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

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

Paper: https://arxiv.org/abs/2012.00451

Computer Vision - Video Understanding

- **Goal:** automating tasks that the human visual system can do
- **Motivation:** video data is plentiful
- **Applications:**
- Video search engines
- Video surveillance
- Self-driving cars
- Describing videos for visually impaired people
- Generation of video content for entertainment
- **Challenge:** How to evaluate video understanding?

Video Question Answering (VideoQA)

VideoQA is a promising proxy task to evaluate video understanding

Open-Ended Question:
Where are the men?

Answer: **Track**

Multiple-Choice Question:
What are the lined up men doing?

Proposal 1: **Running**

Proposal 2: Talking

Proposal 3: Shaving

VideoQA Challenges

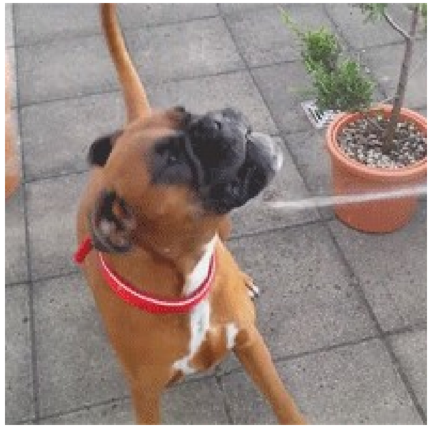
- VideoQA is a difficult task because of the diversity of questions that one may ask, requiring the ability to recognize actions, objects, colors at different spatio-temporal granularities
- Learning is currently the only known approach to handle variability in the data, but it requires lots of training data, and obtaining manually annotated VideoQA data is expensive and not scalable

Question: How many times does the cat lick?

Answer: **7 times**

Question: What does the cat do 3 times?

Answer: **put head down**

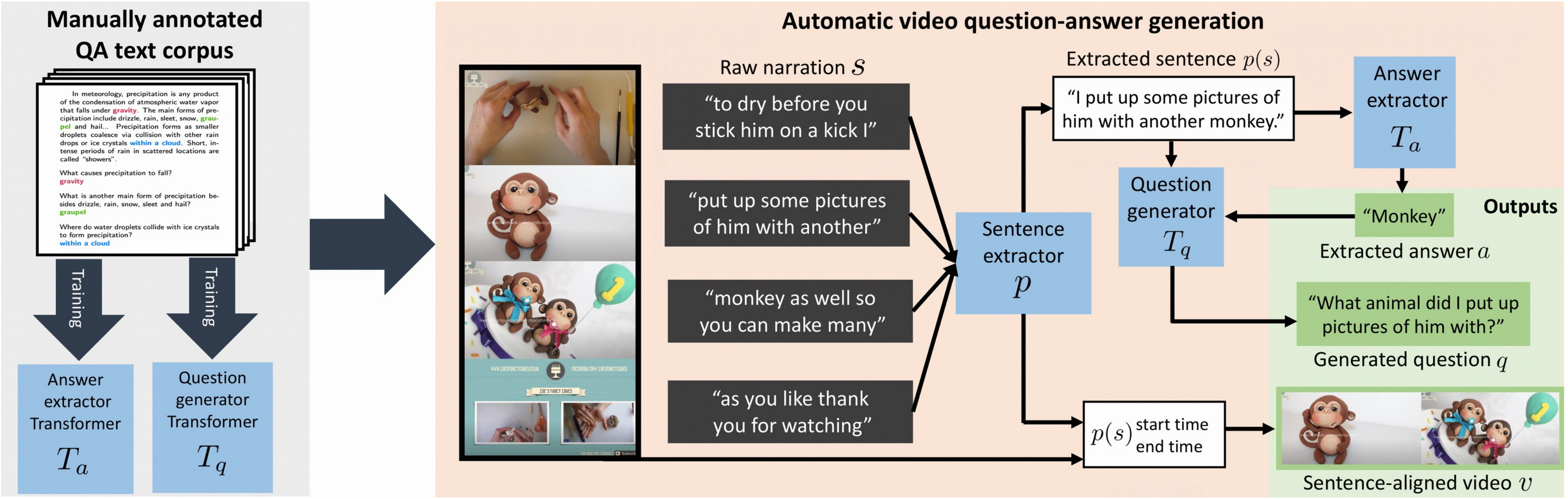


Question: What is the color of the bulldog?

