2021 ICCV OCTOBER 11-17 VIRTUAL

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

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, Cordelia Schmid

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

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

Inria ENS | PSL CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

Video Question Answering (VideoQA)

VideoQA is a promising task to evaluate the ability to understand visual data.

Question: What fruit is shown at the end?

Answer: watermelon

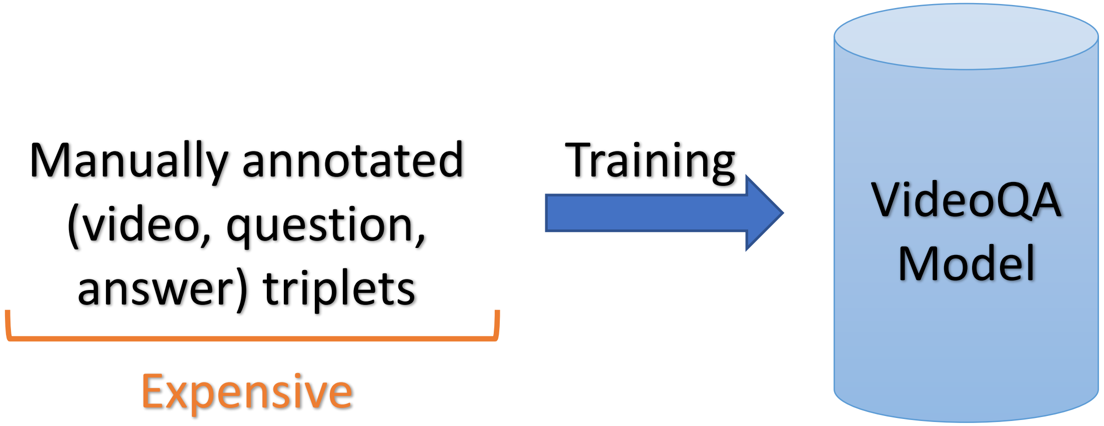
Question: What is the largest object at the right of the man?

Answer: wheelbarrow

Source of the examples: iVQA dataset, see Slide 10

Challenges in VideoQA

- Large diversity of questions and videos
- Manual annotation for VideoQA is expensive
- **Problematic:** How to tackle VideoQA with the least amount of manual supervision possible?



Just Ask idea

- Automatically generate VideoQA training data from narrated videos.
- Rely on text-only annotations and cross-modal supervision.

Speech: The sound is amazing on this piano.

Generated question: What kind of instrument is the sound of?
Generated answer: piano

Weak supervision in narrated videos

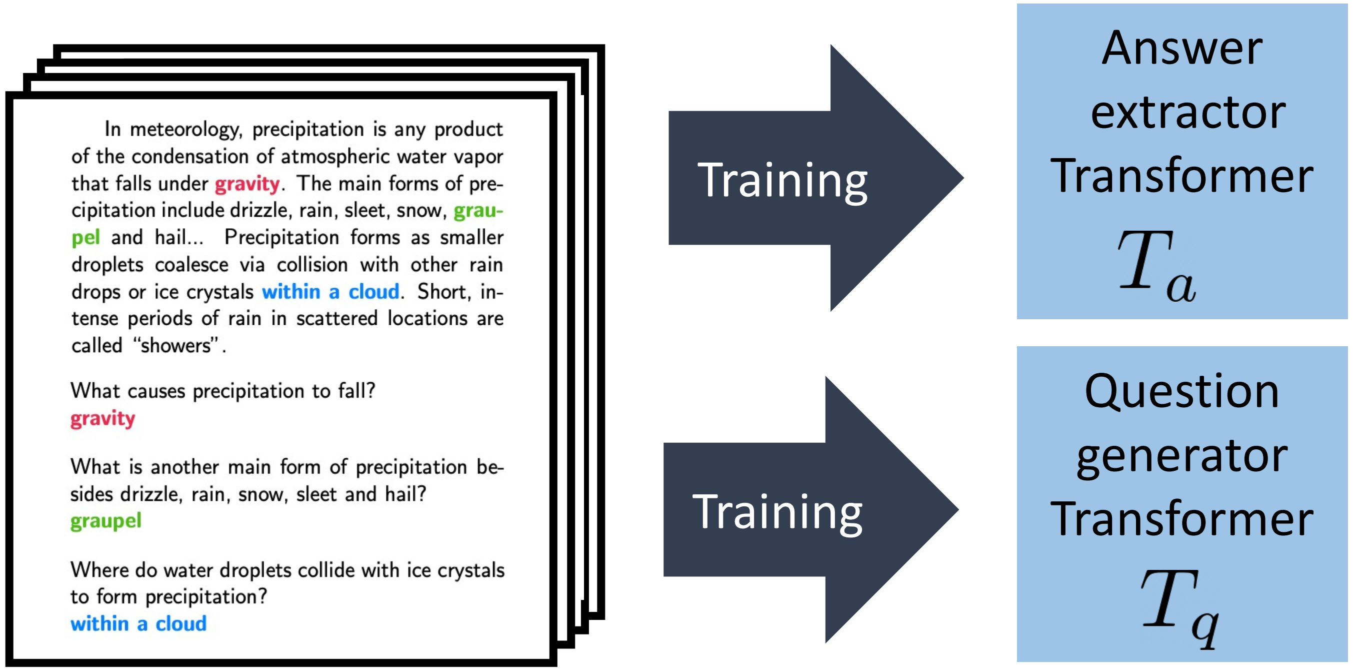
- Narrated videos are easy to obtain at scale.
- **Assumption:** weak correlation between the visual content and the speech [Miech 2019]

