Zero-Shot Video Question Answering

Antoine YANG, Willow (Inria Paris and DI ENS)

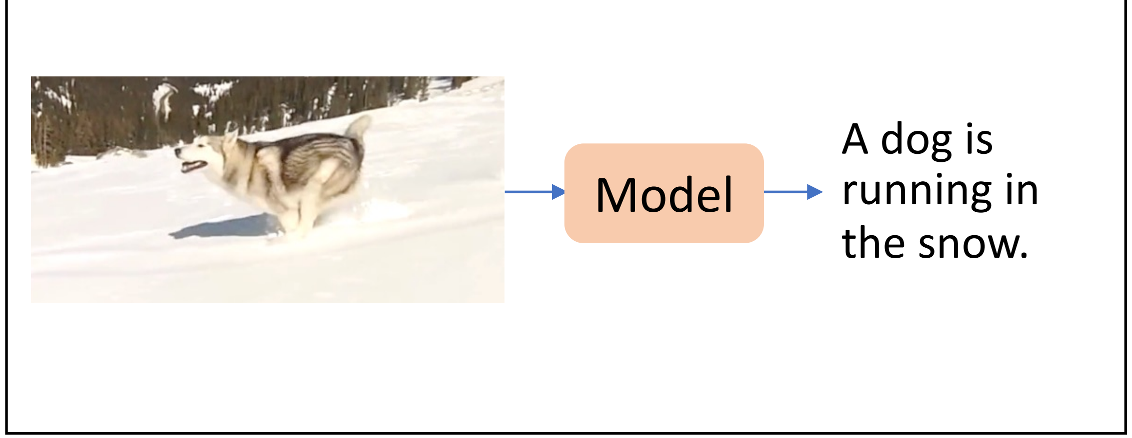
June 2022

Outline

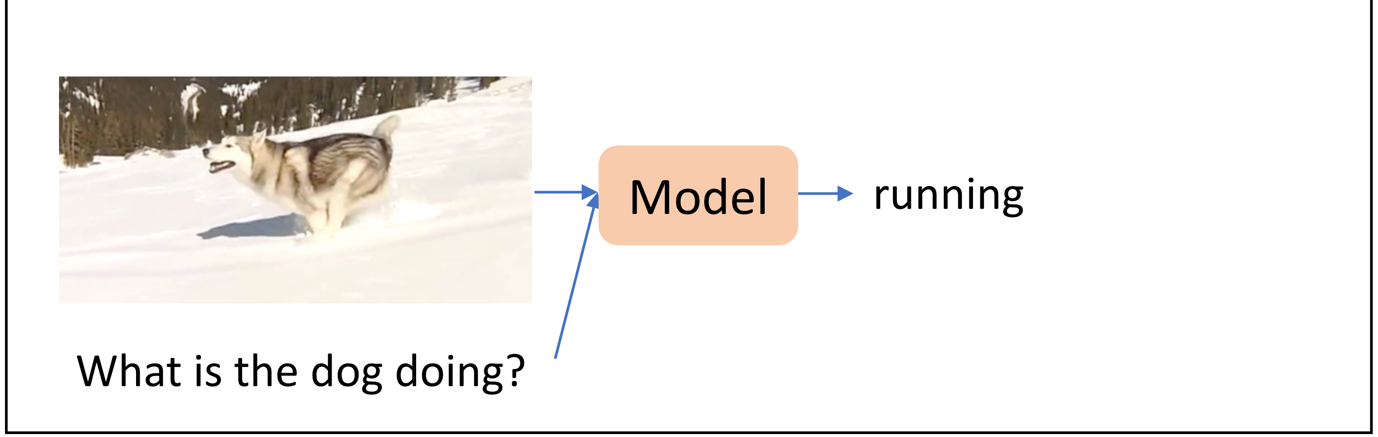
- **Background: Building AI systems that can see and talk**
- Tasks
- Neural architectures
- Training
- **Zero-shot video question answering**
- Just ask: learning to answer questions from narrated videos
- Zero-shot video question answering via frozen bidirectional language models

