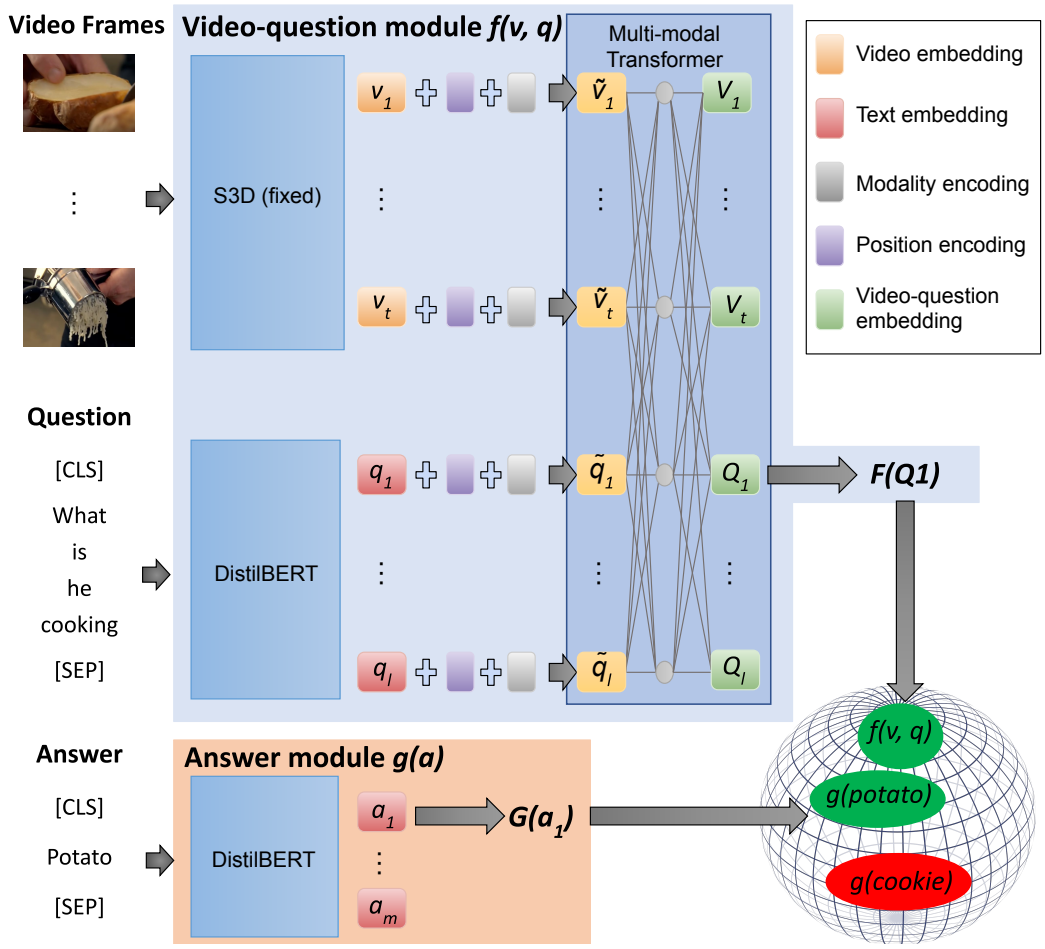


Figure 5: **VideoQA architecture overview.** Our model is composed of a video-question module f based on a multi-modal transformer (top) and an answer module g based on DistilBERT [6] encoder (bottom).

answer triplets (100 for each generation method), 94 were in an agreement by all 5 annotators, 198 in an agreement by at least 4 annotators, and 299 in an agreement by at least 3 annotators. This high agreement among annotators demonstrates the reliability of the results in Table 1 of the main paper.

We further manually classify the 100 video-question-answer triplets obtained with our method by the question type ("Attribute", "Object", "Action", "Counting", "Place", "People", or "Other"), evaluate the quality of generated triplets for different question types and report results in Table 1. Out of the 6 most common categories, we observed that questions related to "Action" lead to the best annotations, "Counting" questions lead to the highest number of QAs unrelated to the video content, and questions related to "Place" lead to the highest number of QA generation errors. Qualitatively, we found that actions are often depicted in the video, while counted quantities (*e.g.* time, weight, length) mentioned in the speech are hard to guess from the video only.