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

Text-only supervision

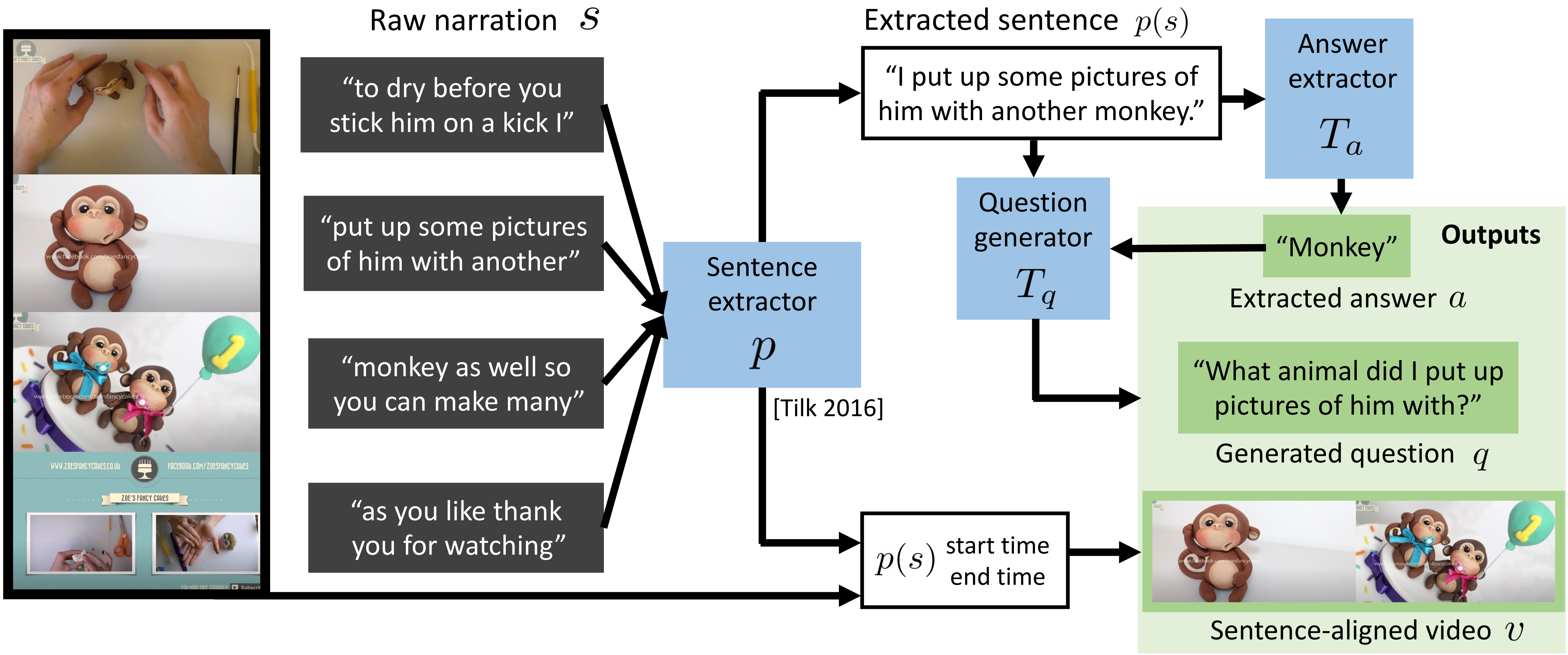
We use language models trained on a text-only question-answering corpus [Raffel 2020, Suraj 2020, Rajpurkar 2016].