Tasks

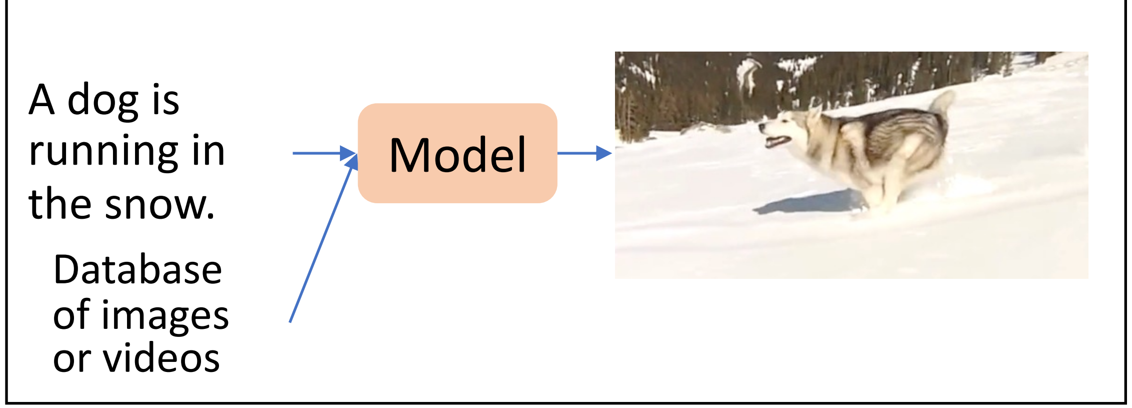
Visual Captioning



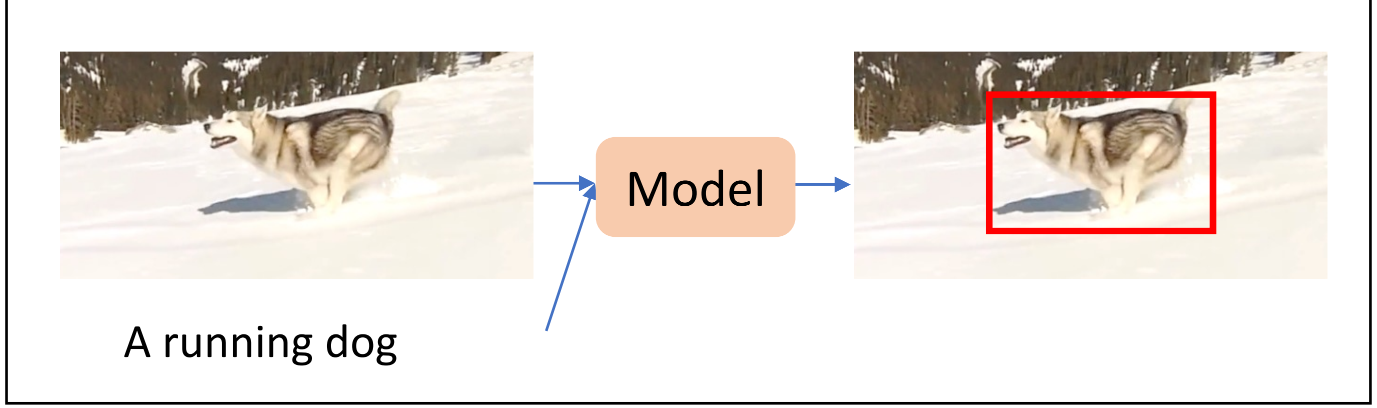
Visual Question Answering (VQA/VideoQA)



Text-to-visual retrieval



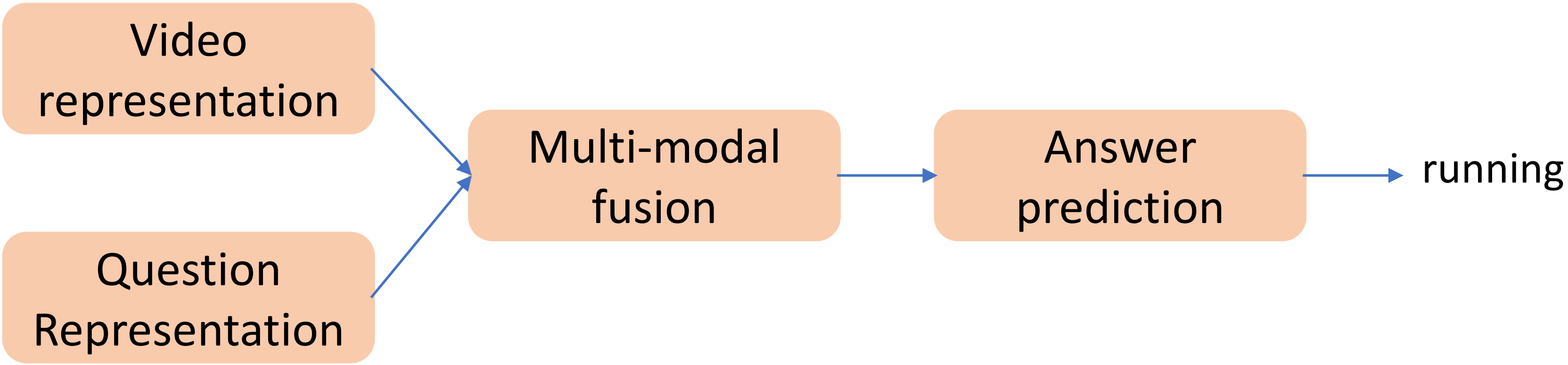
Visual Grounding



A typical neural architecture (VideoQA)

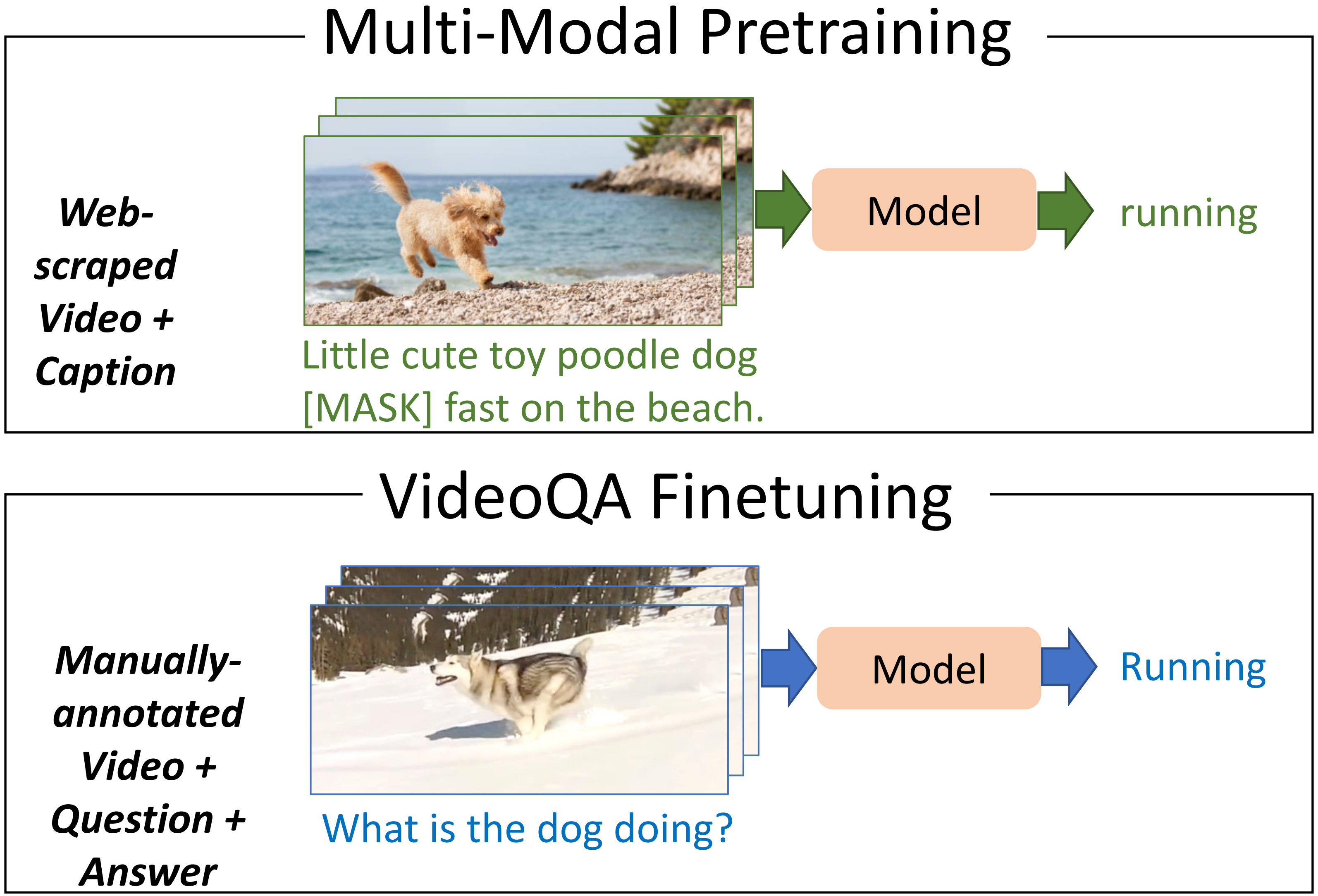
- Typical video representation: pretrained vision transformer [Dosovitskiy 2021]
- Typical question representation: pretrained BERT [Devlin 2019]
- Typical multi-modal fusion: transformer [Vaswani 2017]
- Typical answer prediction module: classifier