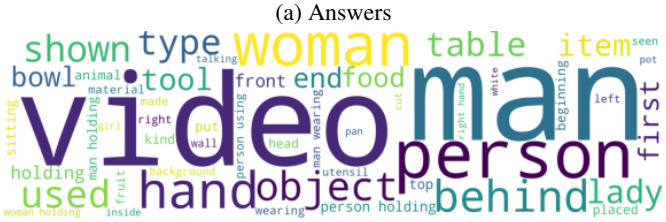
B. VideoQA architecture

Our architecture, shown in Figure 5, has two main modules: (i) a video-question multi-modal transformer (top) and (ii) an answer transformer (bottom). Details are given next, and further implementation details are given in Section D.

Video-question multi-modal transformer. The input video representation, obtained from a fixed S3D model [8], is composed of t features denoted $v = [v_1, ..., v_t] \in \mathbb{R}^{d_v \times t}$ where d_v is the dimension of the video features, and t is the number of extracted features, one per second. The contextualized representation of the question, provided by the DistilBERT model [6], is composed of l token embeddings denoted as $q = [q_1, ..., q_l] \in \mathbb{R}^{d_q \times l}$ where d_q is the dimension of the DistilBERT embedding and l is the number of tokens in the question. The inputs to our video-question multi-modal transformer are then defined as a concatenation of question token embeddings and video features

$$u(v, q) = \left[\tilde{q}_1, ..., \tilde{q}_l, \tilde{v}_1, ..., \tilde{v}_t\right] \in \mathbb{R}^{d \times (l+t)}, \quad (1)$$

(a) Answers



(b) Questions

Figure 6: Word clouds for our iVQA dataset illustrate a vocabulary related to the domains of cooking, hand crafting, or gardening. The frequent occurrence of location and time-specific words (*behind*, *front*, *right*, *left*, *first*, *end*, *beginning*) indicate the presence of the spatial and temporal context within iVQA questions.

where

$$\widetilde{q}_s = dp\left(\sigma\left(W_q q_s + b_q\right) + pos_s + mod_q\right), \qquad (2)$$

and

$$\widetilde{v}_s = dp(\sigma(W_v v_s + b_v) + pos_s + mod_v), \qquad (3)$$

where $W_q \in \mathbb{R}^{d_q \times d}$, $b_q \in \mathbb{R}^d$, $W_v \in \mathbb{R}^{d_v \times d}$, $b_v \in \mathbb{R}^d$ and learnable parameters, $mod_q \in \mathbb{R}^d$ and $mod_v \in \mathbb{R}^d$ are learnt modality encodings for question and video, respectively, and $[pos_1, ..., pos_{l+t}] \in \mathbb{R}^{d \times (l+t)}$ are fixed sinusoidal positional encodings. σ is a Gaussian Error Linear Unit [3] followed by a Layer Normalization [1] and dp refers to Dropout [7].

The multi-modal transformer is a transformer with N layers, h heads, dropout probability p_d, and hidden dimension d_h. The outputs of the multi-modal transformer $[Q_1, ...Q_l, V_1...V_t] \in \mathbb{R}^{d \times (l+t)}$ are contextualized representations over tokens in the question and temporal video representations. Finally, the fused video-question embedding $f(v, q)$ is obtained as

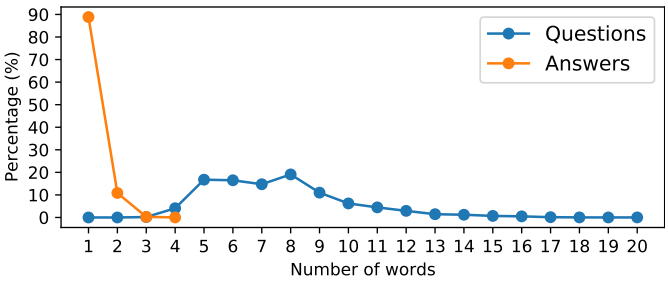
$$F(Q_1) = W_{vq} dp(Q_1) + b_{vq}, \qquad (4)$$

where $W_{vq} \in \mathbb{R}^{d \times d}$, $b_{vq} \in \mathbb{R}^d$ are learnable parameters and Q_1 is the multi-modal contextualized embedding of the [CLS] token in the question, as shown in Figure 5.

