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

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

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

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

[1] Inria Paris / ENS   [2] CIIRC CTU   [+] Now at DeepMind

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

- **Data variability:** VideoQA requires the ability to recognize actions, objects, colors at different spatio-temporal granularities
- **Annotation:** 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?

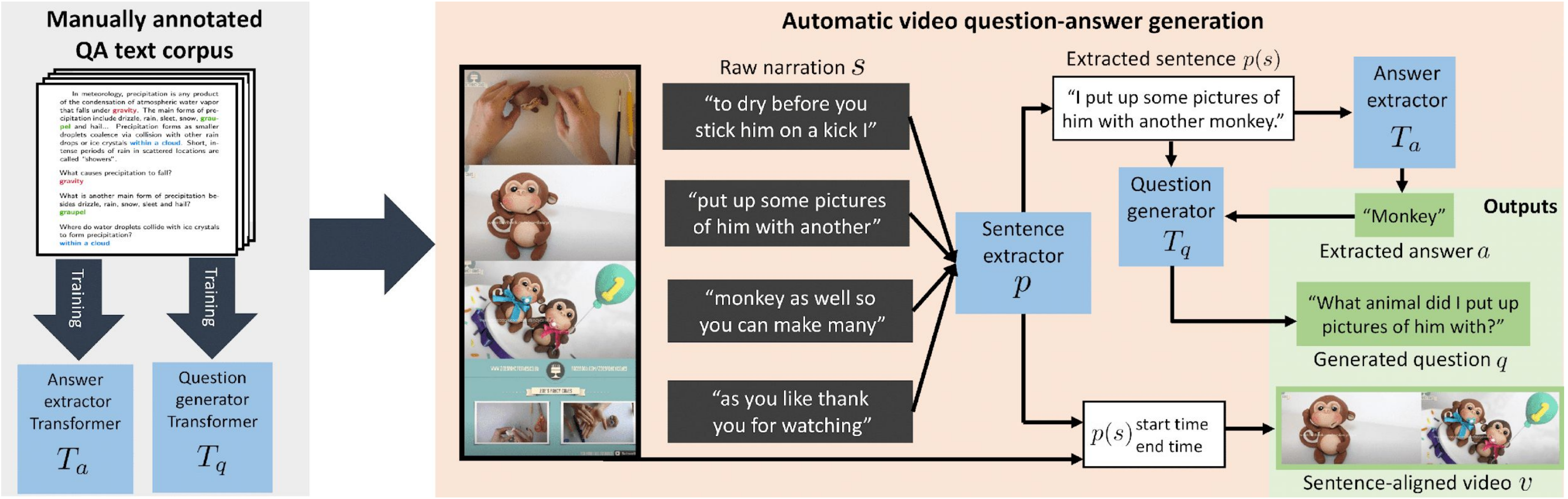
Answer: **put head down**

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

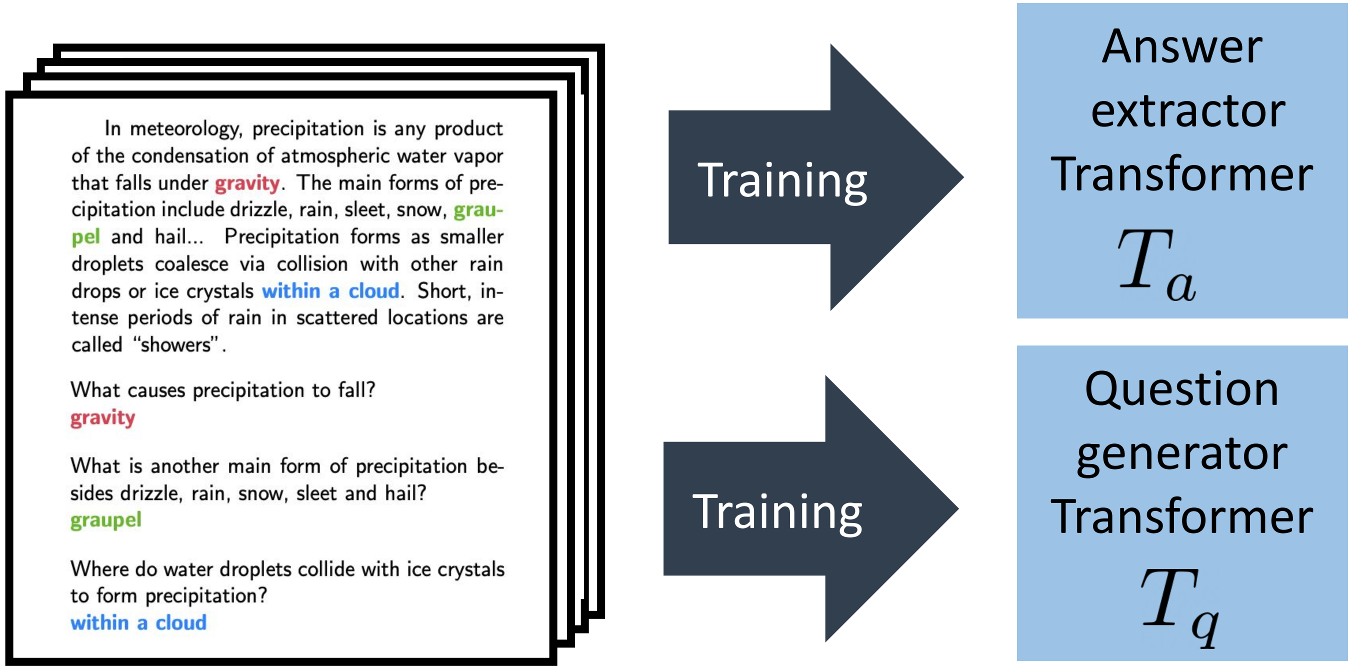
- Narrated videos contain speech, therefore paired (video, speech) data is easy to obtain and abundant
- The weak correlation between the visual content and speech in narrated videos helped improve on other tasks [Miech 2019]