What is the dog doing?



[Dosovitskiy 2021] An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale , Dosovitskiy et al, ICLR 2021.
[Devlin 2019] Bert: Pre-training of deep bidirectional transformers for language understanding, Devlin et al, NAACL 2019.
[Vaswani 2017] Attention is all you need, Vaswani et al, NeurIPS 2017.

A typical training procedure (VideoQA)



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

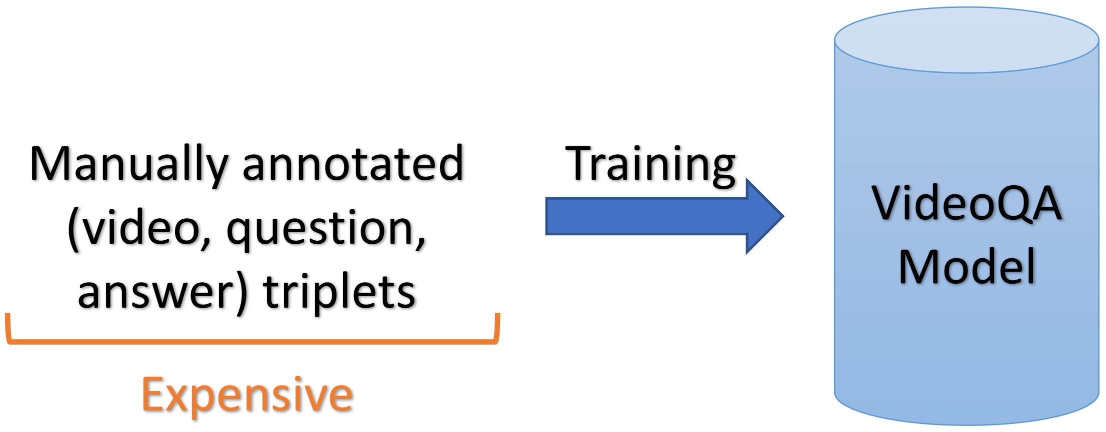
Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, Cordelia Schmid

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

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

Challenges

- SoTA approaches use manual supervision
- **Issues:** Manual annotation for VideoQA is expensive. Large diversity of questions and videos.
- **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.

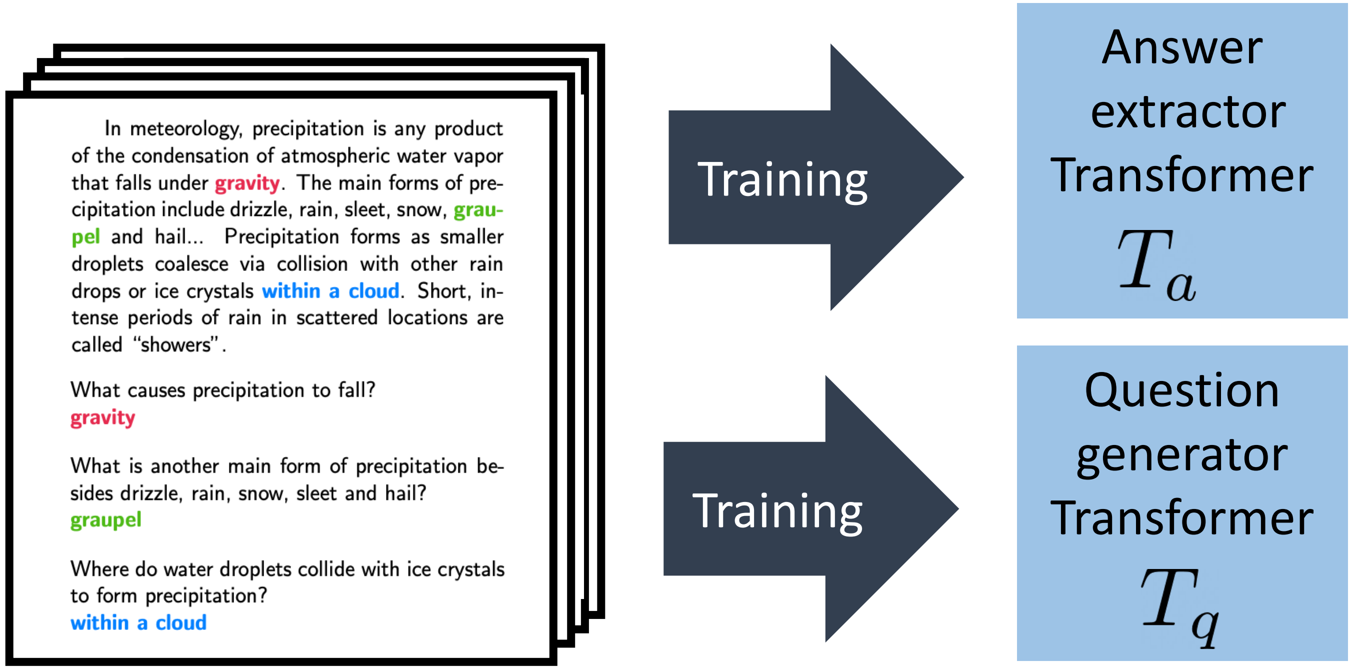
Speech: The sound is amazing on this piano.