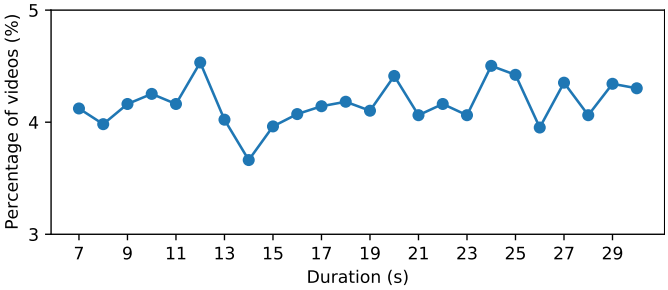
Answer transformer. The contextualized representation of the answer, provided by the DistilBERT model [6], is composed of m token embeddings denoted as $a = [a_1, ..., a_m] \in \mathbb{R}^{d_a \times m}$ where d_a is the dimension of the DistilBERT embedding and m is the number of tokens in the answer. Our answer embedding $g(a)$ is then obtained as

$$G(a_1) = W_a a_1 + b_a, \qquad (5)$$

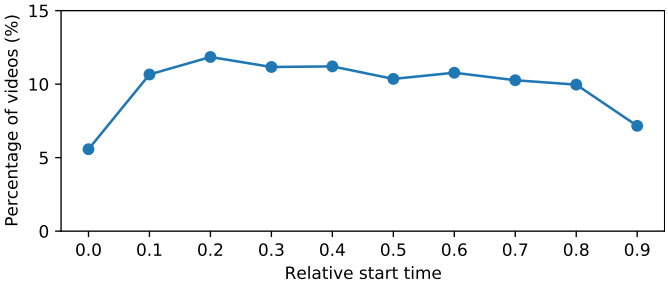
where $W_a \in \mathbb{R}^{d_a \times d}$, $b_a \in \mathbb{R}^d$ are learnable parameters and a_1 is the contextualized embedding of the [CLS] token in the answer, as shown in Figure 5.



