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

ICCV 2021

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

Video Question Answering

- Video question answering is a good task to evaluate a method's ability to understand video content.

Question: What fruit is shown at the end?

Answer: watermelon

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

Answer: wheelbarrow

Challenges

- Manually annotating a large number of videos with a diverse set of questions and answers is extremely time-consuming and costly.

Challenges

- Manually annotating a large number of videos with a diverse set of questions and answers is extremely time-consuming and costly.

Can we do this without relying on manually annotated video data?

Main Idea

- Automatically generate question & answer from narrated YouTube videos.
- Such automatically generated data can then be used to train VideoQA model.

Speech: The sound is amazing on this piano.

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

HowTo100M Dataset

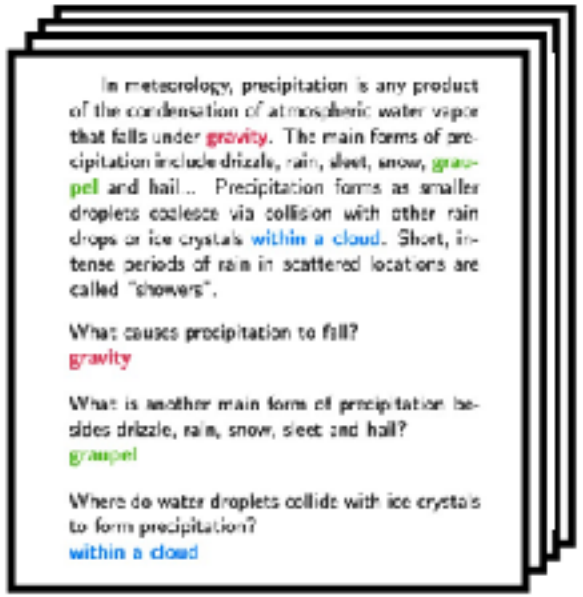
- 136M video clips with automatically transcribed speech narrations.

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

Text-only Supervision

- The authors use language models trained on manually annotated text-only question-answering corpus.