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

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



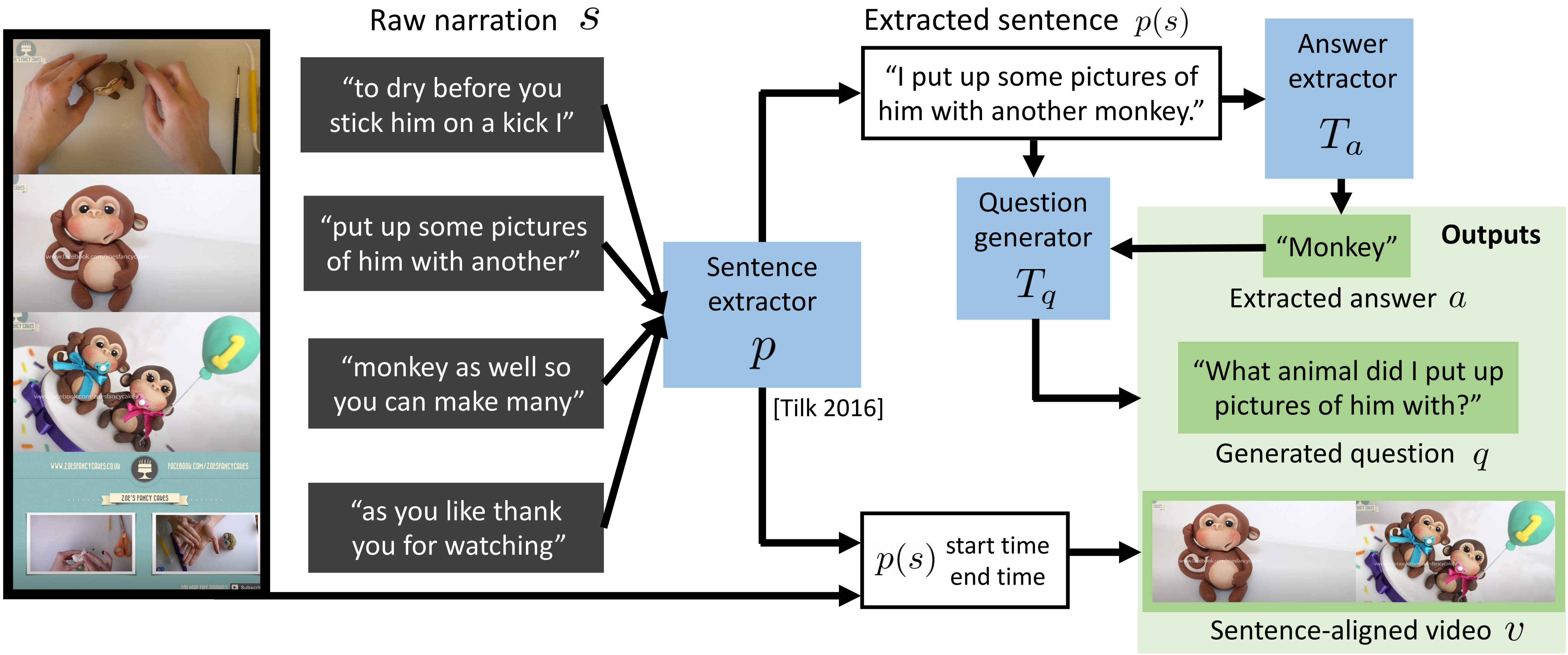
[Raffel 2020] Exploring the limits of transfer learning with a unified text-to-text transformer, Raffel et al, JMLR 2020.
[Suraj 2020] Question Generation, Suraj, GitHub repository 2020.
[Rajpurkar 2016] SQuAD: 100,000+ questions for machine comprehension of text, Rajpurkar et al, arXiv 2016.

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.