(a) Question and answer length



(b) Clip duration



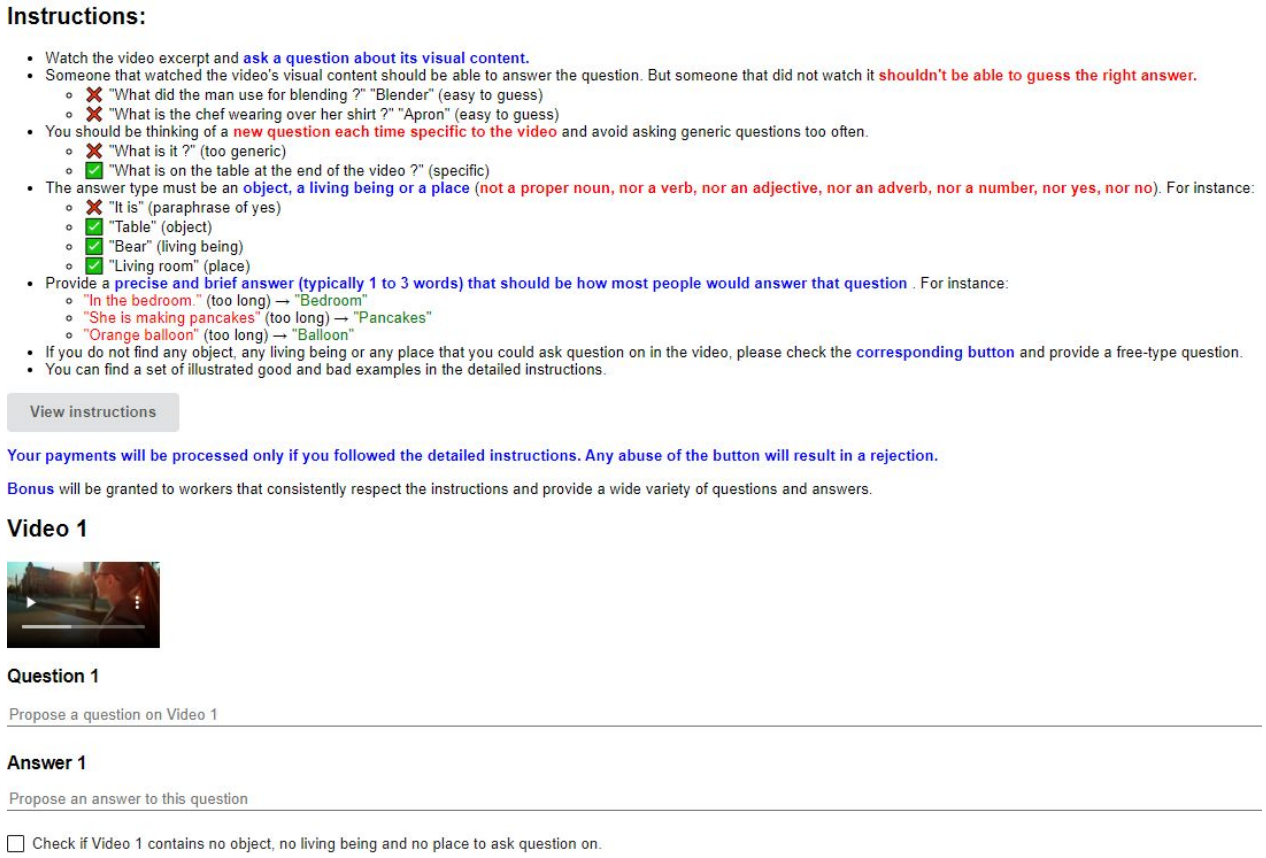
(c) Clip start time in the original video

Figure 7: **Statistics of the iVQA dataset.** (a) Distribution of length of questions and answers. (b) Distribution of video clip duration in seconds. (c) Distribution of video clip relative start time in the original video.