Manually annotated QA text corpus



Training

Answer extractor Transformer T_a

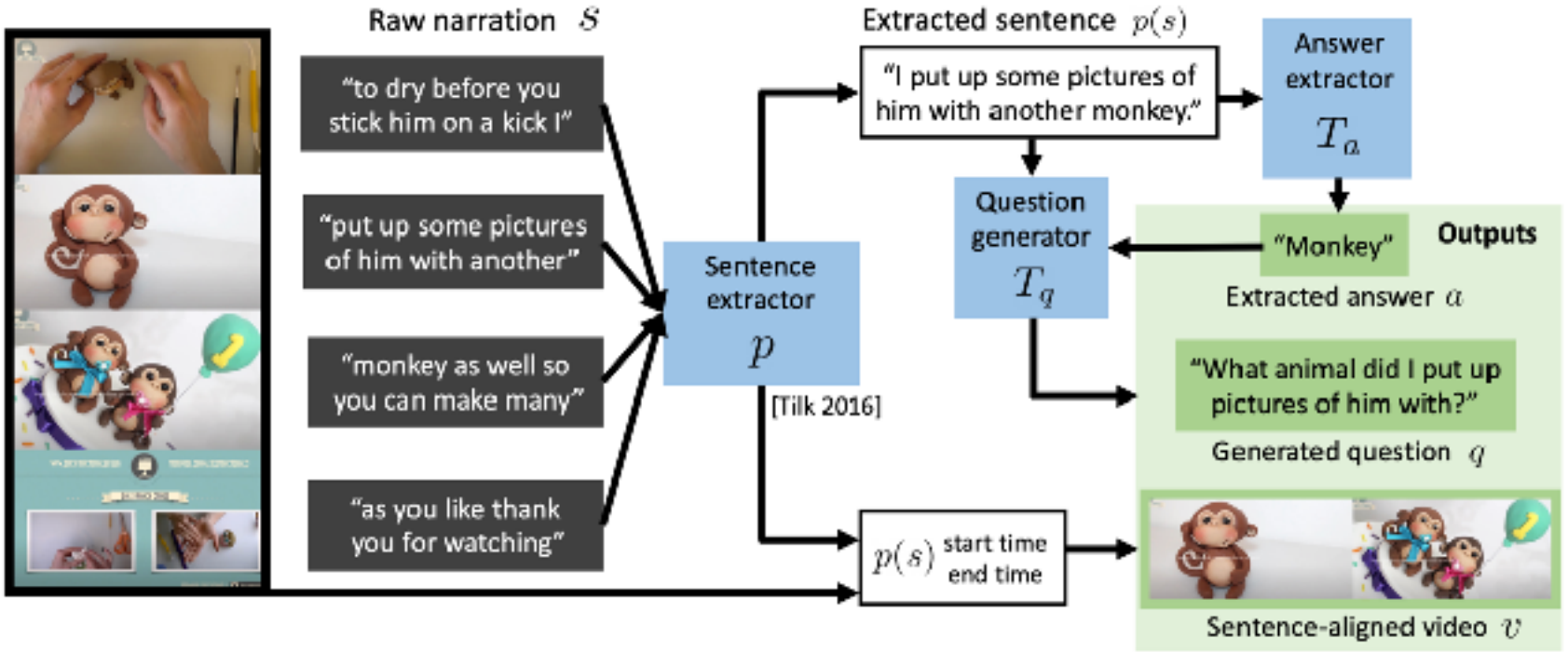
Training

Question generator Transformer T_q

Raffel et al. “Exploring the limits of transfer learning with a unified text-to-text transformer,” JMLR 2020

Generating VideoQA Data

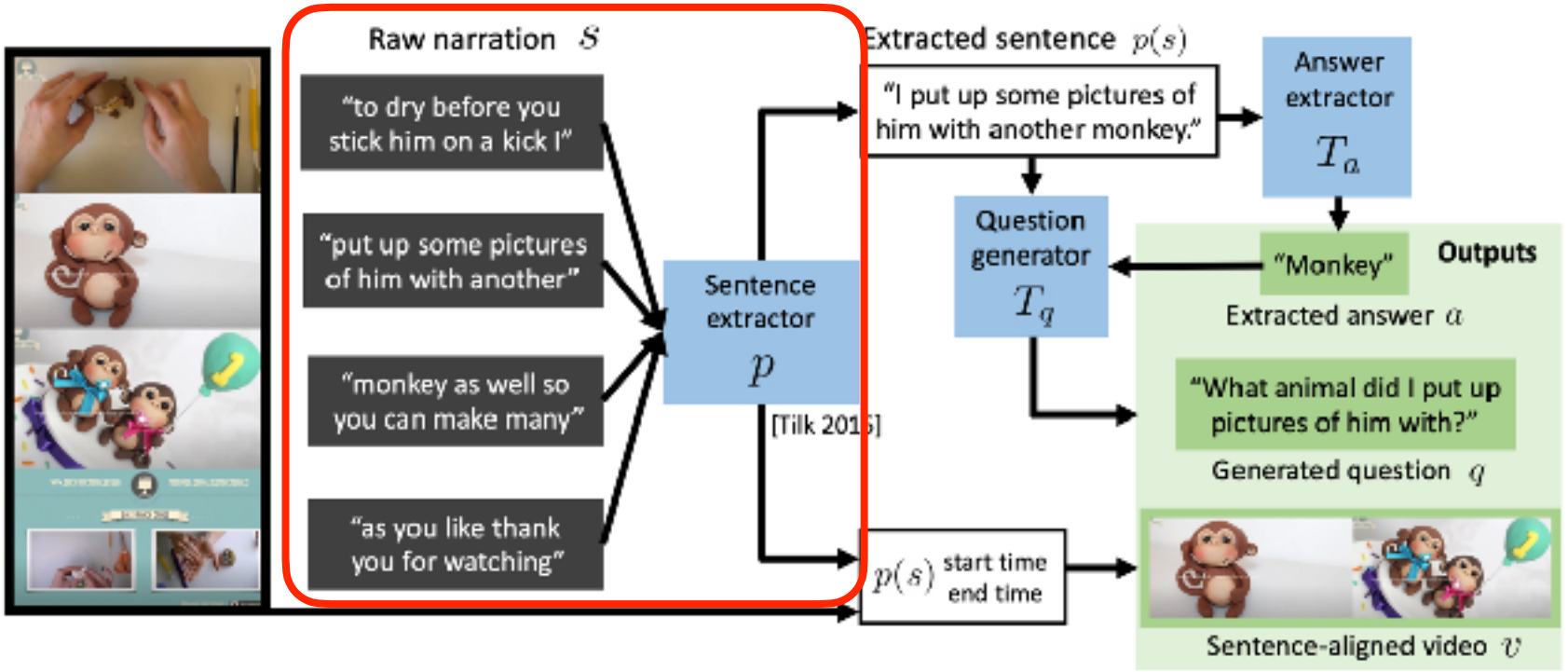
- The authors use language models trained on manually annotated text-only question-answering corpus.