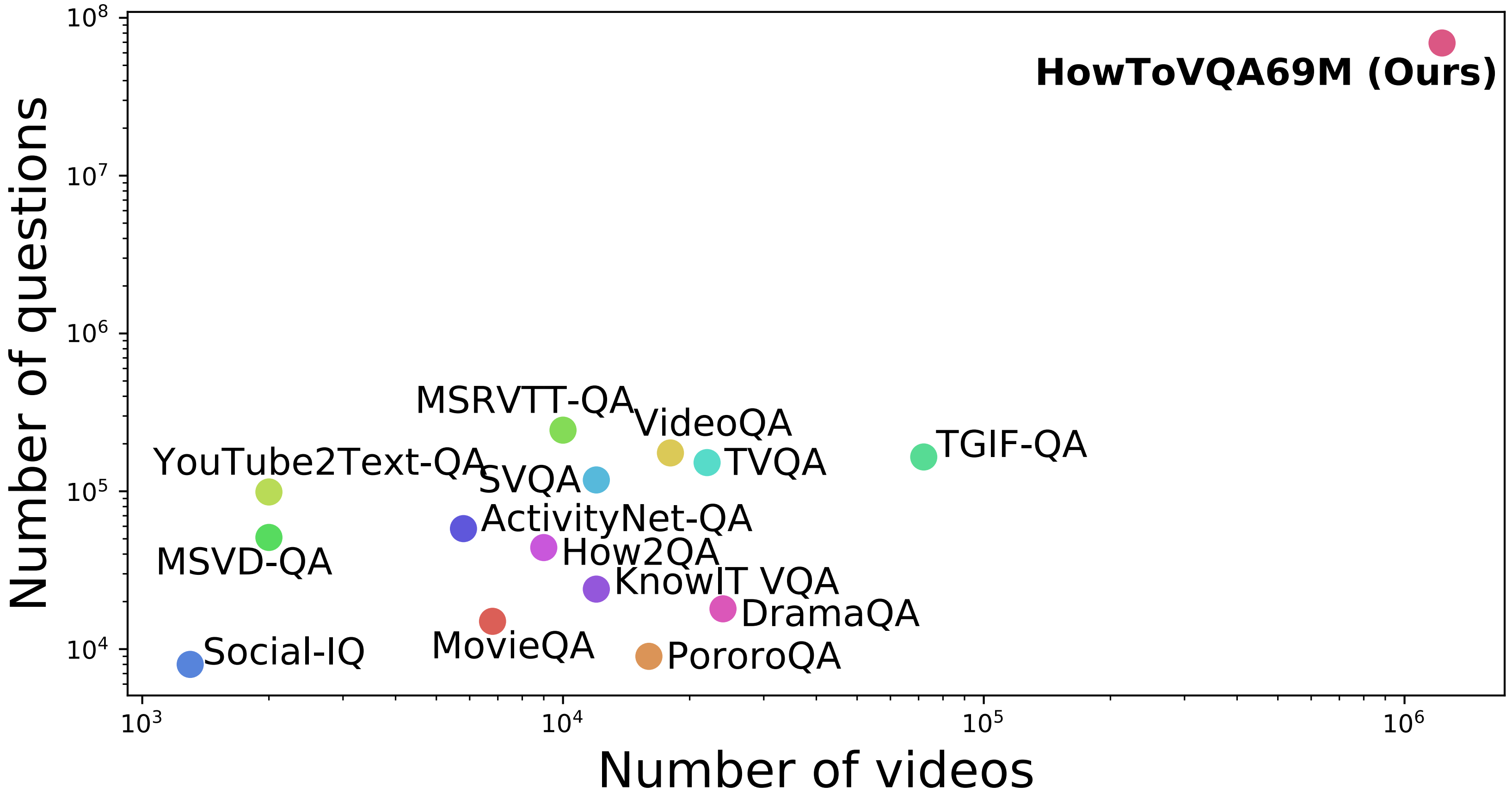
Generating VideoQA data



[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

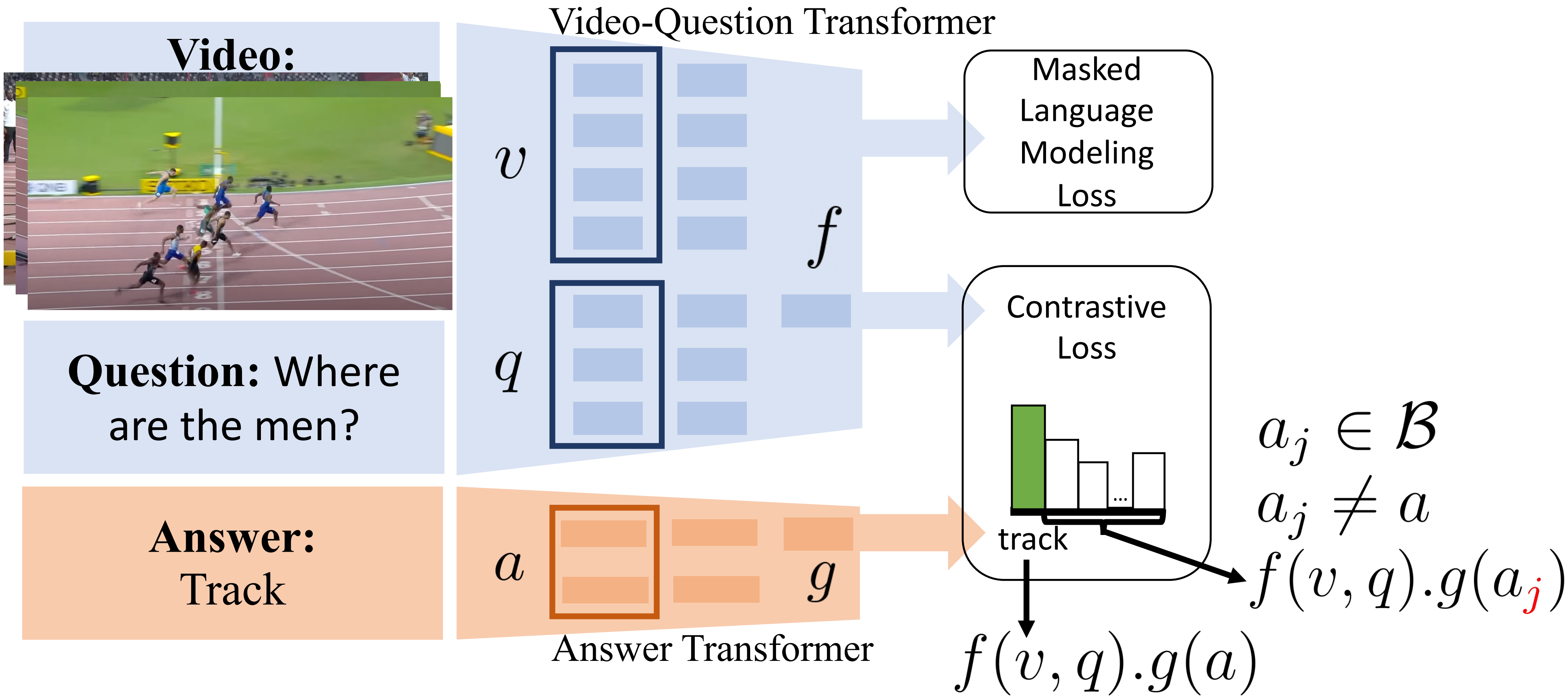
≈ 31%

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



Zero-shot VideoQA: quantitative results

Task definition: no manual supervision of visual data

- Importance of the visual modality
- Importance of generating video-question-answer triplets

Quantitative results on 5 VideoQA datasets:

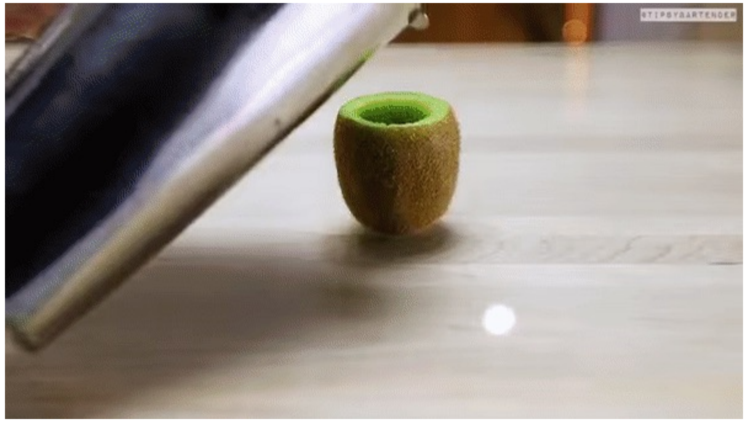
Method	Pretraining Data	iVQA		MSRVTT-QA		MSVD-QA		ActivityNet-QA		How2QA
		Top-1	Top-10	Top-1	Top-10	Top-1	Top-10	Top-1	Top-10	Top-1
Random	∅	0.09	0.9	0.02	0.2	0.05	0.5	0.05	0.5	25.0
QA-T	HowToVQA69M	4.4	23.2	2.5	6.5	4.8	15.0	11.6	45.8	38.4
VQA-T	HowTo100M	1.9	11.9	0.3	3.4	1.4	10.4	0.3	1.9	46.2
VQA-T (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.