Manually annotated QA text corpus



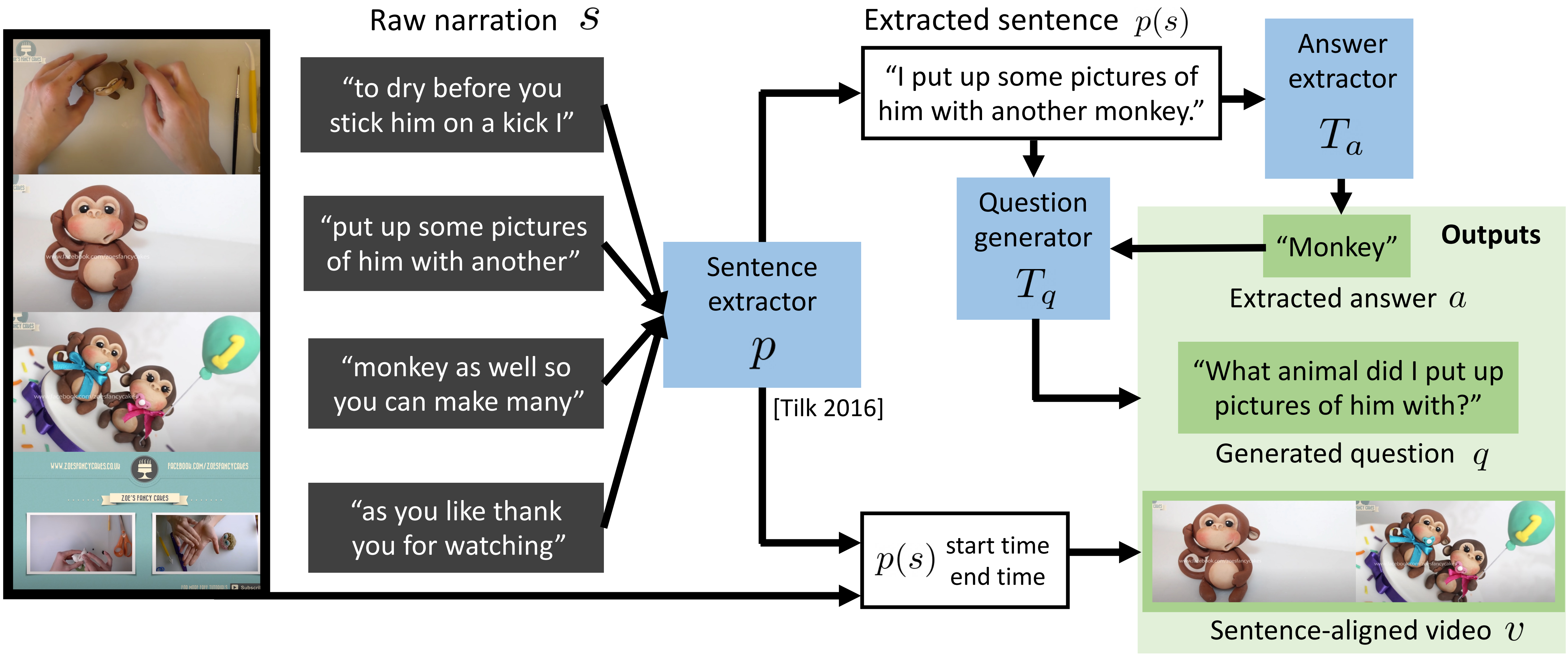
[Raffel 2020] Exploring the limits of transfer learning with a unified text-to-text transformer, Raffel et al, JMLR 2020.
[Suraj 2020] Question Generation, Suraj, GitHub repository 2020.
[Rajpurkar 2016] SQuAD: 100,000+ questions for machine comprehension of text, Rajpurkar et al, arXiv 2016.