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

Generating VideoQA Data

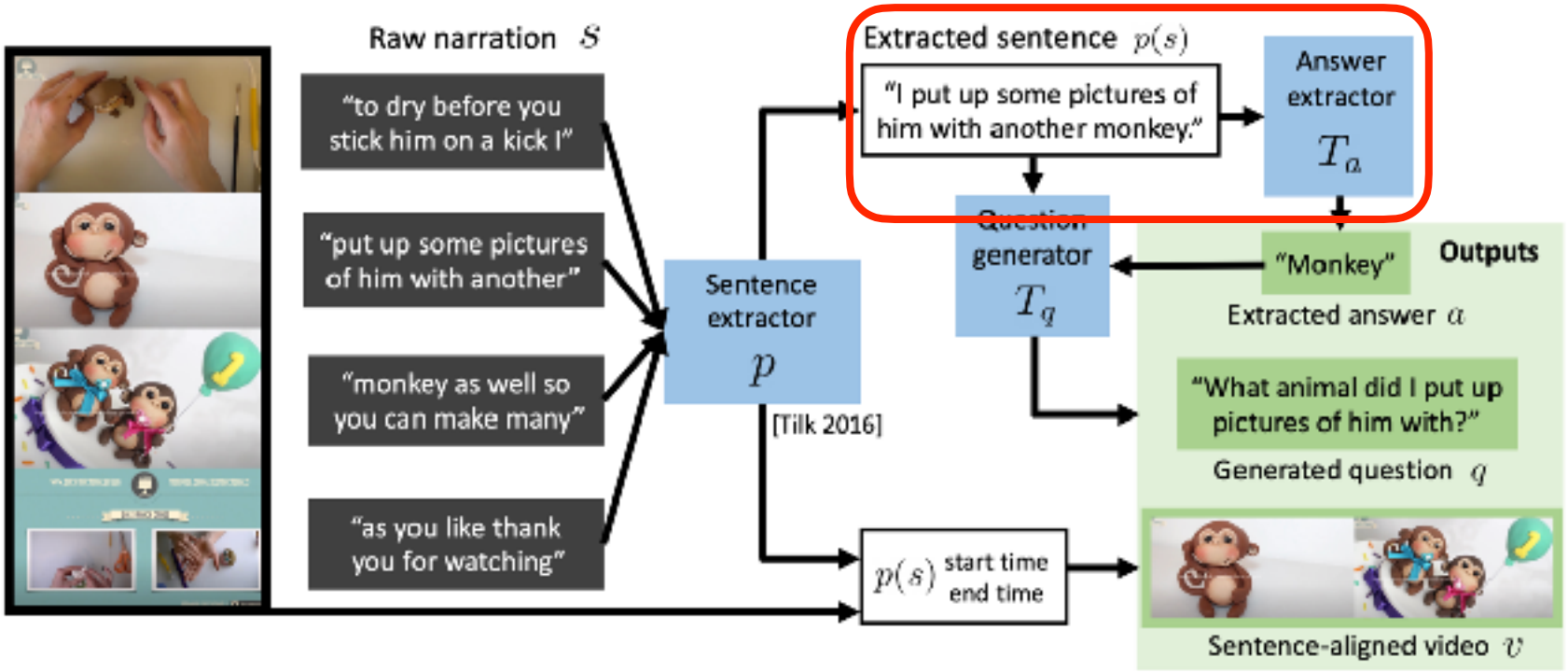
- The authors use language models trained on manually annotated text-only question-answering corpus.