Zero-shot VideoQA: qualitative results

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

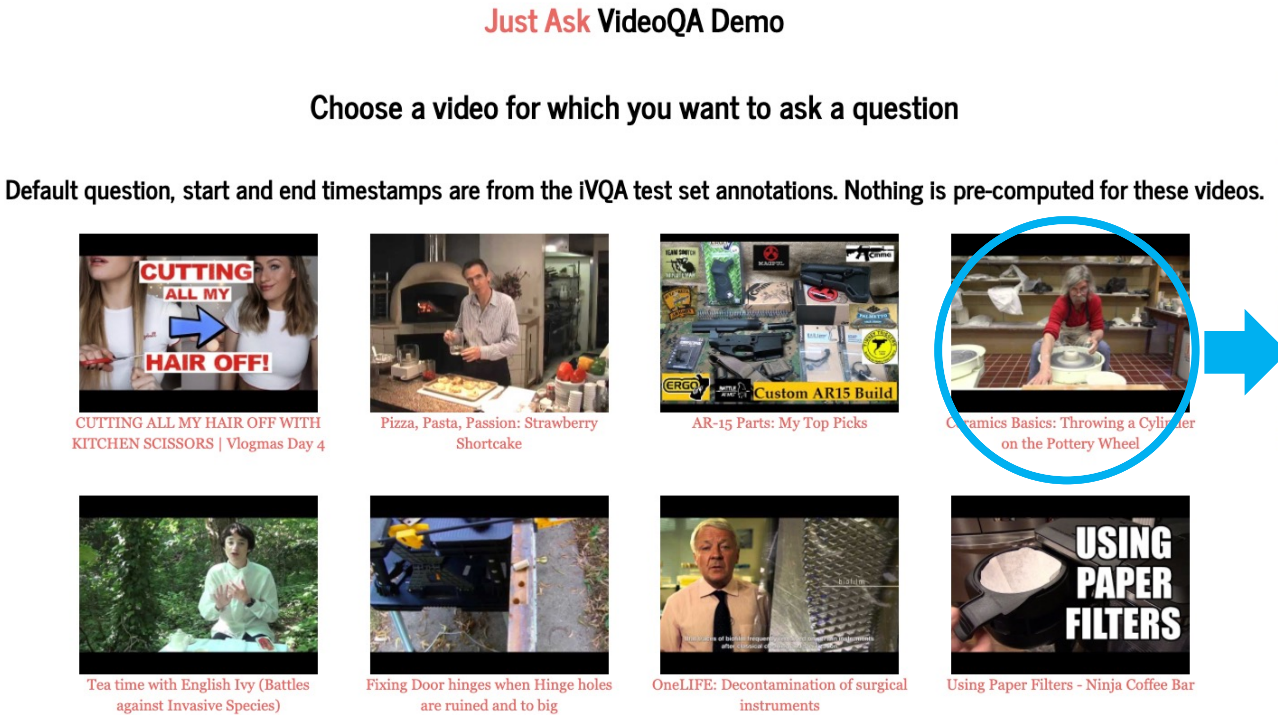
Question: What is the largest object at the right of the man?
GT answer: wheelbarrow
QA-T (HowToVQA69M): statue
VQA-T (HowTo100M): trowel
Just Ask: wheelbarrow



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

Source of the examples: iVQA dataset

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



Results after finetuning

SoTA on 4 existing VideoQA datasets

Method	Pretraining data	MSRVTT-QA	MSVD-QA
E-SA [87]		29.3	27.6
ST-TP [35]		30.9	31.3
AMU [87]		32.5	32.0
Co-mem [27]		32.0	31.7
HME [23]		33.0	33.7
LAGCN [33]		—	34.3
HGA [37]		35.5	34.7
QueST [36]		34.6	36.1
HCRN [42]		35.6	36.1
ClipBERT [44]	COCO [15]+ Visual Genome [41]	37.4	—
SSML [6]	HowTo100M	35.1	35.1
CoMVT [68]	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 [94]		31.8	—
MAR-VQA [105]		34.6	—
HERO [48]	HowTo100M + TV Dataset	—	74.1
CoMVT [68]	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).

Zero-Shot Video Question Answering via Frozen Bidirectional Language Models

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

Project page: https://antoyang.github.io/frozenbilm.html

Paper: on arXiv by end of June

Challenges

- SoTA models for zero-shot VQA are based on *frozen* autoregressive language models
- **Issues:** They require billion parameters to work well => hard to train and deploy in practice.
- **Problematic:** Can we tackle zero-shot VideoQA with lighter models?
- **Idea:** Use bidirectional language models!