[Miech 2019] HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, A. Miech et al

Text-only supervision for automatic generation of VideoQA data

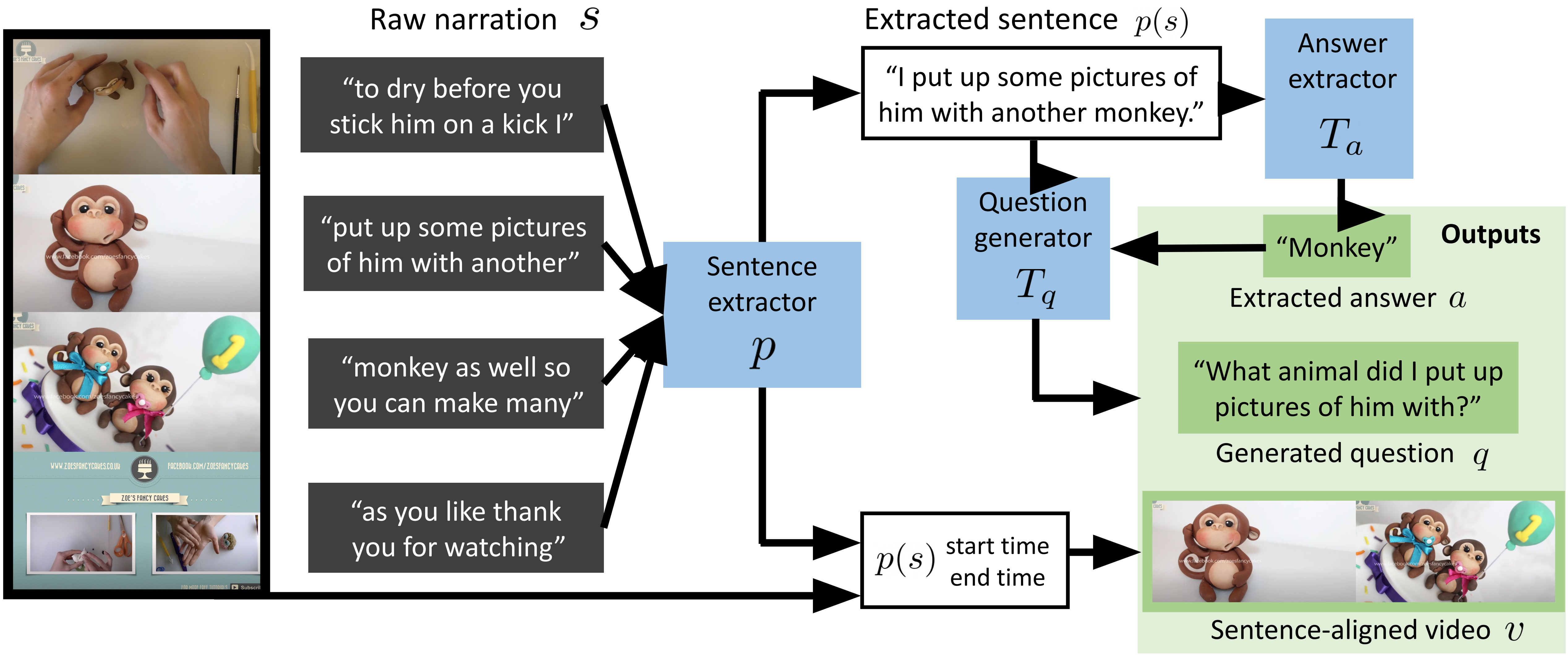
To generate VideoQA data, we rely on language models [Raffel 2020] trained on text-only annotations

Manually annotated QA text corpus



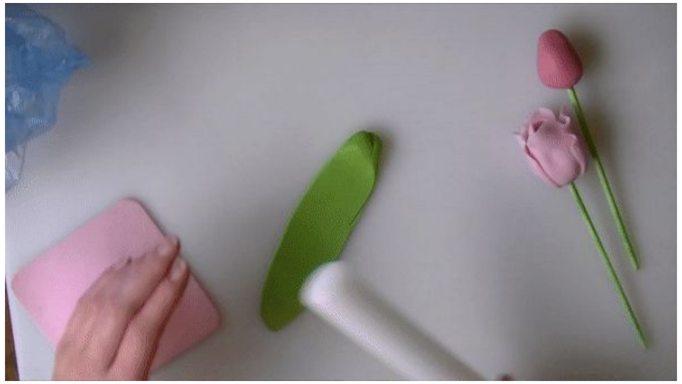
[Raffel 2020] Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, C. Raffel et al