Answer: **brown**

Just Ask: Method overview

- We automatically generate large-scale VideoQA data from narrated videos, relying on language models trained on text-only annotations
- We show how VideoQA models can benefit from such data, by tackling VideoQA without any manual supervision of visual data (*zero-shot*) or by finetuning our pretrained model

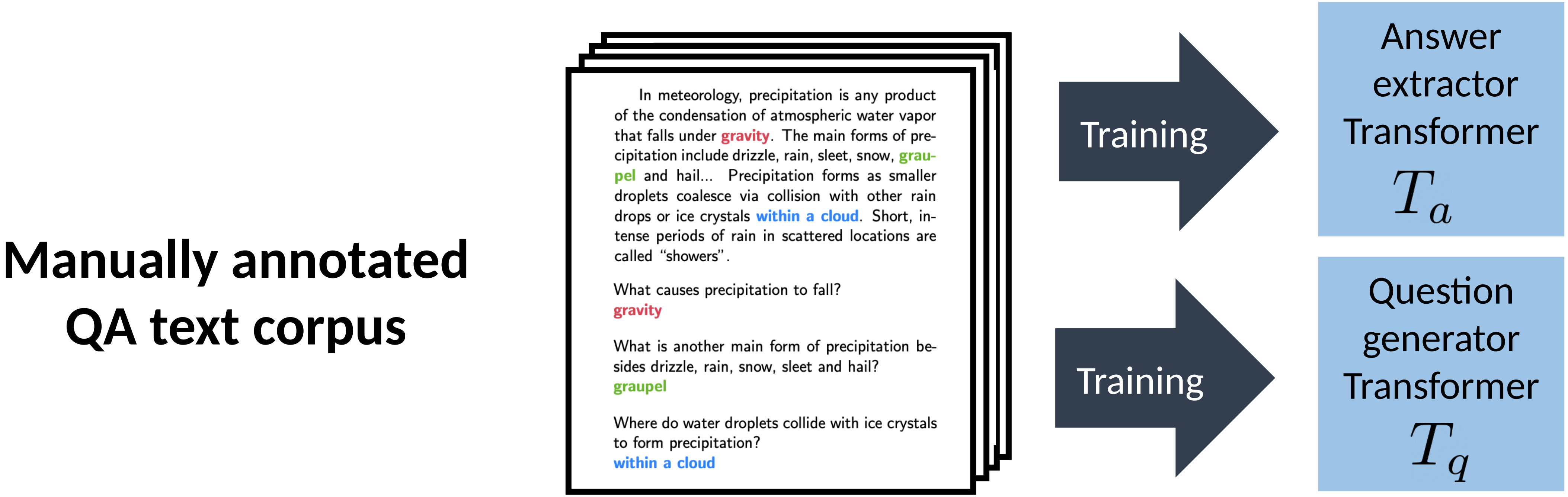


Weak supervision

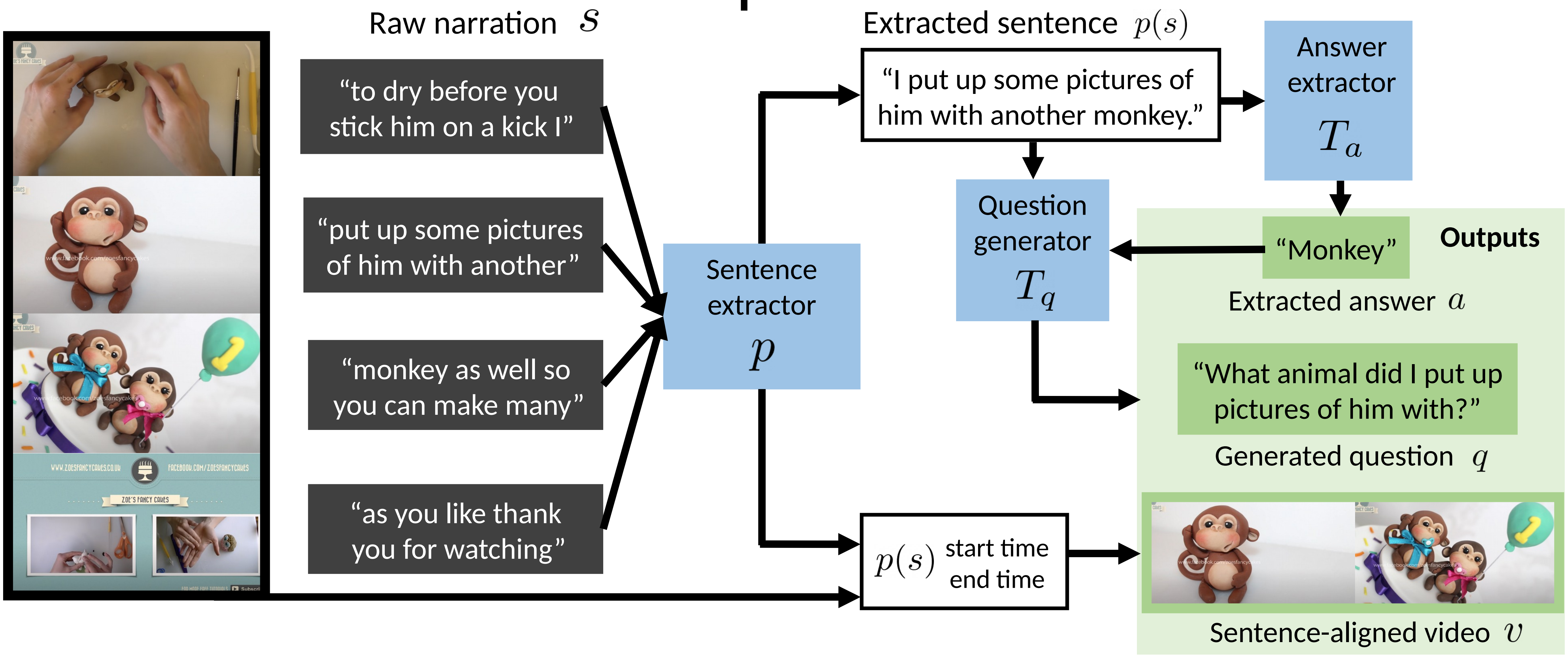
- Narrated videos contain speech, therefore paired (video, speech) data is easy to obtain and abundant.
- The weak correlation between the visual content and speech in narrated videos helped improve on other tasks [Miech 2019]

Text-only supervision for automatic generation of VideoQA data

To generate VideoQA data, we rely on cross-modal supervision and language models [Raffel 2020] trained on text-only annotations



Generating video-question-answer triplets



HowToVQA69M: a large-scale VideoQA training dataset

We apply our generation pipeline to the videos from HowTo100M [Miech 2019] and obtain HowToVQA69M, a large-scale and noisy VideoQA dataset

Speech: So you bring it to a point and we'll, just cut it off at the bottom.
Generated question: What do we do at the bottom?
Generated answer: cut it off