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

Generating VideoQA Data

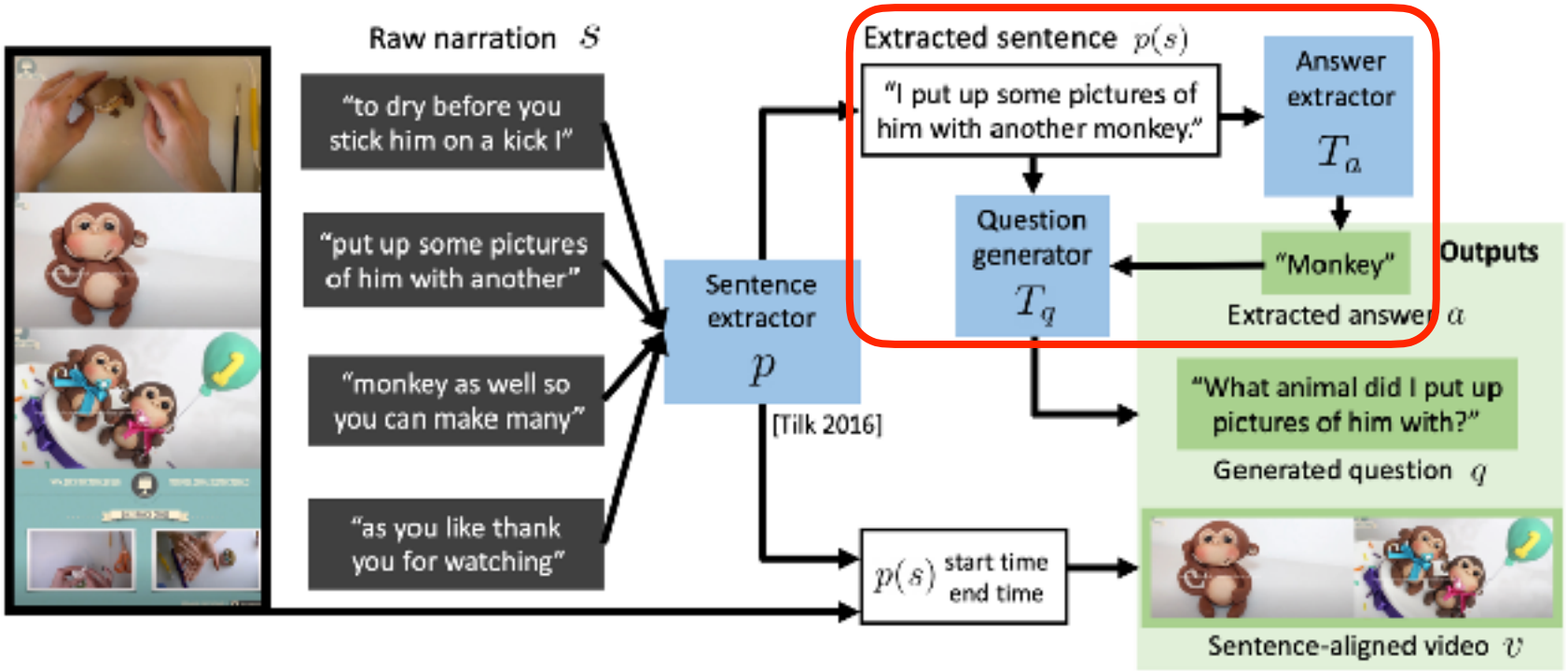
- The authors use language models trained on manually annotated text-only question-answering corpus.