Generating VideoQA data



[Tilk 2016] Bidirectional recurrent neural network with attention mechanism for punctuation restoration, Tilk et al, Interspeech 2016.

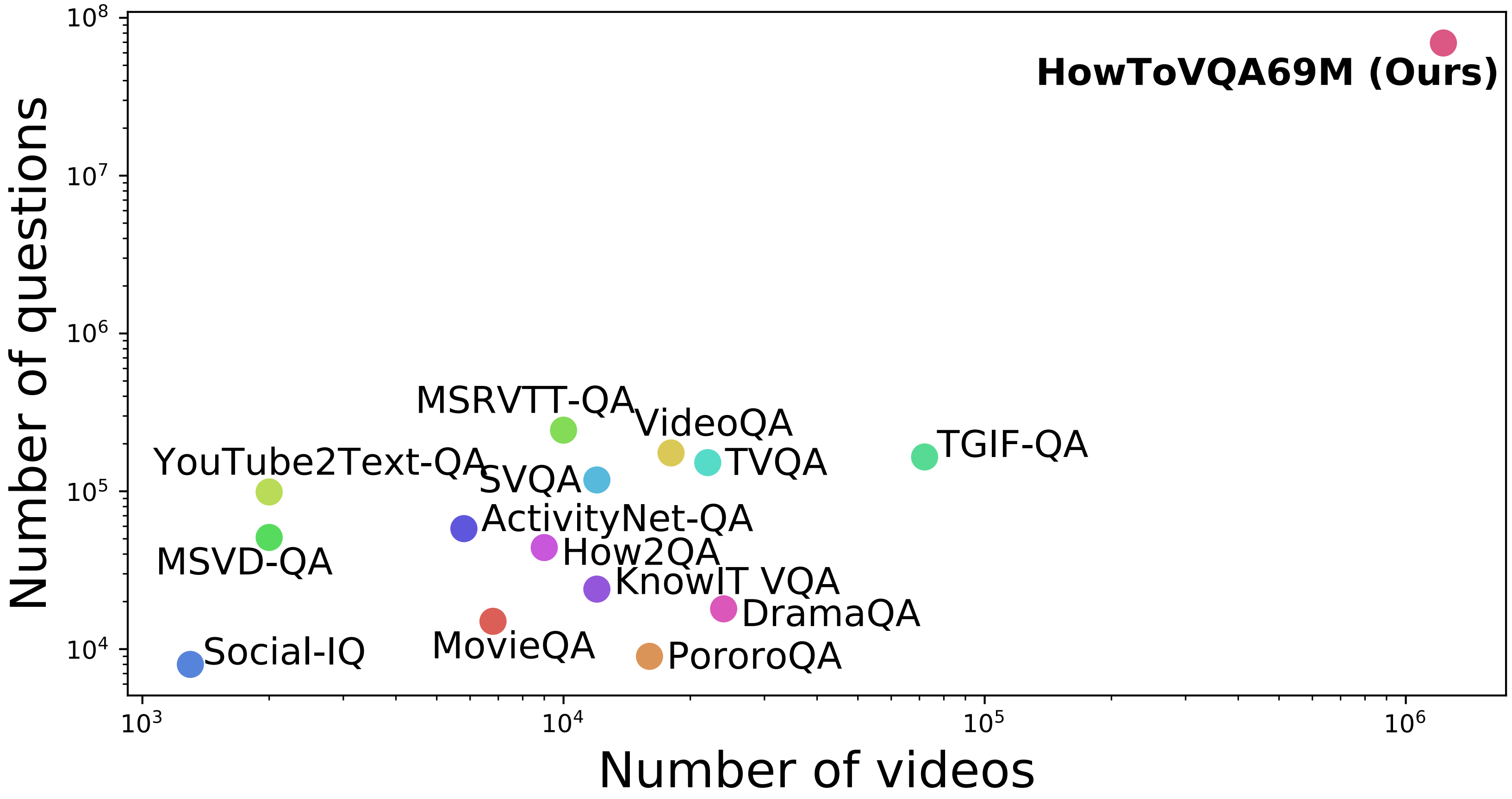
Generating VideoQA data



[Tilk 2016] Bidirectional recurrent neural network with attention mechanism for punctuation restoration, Tilk et al, Interspeech 2016.

HowToVQA69M: a large-scale VideoQA dataset

- Generated by applying our pipeline to HowTo100M [Miech 2019]
- 69M video-question-answer triplets