✓

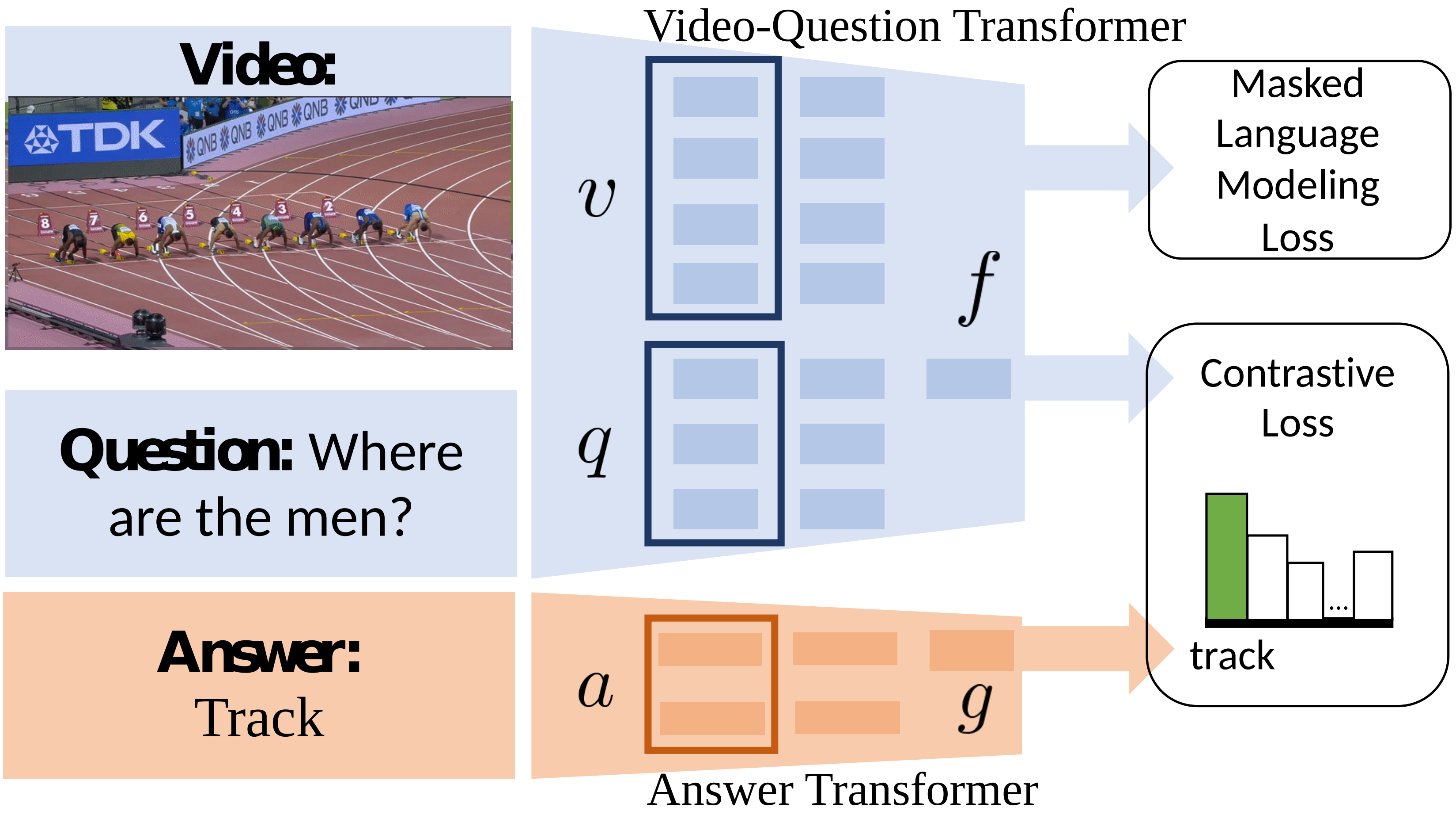
Speech: Do it on the other side, and you've peeled your orange.
Generated question: What color did you peel on the other side?
Generated answer: orange

Wrong QA Generation

Speech: You can't miss this…
Generated question: What can't you do?
Generated answer: miss

Weak video-speech correlation

VideoQA model and training procedure on HowToVQA69M



iVQA: a new VideoQA evaluation benchmark

- We manually collected an open-ended VideoQA dataset based on HowTo100M narrated videos
- It contains 10K videos, each annotated with 1 question and 5 corresponding correct answers

Question: What shape is the handcraft item in the end?