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

Generating VideoQA Data

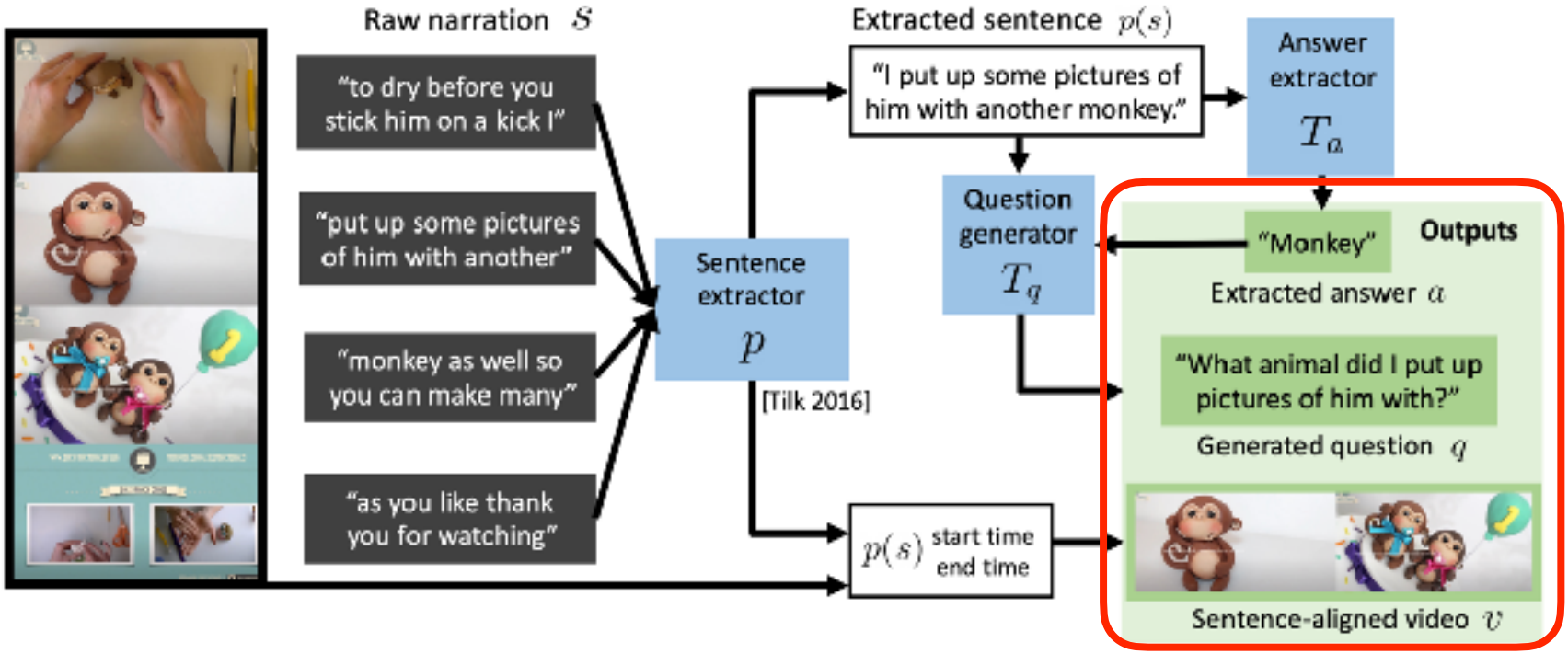
- The authors use language models trained on manually annotated text-only question-answering corpus.



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

Generating VideoQA Data

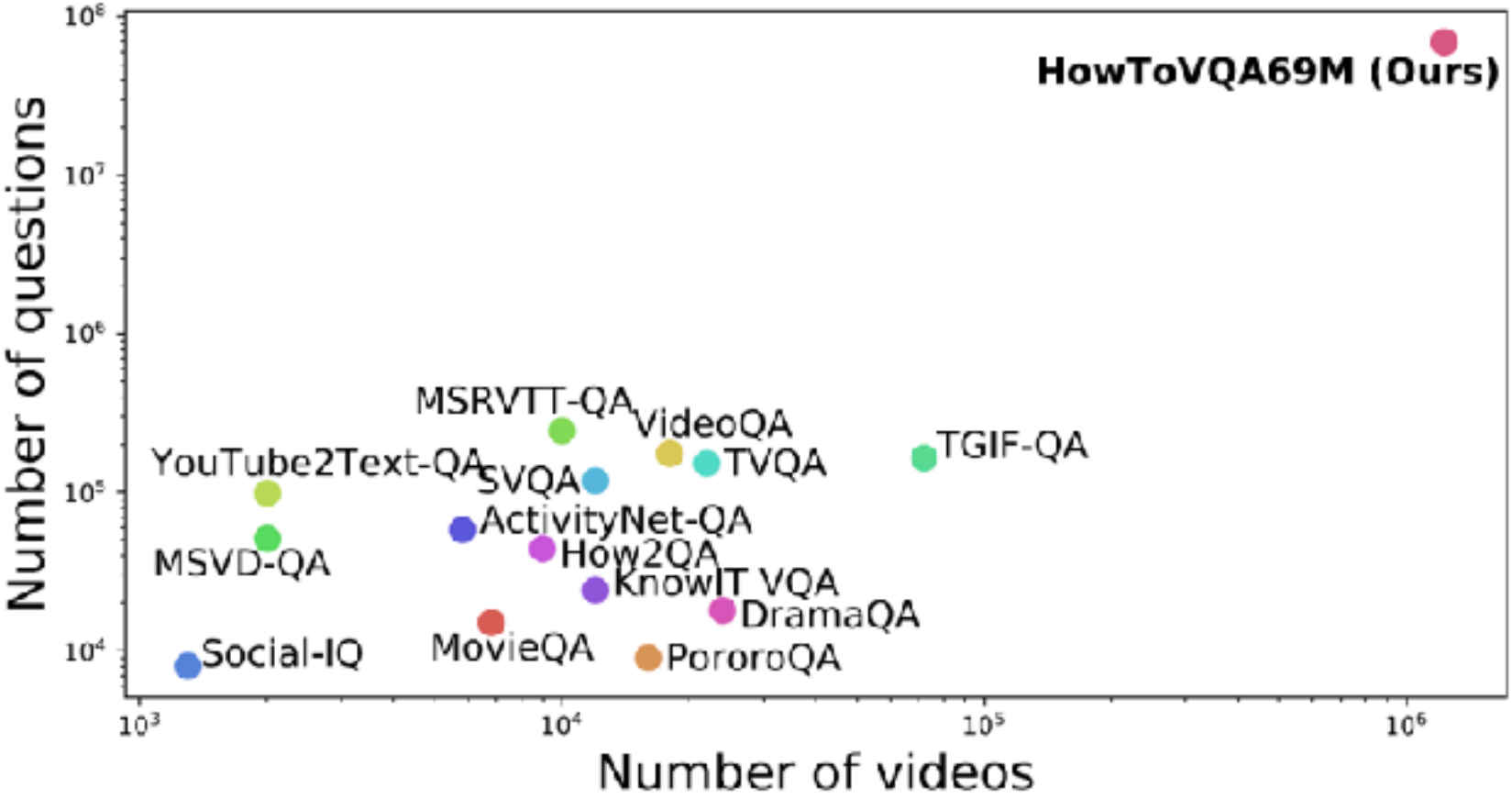
- The authors use language models trained on manually annotated text-only question-answering corpus.