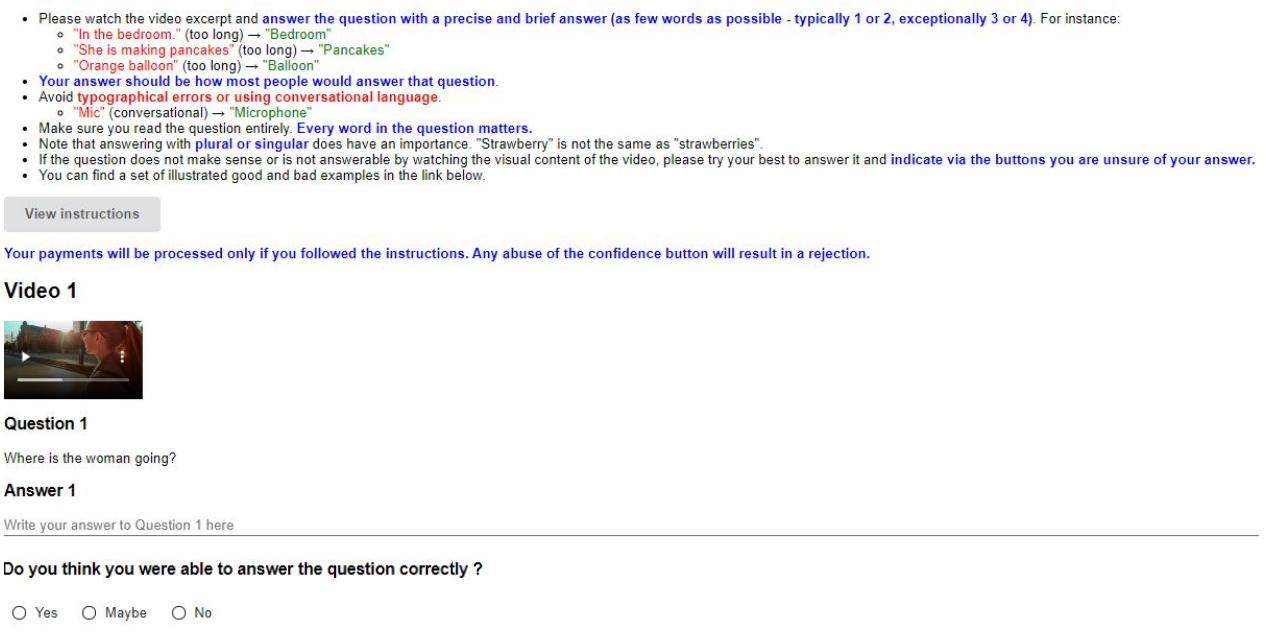
C. Details of the iVQA dataset

C.1. Data Collection

The Amazon Mechanical Turk interfaces used for collecting the question and answer annotations, are shown in Figure 8. An emphasis was placed on collecting visually grounded questions about objects and scenes that could not be easily guessed without watching the video, and collecting short answers in order to maximize the chance for consensus between annotators, *i.e.*, having multiple annotators giving exactly the same answer.



(a) Collection interface for questions. Note that the answer provided by the question annotator is only used to ensure that the provided question follows the given instructions, but is not included in iVQA. Answers are collected separately, see Figure 8b.



(b) Collection interface for answers. Five different answer annotators provide an answer annotation for each collected question.

Figure 8: Amazon Mechanical Turk interfaces for collecting questions (Figure 8a) and answers (Figure 8b) for the iVQA dataset. For readability, the videos shown in these Figures are shrinked, and only one annotation example is shown.

Pretraining Data	Zero-shot				Finetune			
	iVQA	MSRVTT-QA	ActivityNet-QA	How2QA	iVQA	MSRVTT-QA	ActivityNet-QA	How2QA
∅	—	—	—	—	23.0	39.6	36.8	80.8
MSRVTT-QA	8.6	—	1.7	42.5	25.2	—	37.5	80.0
ActivityNet-QA	5.5	2.7	—	40.8	24.0	39.9	—	80.7
HowToVQA69M	**12.2**	**2.9**	**12.2**	**51.1**	**35.4**	**41.5**	**38.9**	**84.4**

Table 2: Comparison of our training on HowToVQA69M with cross-dataset transfer using the previously largest open-ended VideoQA dataset (MSRVTT-QA) and the largest manually annotated open-ended VideoQA dataset (ActivityNet-QA).

Pretraining Data	Finetuning	MSRVTT-QA				MSVD-QA				ActivityNet-QA			
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
	✓	**68.4**	44.1	32.9	8.1	71.2	53.7	28.9	8.8	65.6	49.0	25.7	3.9
HowTo100M	✓	65.2	46.4	34.9	10.6	**74.8**	58.8	30.6	10.5	**67.5**	**53.3**	25.9	4.1
HowToVQA69M	✗	0.2	6.4	2.4	3.0	9.3	9.0	6.9	4.8	36.3	5.7	3.7	1.5
HowToVQA69M	✓	66.9	**46.9**	**36.0**	**11.5**	74.7	**59.0**	**35.0**	**14.1**	66.3	53.0	**28.0**	**5.0**

Table 3: Results of our *VQA-T* model with different training strategies, on subsets of MSRVTT-QA, MSVD-QA and ActivityNet-QA, corresponding to four quartiles with Q1 and Q4 corresponding to samples with the most frequent and the least frequent answers, respectively.