Noise in HowToVQA69M

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

≈ 30%

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

QA Generation error

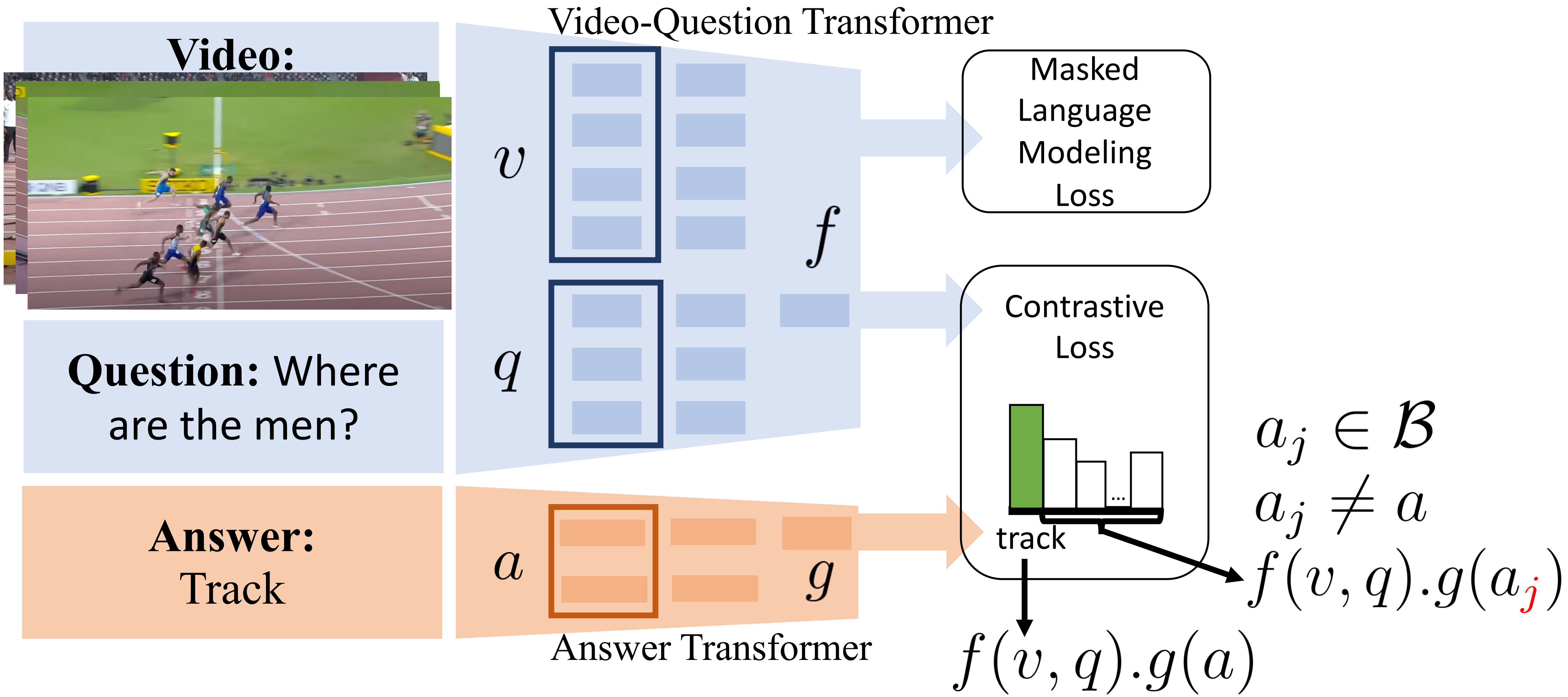
≈ 31%

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

QA unrelated to video

≈ 39%

VideoQA model and training procedure



iVQA: a new VideoQA benchmark

- 10K videos from HowTo100M
- Manually collected
- 10K open-ended questions
- 5 correct answers per question
- Exclusion of non-visual questions to reduce language bias

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

Answers

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

Zero-shot VideoQA: quantitative results