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

HowToVQA69M

- A large scale videoQA dataset consisting of 69M video clip, question and answer triplets.



Noise in HowToVQA69M

- HowToVQA69M annotations are relatively noisy.

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%

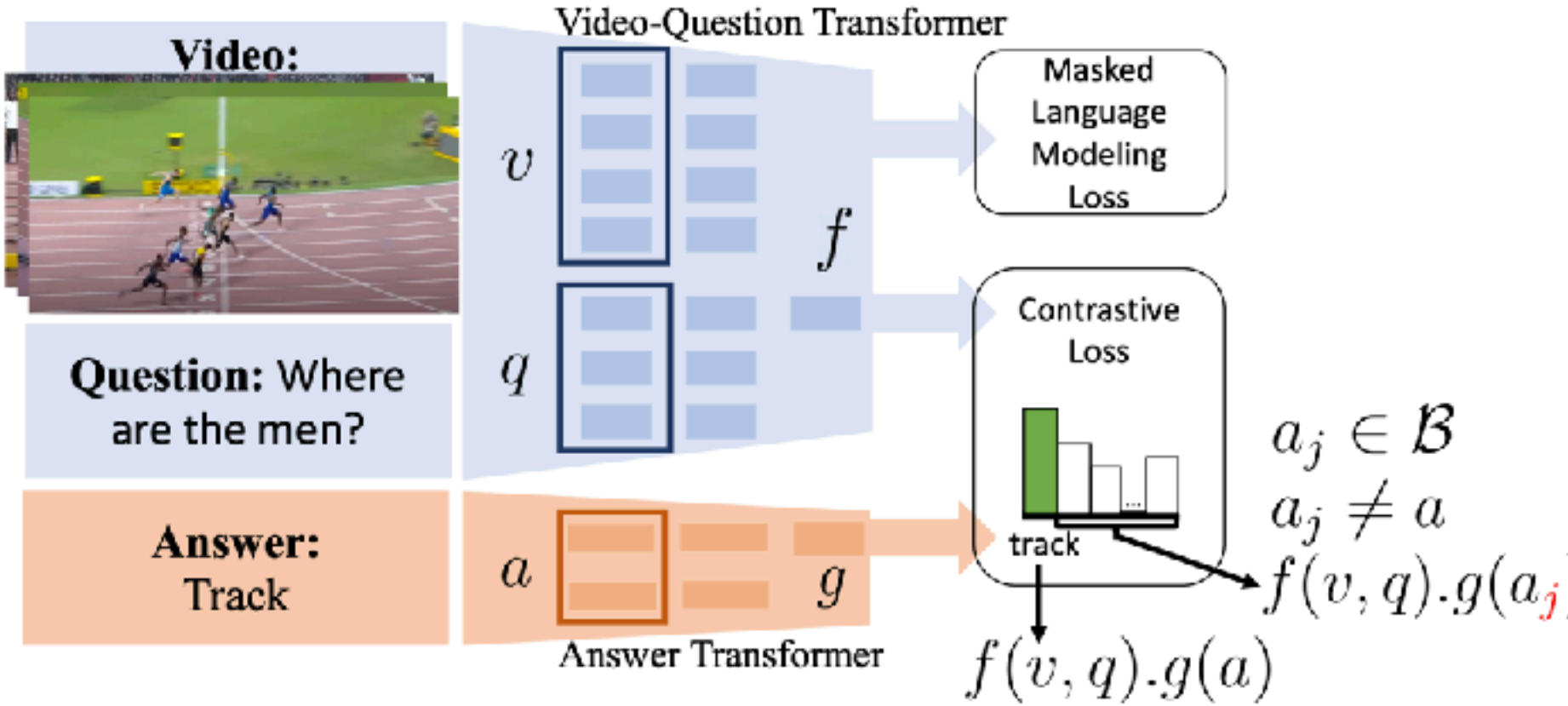
Speech: You can't miss this...
Generated question: What can't you do?
Generated answer: miss