C.2. Statistical Analysis

Word clouds for questions and answers in iVQA, shown in Figure 6, demonstrate the relation of iVQA to the domains of cooking, hand crafting and gardening. These word clouds also indicate that questions in iVQA often require spatial reasoning (*behind*, *front*, *right*, *left*) and temporal understanding (*first*, *end*, *left*, *beginning*) of the video. The most frequent answer (*spoon*) in iVQA corresponds to 2% of all answers in the dataset. In contrast, the most frequent answers in other VideoQA datasets account for more than 9% of all answers in these datasets (we have verified this for MSRVTT-QA, MSVD-QA and ActivityNet-QA). As a consequence, the *most frequent answer baseline* is significantly lower for our iVQA dataset compared to other VideoQA datasets. Figure 7 shows the distributions of question length, answer length, clip duration and clip relative start time in the original video. Clip duration and start time distributions are almost uniform because we randomly sampled them to obtain the clips, which results in a high video content diversity. Answers are in great majority one or two words as a result of our collection procedure.

We observe that 27.0% of questions lead to a perfect consensus among the five answer annotators, 48.4% of questions lead to a consensus among at least four annotators, and 77.3% lead to a consensus among at least three annotators, while only six questions do not lead to a consensus between at least two annotators, justifying the defined accuracy metric. Additionally, 27.5% of questions have two different answers that had a consensus between at least two annotators.

D. Additional experimental details

VideoQA generation. The input sequence to the answer extractor and question generation transformers are truncated and padded up to a maximum of 32 tokens. The question decoding is done with beam search keeping track of the 4 most probable states at each level of the search tree. We have used the original captions (including stop words) from the HowTo100M dataset [5] and removed word repetitions from adjacent clips.

VideoQA model. We use the following hyperparameters: $l = 20$, $t = 20$, $m = 10$, $d = 512$, $d_h = 2048$, $N = 2$, $H = 8$, $p_d = 0.1$, $d_q = d_a = 768$, $d_v = 1024$. The video features are sampled at equally spaced timestamps, and padded to length t. Sequences of question and answer tokens are truncated and padded to length l and m, respectively. Attention is computed only on non-padded sequential video and question features.

VideoQA datasets. For MSRVTT-QA and MSVD-QA, we follow [4] and use a vocabulary made of the top 4000 training answers for MSRVTT-QA, and all 1852 training answers for MSVD-QA. For our iVQA dataset and ActivityNet-QA, we consider all answers that appear at least twice in the training set, resulting in 2348 answers for iVQA and 1654 answers for ActivityNet-QA.

Training. We use a cosine annealing learning rate schedule with initial values of 5×10^{-5} and 1×10^{-5} for pretraining and finetuning, respectively. For finetuning, we use the Adam optimizer with batch size of 256 and training runs for 20 epochs. The final model is selected by the best performance on the validation set.

Masked Language Modeling. For the masked language modeling objective, a token is corrupted with a probability 15%, and replaced 80% of the time with [MASK], 10% of the time with the same token and 10% of the time with a randomly sampled token. To guess which token is masked, each sequential question output Q_i of the multimodal transformer is classified in a vocabulary of 30,522 tokens, and we use a cross-entropy loss.

Pretraining on HowTo100M. For video-text cross-modal matching, we sample one video negative and one text neg-

Pretraining Data	Finetuning	MSRVTT-QA						MSVD-QA					
		What	Who	Number	Color	When	Where	What	Who	Number	Color	When	Where
	✓	33.4	49.8	83.1	50.5	78.5	40.2	31.5	54.9	**82.7**	50.0	74.1	46.4
HowTo100M	✓	34.3	50.2	82.7	**51.8**	80.0	41.5	34.3	58.6	82.4	**62.5**	**77.6**	**50.0**
HowToVQA69M	✗	1.8	0.7	66.3	0.6	0.6	4.5	7.8	1.7	74.3	18.8	3.5	0.0
HowToVQA69M	✓	**35.5**	**51.1**	**83.3**	49.2	**81.0**	**43.5**	**37.9**	**58.0**	80.8	**62.5**	**77.6**	46.4

Table 4: Effect of our pretraining per question type on MSRVTT-QA and MSVD-QA.