Task definition: no manual supervision of visual data

Our model (iii) outperforms:

- Its language-only variant (i) -> importance of multi-modality in HowToVQA69M
- Its variant trained on HowTo100M (ii) -> benefit of HowToVQA69M to train VideoQA models

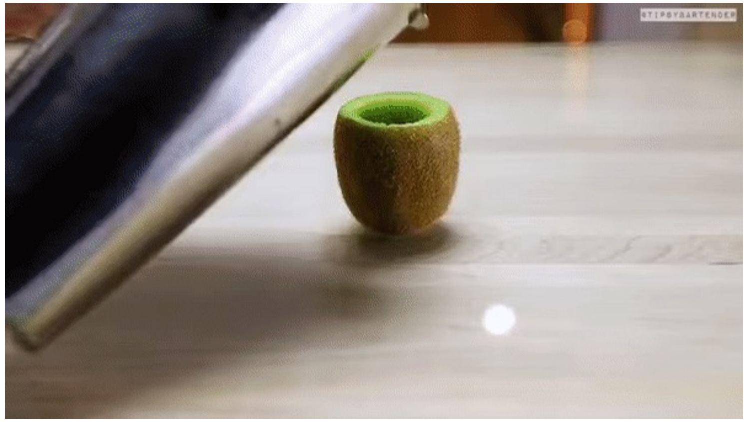
Quantitative results on 5 VideoQA datasets:

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

Zero-shot VideoQA: qualitative results

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

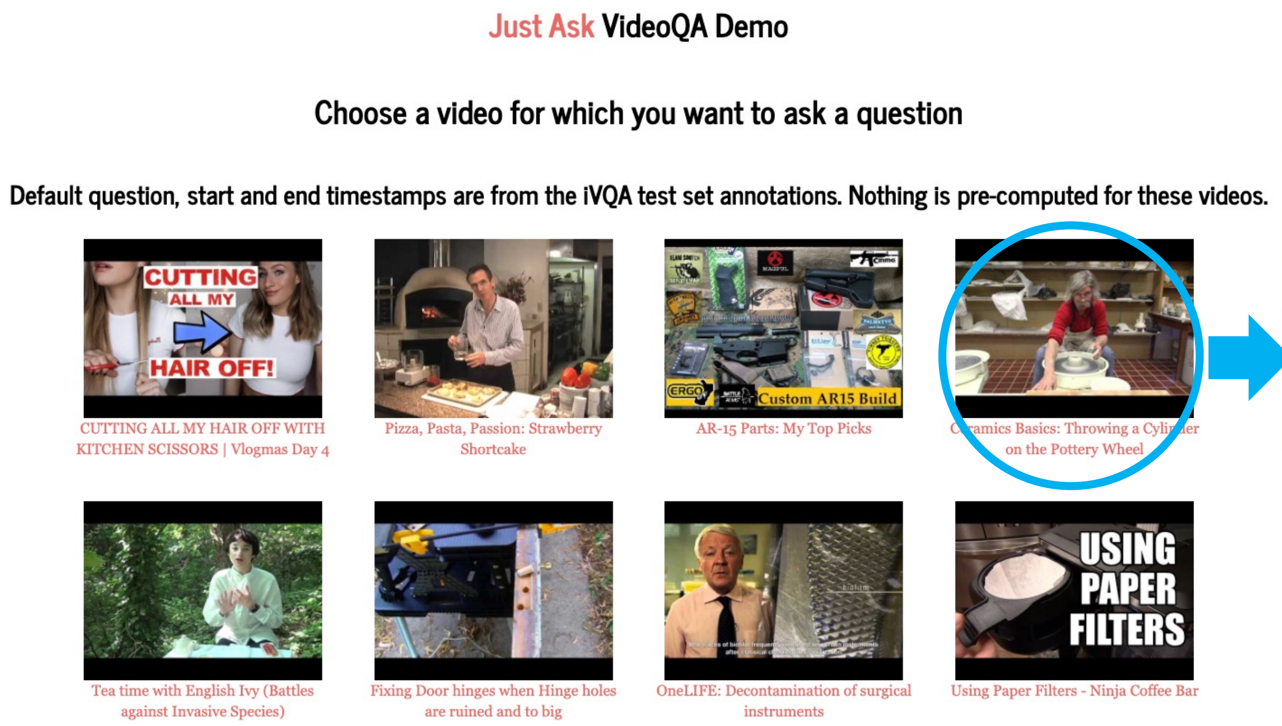
Question: What is the largest object at the right of the man?
GT answer: wheelbarrow
QA-T (HowToVQA69M): statue
VQA-T (HowTo100M): trowel
Ours: wheelbarrow