QA unrelated to video

≈ 39%

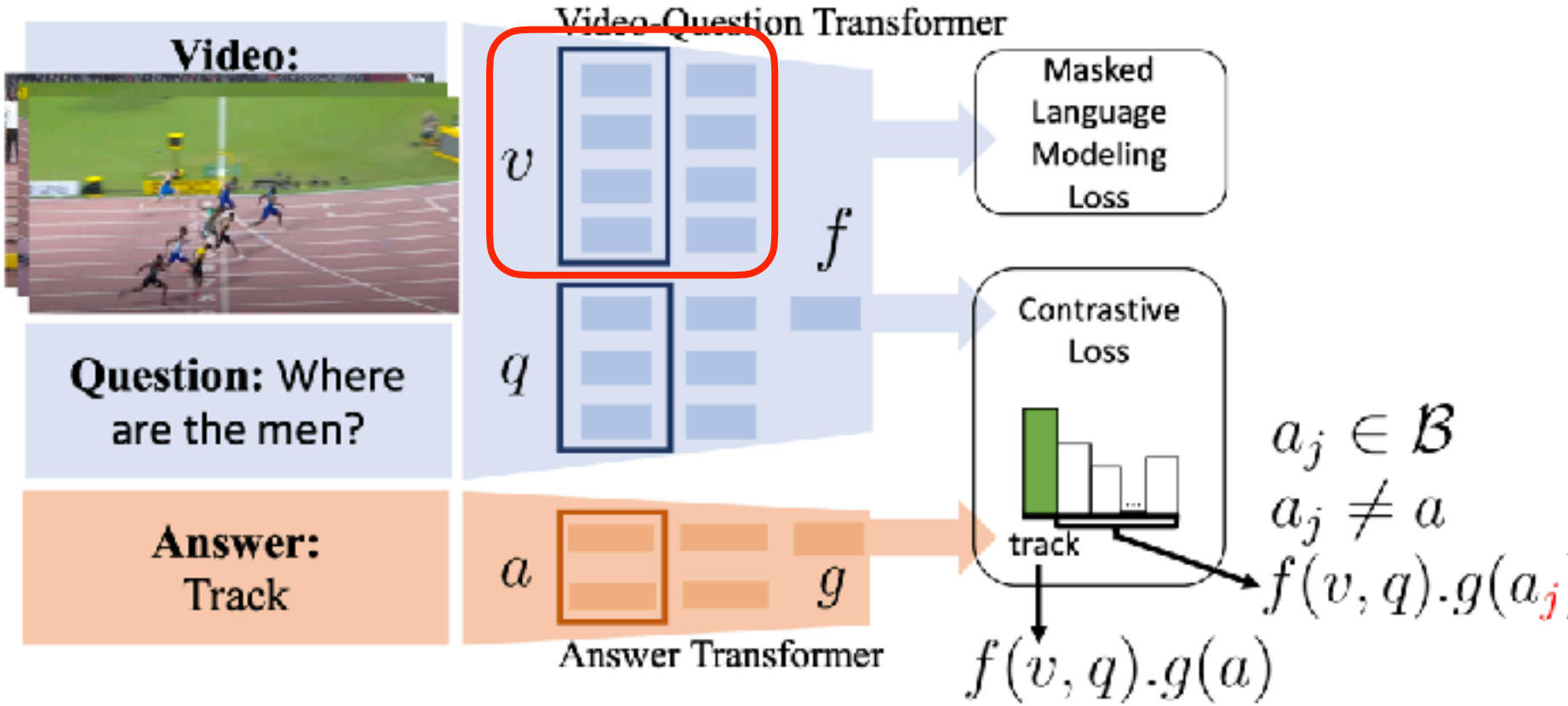
VideoQA model

- The model is composed of two modules: (i) a video-question module, and (ii) an answer embedding module.