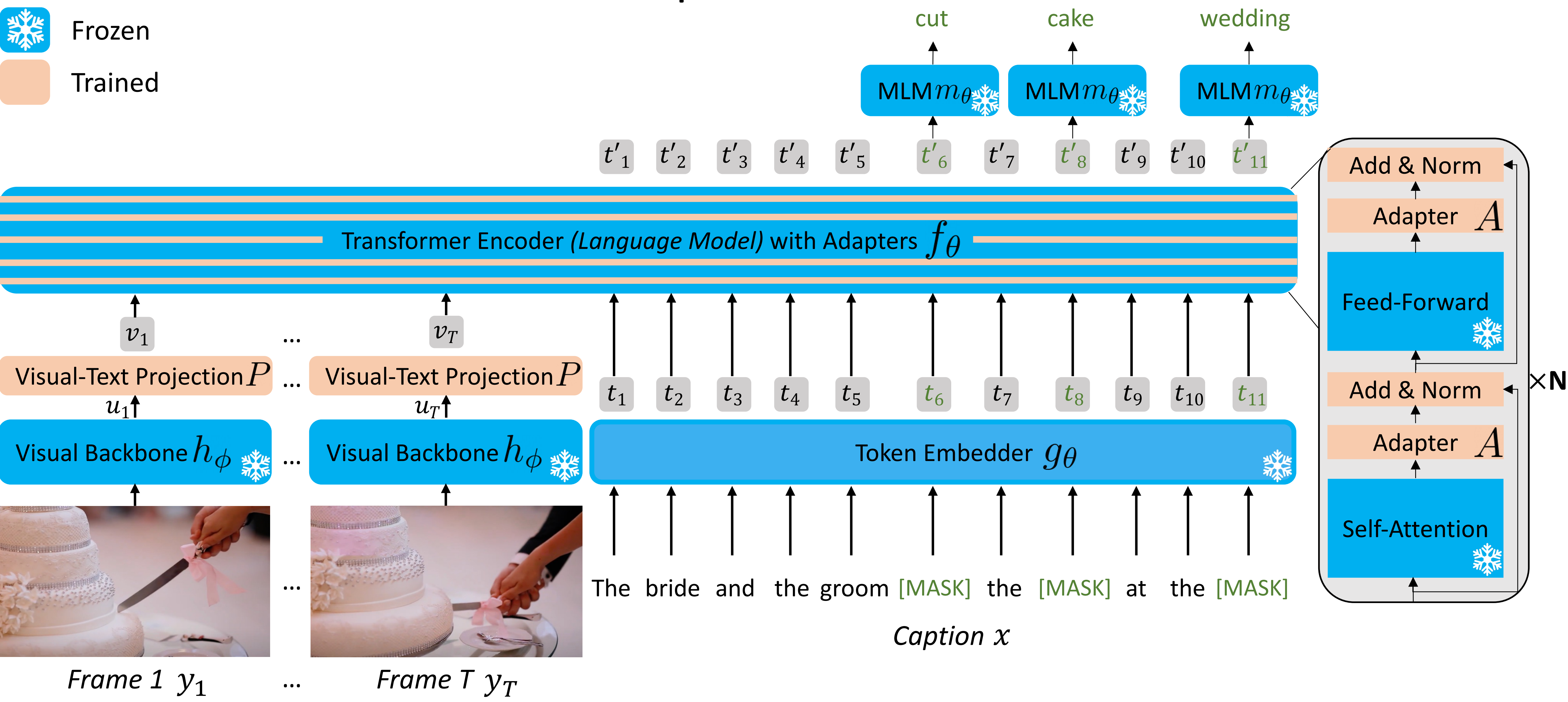
Autoregressive language models

The -> dog
The dog -> is
The dog is -> running
The dog is running -> in
The dog is running in -> the
The dog is running in the -> snow
The dog is running in the snow -> EOS

Bidirectional language models (BiLM)

The dog is [MASK] in the snow -> running

Multi-modal adaptation of a *Frozen* BiLM



Training data: videos with alt-text description

- Videos with alt-text description are easy to obtain at scale.
- Such data is less noisy than narrated videos [Bain 2021].

Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking.

Female cop talking on walkietalkie, responding emergency call, crime prevention

Billiards, concentrated young woman playing in club.

[Bain 2021] Frozen in Time: A Joint Video and Text Encoder for End-to-End Retrieval, Bain et al, ICCV 2021.

Zero-shot inference through unmasking

- Open-ended VideoQA:

  “[CLS] Question: <Question>? Answer: [MASK]. Subtitles: <Subtitles> [SEP]”

- Multiple-choice VideoQA:

  “[CLS] Question: <Question>? Is it '<Answer Candidate>'? [MASK]. Subtitles: <Subtitles> [SEP]”

- Video-conditioned fill-in-the-blank:

  “[CLS] <Sentence with a [MASK] token>. Subtitles: <Subtitles> [SEP]”

Ablation: Modalities

- Vision is essential.
- Speech helps.

	Visual	Speech	Fill-in-the-blank LSMDC	iVQA	MSRVTT-QA	Open-ended MSVD-QA	ActivityNet-QA	TGIF-QA	Multiple-choice How2QA	TVQA
1.	✗	✗	47.9	11.0	6.4	11.3	22.6	32.3	29.6	23.2
2.	✗	✓	49.8	13.2	6.5	11.7	23.1	32.3	45.9	44.1
3.	✓	✗	50.9	26.2	**16.9**	33.7	**25.9**	**41.9**	41.9	29.7
4.	✓	✓	**51.5**	**26.8**	16.7	**33.8**	**25.9**	**41.9**	**58.4**	**59.2**

Table 2: Impact of the visual and speech modalities on zero-shot VideoQA. Rows 1 and 2 report results for a pretrained language model without any visual input. Rows 3 and 4 give results for a *FrozenBiLM* model pretrained on WebVid10M.

Ablation: Model Training

- Freezing the pretrained BiLM considerably helps.
- Adapters help.

	LM Pretraining	*Frozen* LM	Adapters	Fill-in-the-blank LSMDC	Open-ended iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	Multiple-choice How2QA	TVQA
1.	✗	✗	✗	0.5	0.3	0.1	0.0	0.5	0.0	32.4	20.7
2.	✓	✗	✗	37.1	21.0	**17.6**	31.9	20.7	30.7	45.7	45.6
3.	✓	✓	✗	50.7	**27.3**	16.8	32.2	24.7	41.0	53.5	53.4
4.	✓	✓	✓	**51.5**	26.8	16.7	**33.8**	**25.9**	**41.9**	**58.4**	**59.2**

Table 1: The effect of initializing and training various parts of our model evaluated on zero-shot VideoQA. All models are trained on WebVid10M and use multi-modal inputs (video, speech and question) at inference.

Bidirectional vs autoregressive frameworks

Bidirectional models perform better, train faster and require less parameters.