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

Source of the examples: iVQA dataset

Online Demo
http://videoqa.paris.inria.fr/



Results after finetuning

State-of-the-art results on 4 existing VideoQA datasets

Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HCRN [Le 2020]	∅	-	35.6	36.1	-	-
SSML [Amrani 2021]	HowTo100M	-	35.1	35.1	-	-
HERO [Li 2020]	HowTo100M + TV	-	-	-	-	74.1
ClipBERT [Lei 2021]	COCO + VG	-	37.4	-	-	-
CoMVT [Seo 2021]	HowTo100M	-	39.5	42.6	38.8	82.3
Ours (∅)	∅	23.0	39.6	41.2	36.8	80.8
Ours (HowTo100M)	HowTo100M	28.1	40.4	43.5	38.1	81.9
Ours	HowToVQA69M	**35.4**	**41.5**	**46.3**	**38.9**	**84.4**

[Le 2020] Hierarchical conditional relation networks for video question answering, Le et al, CVPR 2020.
[Amrani 2021] Noise estimation using density estimation for self-supervised multimodal learning, Amrani et al, AAAI 2021.
[Li 2020] HERO: Hierarchical encoder for video+language omni-representation pre-training, Li et al, EMNLP 2020.
[Lei 2021] Less is more: Clipbert for video-and-language learning via sparse sampling, Lei et al, CVPR 2021.
[Seo 2021] Look before you speak: Visually contextualized utterances, Seo et al, CVPR 2021.

Results for rare answers

- Training on a downstream VideoQA dataset -> large improvements for most frequent answers
- Pretraining on HowToVQA69M -> significant improvements *both* for common and rare answers

Results on subsets of iVQA:

Most frequent answers　　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**

Comparison of generation methods

- [Heilman 2010] was previously used to generate VideoQA data from video descriptions [Xu 2017].
- We compare against [Heilman 2010] by applying it in our case.
- Our generation leads to better downstream VideoQA results.

	Zero-Shot			Finetuning		
Generation method	iVQA	ActivityNet-QA	How2QA	iVQA	ActivityNet-QA	How2QA
[Heilman 2010]	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**

Speech: This is classic premium chicken, grilled sandwich.
[Heilman 2010]: What is classic premium chicken, grilled sandwich? this
Ours: What type of sandwich is this? classic premium chicken, grilled sandwich

[Heilman 2010] Good question! Statistical ranking for question generation, Heilman et al, ACL 2010.
[Xu 2017] Video question answering via gradually refined attention over appearance and motion, Xu et al, ACM 2017.

Ablation: pretraining losses

MLM	Sampling without answer repetition	Zero-Shot		Finetuning	
		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**

=> Best results with MLM and our contrastive loss

Ablation: scale

Pretraining data size	Zero-Shot		Finetuning	
	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**

=> Scale matters.

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. After finetuning, our model improves the state-of-the-art on 4 VideoQA datasets.