Pretraining Data	Finetuning	Motion	Spatial	Temporal	Yes-No	Color	Object	Location	Number	Other
	✓	23.4	16.1	3.8	65.6	31.3	26.4	33.7	48.0	33.6
HowTo100M	✓	26.6	**17.7**	3.5	**67.5**	32.8	25.3	34.0	**50.5**	35.8
HowToVQA69M	✗	2.3	1.1	0.3	36.3	11.3	4.1	6.5	0.2	4.7
HowToVQA69M	✓	**28.0**	17.5	**4.9**	66.3	**34.3**	**26.7**	**35.8**	50.2	**36.8**

Table 5: Effect of our pretraining per question type on ActivityNet-QA.

Method	iVQA	MSRVTT QA	MSVD QA	ActivityNet QA	How2QA
QA-T	14.1	32.8	32.6	30.4	76.6
VQA-T	23.0	39.6	41.2	36.8	80.8

Table 6: Comparison of *QA-T* and *VQA-T* models trained from scratch (without pretraining) on downstream datasets.

ative per (positive) video-text pair, and use a binary cross-entropy loss. The cross-modal matching module is used to perform zero-shot VideoQA for the variant *VQA-T* trained on HowTo100M, by computing scores for $f(v, [q, a])$ for all possible answers a, for each video-question pair (v, q). We aggregate adjacent clips from HowTo100M to have at least 10 second clips and at least 10 narration words.

E. Additional experiments

E.1. Comparison to cross-dataset transfer

We define cross-dataset transfer as a procedure where we pretrain our VideoQA model on a VideoQA dataset and then finetune and test it on another VideoQA dataset. The training follows the procedure described for finetuning in the main paper (Section 4.2). We report results for cross-dataset transfer in Table 2. Note that we do not use MSVD-QA as downstream dataset as its test set has been automatically generated with the same method [2] as MSRVTT-QA. As can be observed, our approach with pretraining on HowToVQA69M significantly outperforms cross-dataset transfer models using the previously largest VideoQA dataset (MSRVTT-QA), or the largest manually annotated VideoQA dataset (ActivityNet-QA), both for the zero-shot and finetuning settings, on all four downstream datasets. We emphasize that our dataset is generated relying on text-only annotations, while MSRVTT-QA was generated using manually annotated video descriptions and ActivityNet-QA was manually collected. These results further demonstrate the benefits of our HowToVQA69M dataset.

E.2. Results for rare answers and per question type

Results for different answers frequencies are presented for the iVQA dataset in the main paper (Section 6.4). Here, we show results for MSRVTT-QA, MSVD-QA and ActivityNet-QA datasets in Table 3. As for iVQA, we observe that our model pretrained on our HowToVQA69M dataset, after finetuning, shows the best results for quartiles corresponding to rare answers (Q3 and Q4), notably in comparison with the model trained from scratch or the model pretrained on HowTo100M. We also find that our pretrained model, in the zero-shot setting, performs similarly across the different quartiles, with the exception of ActivityNet-QA, which includes in its most common answers *yes*, *no*. Note that in order to have a consistent evaluation with other experiments, we keep the same train vocabulary at test time. This implies that a significant part of answers in the test set is considered wrong because the answer is not in the vocabulary. This represents 16% of answers in iVQA, 3% of answers in MSRVTT-QA, 6% for MSVD-QA and 19% for ActivityNet-QA. Note, however, that our joint embedding framework could allow for different vocabularies to be used at the training and test time.

We also present results per question type for MSRVTT-QA, MSVD-QA and ActivityNet-QA in Tables 4 and 5. Compared to the model trained from scratch or the model pretrained on HowTo100M, we observe consistent improvements for most categories.

E.3. Comparison between *QA-T* and *VQA-T* on different datasets.

We show in Table 6 that *QA-T* is a strong baseline compared to *VQA-T* on existing VideoQA datasets, when both are trained from scratch. However, on iVQA, *VQA-T* improves more over *QA-T* than in other datasets, as measured by absolute improvement in top-1 accuracy. This suggests that the visual modality is more important in iVQA than in other VideoQA datasets.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[2] Michael Heilman and Noah A Smith. Good question! Statistical ranking for question generation. In *ACL*, 2010. 1, 2, 8

[3] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. 5

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

[5] 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. 7

[6] 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. 4, 5

[7] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014. 5

[8] 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. 4