Method	Language Model	# LM params	Train time (GPUH)	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA
Autoregressive	1. GPT-Neo-1.3B	1.3B	200	6.6	4.2	10.1	17.8	14.4
	2. GPT-Neo-2.7B	2.7B	360	9.1	7.7	17.8	17.4	20.1
	3. GPT-J-6B	6B	820	21.4	9.6	26.7	24.5	37.3
Bidirectional	4. BERT-Base	**110M**	**24**	12.4	6.4	11.7	16.7	23.1
	5. BERT-Large	340M	60	12.9	7.1	13.0	19.0	21.5
	6. DeBERTa-V2-XLarge	890M	160	**27.3**	**16.8**	**32.2**	**24.7**	**41.0**

Table 4: Comparison of autoregressive language models (top) and bidirectional language models (bottom) for zero-shot VideoQA. All variants are trained on WebVid10M for the same number of epochs.

Zero-shot quantitative results

SoTA on 8 datasets spanning fill-in-the-blank, open-ended VideoQA and multiple-choice VideoQA.

Method	Training Data	Fill-in-the-blank	Open-ended					Multiple-choice	
		LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
Random	—	0.1	0.1	0.1	0.1	0.1	0.1	25	20
CLIP ViT-L/14 [68]	400M image-texts	1.2	9.2	2.1	7.2	1.2	<u>3.6</u>	47.7	<u>26.1</u>
Just Ask [97]	HowToVQA69M + WebVidVQA3M	—	<u>13.3</u>	5.6	<u>13.5</u>	<u>12.3</u>	—	<u>53.1</u>	—
Reserve [105]	YT-Temporal-1B	<u>31.0</u>	—	<u>5.8</u>	—	—	—	—	—
FrozenBiLM (Ours)	WebVid10M	**51.5**	**26.8**	**16.7**	**33.8**	**25.9**	**41.9**	**58.4**	**59.7**

Table 5: Comparison with the state of the art for zero-shot VideoQA.

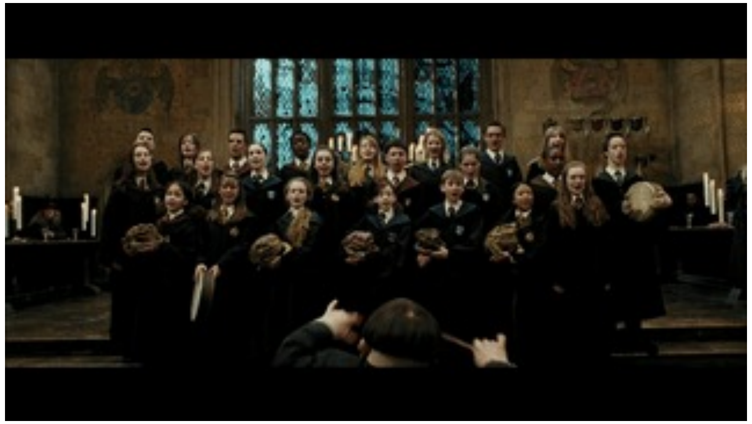
Zero-shot qualitative results (open-ended)

Question: What is the man holding at the start of the video?
GT answer: guitar, electric guitar
Just Ask: typewriter
UnFrozenBiLM: beer
FrozenBiLM (text-only): scissors
FrozenBiLM: guitar

Question: What item hanging on the wall features a tree?
GT answer: quilt
Just Ask: christmas tree
UnFrozenBiLM: fabric
FrozenBiLM (text-only): tree
FrozenBiLM: quilt

Question: Which category of sports does this sport belong to?
GT answer: surfing
Just Ask: second
UnFrozenBiLM: swimming
FrozenBiLM (text-only): 1
FrozenBiLM: surfing

Zero-shot qualitative results (fill-in-the-blank)



Sentence: Each singer in the front row ____ a huge toad.
GT answer: holds
UnFrozenBiLM: plays
FrozenBiLM (text-only): wears
FrozenBiLM: holds

Sentence: Someone ____ him to the truck and across the street.
GT answer: chases
UnFrozenBiLM: follow
FrozenBiLM (text-only): drags
FrozenBiLM: chases

Sentence: A woman wraps food in newspapers and brings it over to their ____.
GT answer: table
UnFrozenBiLM: man
FrozenBiLM (text-only): home
FrozenBiLM: table

Zero-shot qualitative results (multiple-choice)

Question: When did the chef flipped over the layer of rice and seaweed?
GT answer: holds
A0: after she sprinkled sesame
A1: after she added cucumber
A2: after she added fish
A3: after she cut the cucumbers

UnFrozenBiLM: A3
FrozenBiLM (text-only):A1
FrozenBiLM: A0

Results after finetuning

- Freezing the BiLM also helps in the fully-supervised setting.
- SoTA on 6 out of 8 datasets + high parameter efficiency.

Method	# Trained Params	Fill-in-the-blank	Open-ended					Multiple-choice	
		LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
HCRN [42]	44M	—	—	35.4	36.8	—	57.9	—	71.4
HERO [51]	119M	—	—	—	—	—	—	74.1	73.6
ClipBERT [45]	114M	—	—	37.4	—	—	60.3	—	—
Just Ask [97]	157M	—	35.4	41.8	47.5	39.0	—	85.3	—
SiaSamRea [102]	—	—	—	41.6	45.5	39.8	60.2	84.1	—
MERLOT [104]	223M	52.9	—	43.1	—	41.4	**69.5**	—	78.7
Reserve [105]	644M	—	—	—	—	—	—	—	**86.1**
VIOLET [19]	198M	53.7	—	43.9	47.9	—	68.9	—	—
All-in-one [90]	110M	—	—	46.8	48.3	—	66.3	—	—
UnFrozenBiLM (Ours)	890M	58.9	37.7	45.0	53.9	**43.2**	66.9	**87.5**	79.6
FrozenBiLM (Ours)	**30M**	**63.5**	**39.6**	**47.0**	**54.8**	**43.2**	68.6	86.7	82.0

Table 6: Comparison with the state of the art, and the variant *UnFrozenBiLM* which does not freeze the language model weight, on fully-supervised benchmarks.

Conclusion

- Zero-shot video question answering can be tackled by generating training data using language models and narrated videos
- It can also be efficiently tackled without data generation procedure using frozen bidirectional language models