Answers	shell	✓	2 annotators
	spiral	✓	2 annotators
	heart	✓	1 annotator

Zero-shot VideoQA with *no manual supervision of visual data*

We evaluate our VideoQA model VQA-T pretrained on HowToVQA69M with the following baselines:

- QA-T pretrained on HowToVQA69M: language-only variant, not using the visual modality
- VQA-T pretrained on HowTo100M: common pretraining approach for multi-modal transformers

Quantitative results on 5 VideoQA datasets:

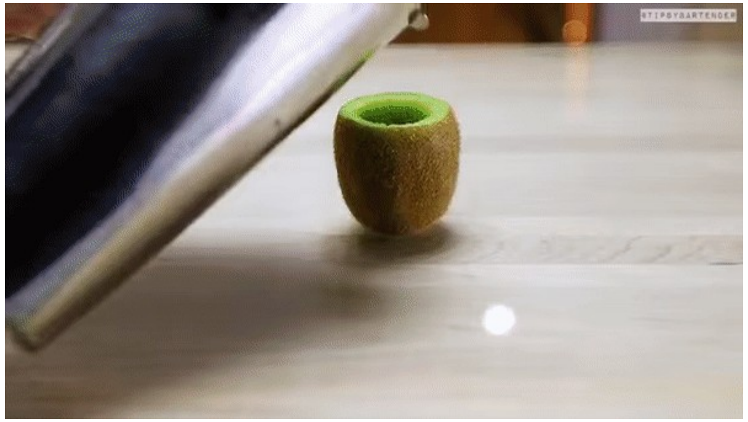
Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
Random	∅	0.09	0.02	0.05	0.05	25.0
QA-T	HowToVQA69M	4.4	2.5	4.8	11.6	38.4
VQA-T	HowTo100M	1.9	0.3	1.4	0.3	46.2
VQA-T	HowToVQA69M	**12.2**	**2.9**	**7.5**	**12.9**	**51.1**

Zero-shot VideoQA with *no manual supervision of visual data*

Qualitative examples on iVQA:

Question: What is the man cutting?
GT answer: pipe
QA-T (HowToVQA69M): onion
VQA-T (HowTo100M): knife holder
Ours: pipe

Question: What is the largest object at the right of the man?
GT answer: wheelbarrow
QA-T (HowToVQA69M): statue
VQA-T (HowTo100M): trowel
Ours: wheelbarrow



Question: What fruit is shown in the end?
GT answer: watermelon
QA-T (HowToVQA69M): pineapple
VQA-T (HowTo100M): slotted spoon
Ours: watermelon

Zero-shot VideoQA: failure cases

Qualitative examples on iVQA:

Question: What are standing up behind the man on his right?
GT answer: guitars
Ours: strings

Question: In what room does the video take place?
GT answer: kitchen
Ours: dining room

Benefits of HowToVQA69M pretraining

Comparison with state-of-the-art on 4 VideoQA datasets:

Method	Pretraining Data	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HCRN [Le 2020]	∅	35.6	36.1	-	-
SSML [Amrani 2020]	HowTo100M	35.1	35.1	-	-
HERO [Li 2020]	HowTo100M	-	-	-	74.1
ClipBERT [Lei 2021]	COCO + VG	37.4	-	-	-
CoMVT [Seo 2021]	HowTo100M	39.5	42.6	38.8	82.3
Ours (∅)	∅	39.6	41.2	36.8	80.8
Ours	HowToVQA69M	**41.5**	**46.3**	**38.9**	**84.4**

Results for rare answers

Results on subsets of iVQA corresponding to four quartiles with Q1 and Q4 corresponding to samples with most frequent and least frequent answers:

Pretraining Data	Finetuning	Q1	Q2	Q3	Q4
∅	✓	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
HowToVQA69M	✓	**47.9**	**28.1**	**15.6**	**8.5**

=> VideoQA specific pretraining on additional large-scale, diverse data helps improve generalization

Open research directions

- Reduce the domain gap between the question-answer generator trained on text-only data (SQuADv1 in our case) and the text data used for VideoQA generation (speech in our case)
- Automatic cleaning of generated data
- Generalization of VideoQA models to other tasks
- Creation of VideoQA datasets that are closer to potential applications
- End-to-end learning of VideoQA models

Conclusion

- We automatically generate a large-scale VideoQA dataset, HowToVQA69M, using text-only supervision and videos with readily-available narration
- We manually collect iVQA, a new VideoQA benchmark with redundant annotations and reduced language bias
- We show that our VideoQA model highly benefits from training on HowToVQA69M in a new zero-shot VideoQA setting; additionally, after finetuning, our model improves the state-of-the-art on 4 VideoQA datasets