Generating video-question-answer triplets



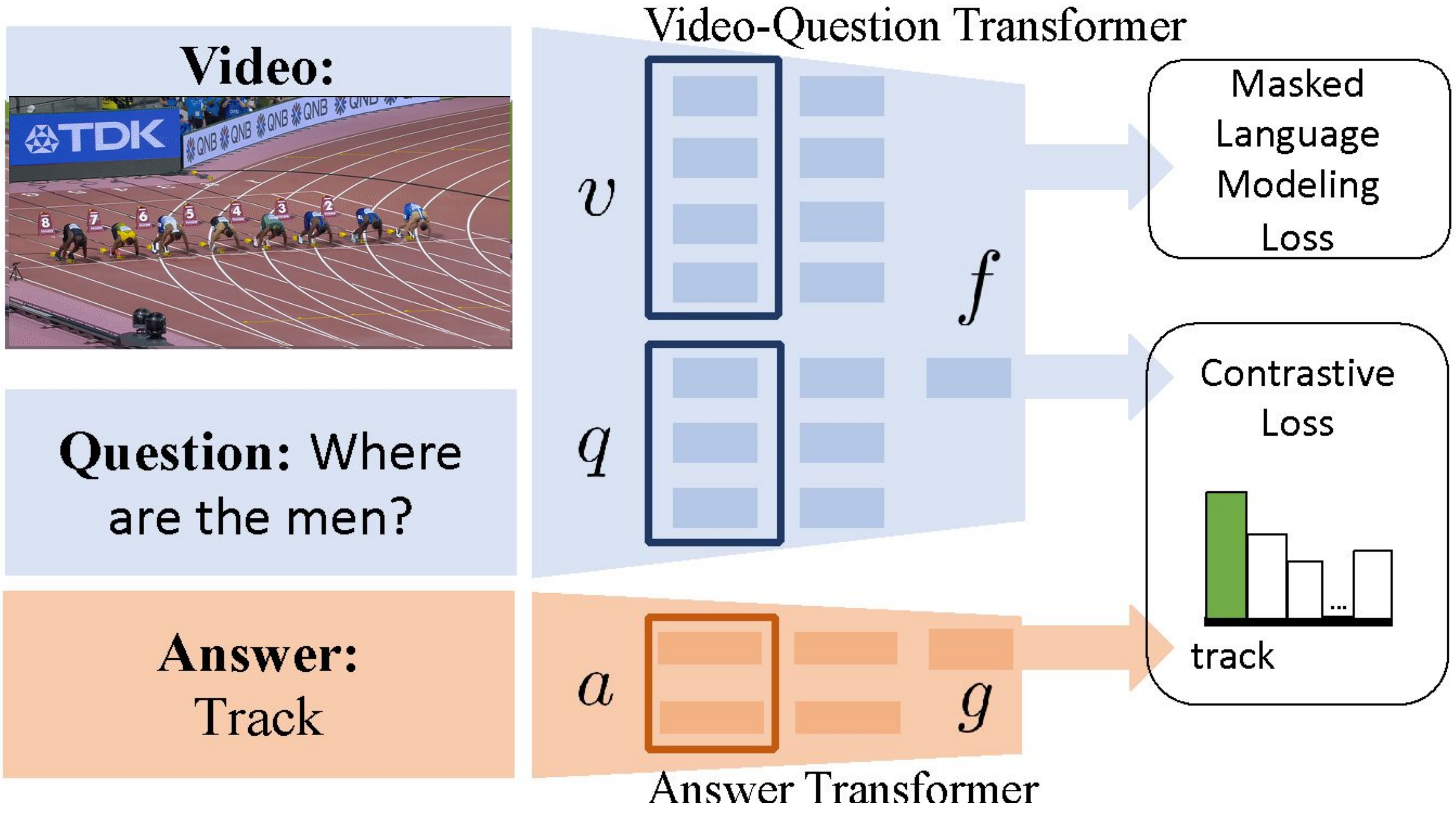
HowToVQA69M: a large-scale VideoQA training dataset

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



Input Speech:	So you bring it to a point and we'll, just cut it off at the bottom.	Do it on the other side, and you've peeled your orange.	You can't miss this…
Generated outputs:	**Question:** What do we do at the bottom? **Answer:** cut it off	**Question:** What color did you peel on the other side? **Answer:** orange	**Question:** What can't you do? **Answer:** miss
	✓	**Incorrect QA Generation**	**Weak video-speech correlation**

VideoQA model (*VQA-T*) 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

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**

Conclusion

- We automatically generate a large-scale VideoQA dataset, HowToVQA69M, using text-only supervision and videos with readily-available narration
- 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