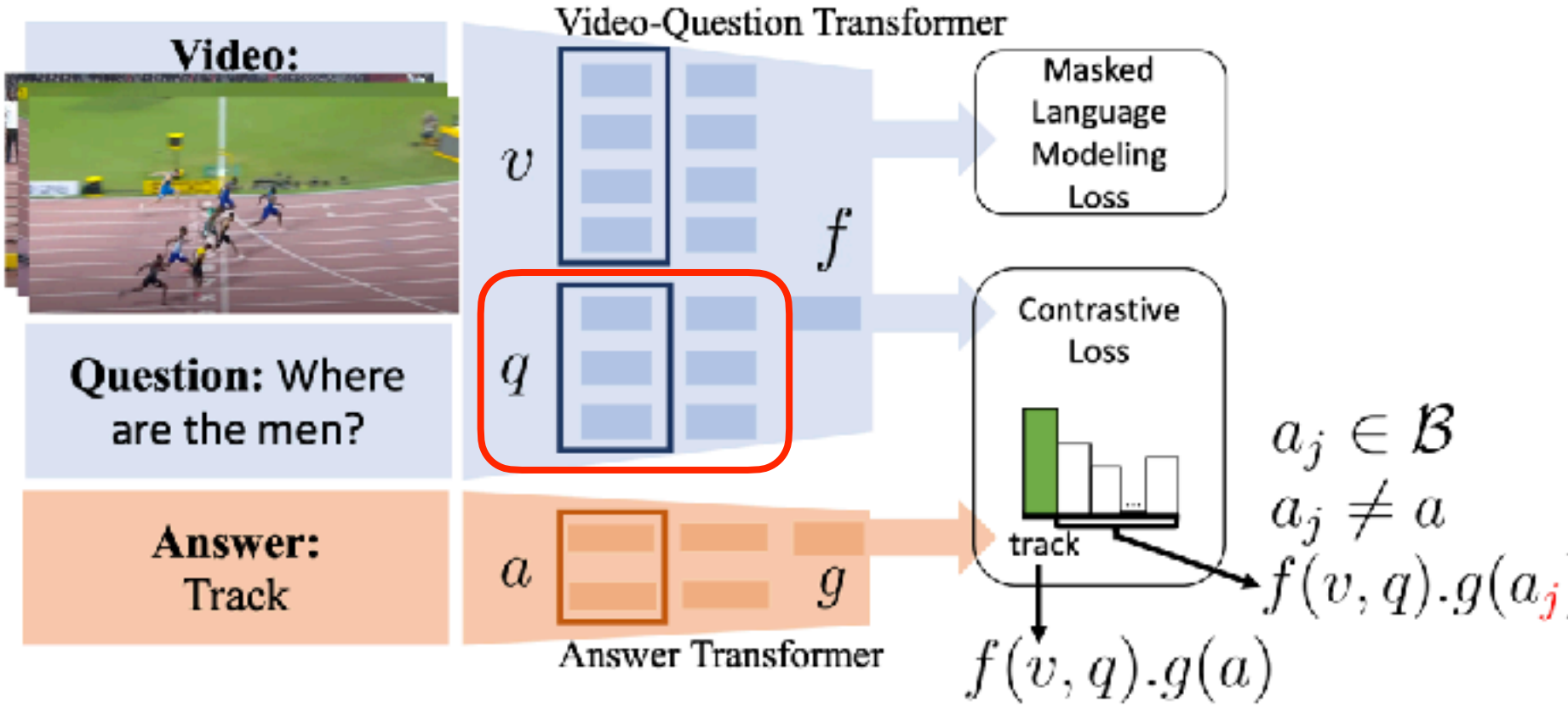
VideoQA model

- The model is composed of two modules: (i) a video-question module, and (ii) an answer embedding module.



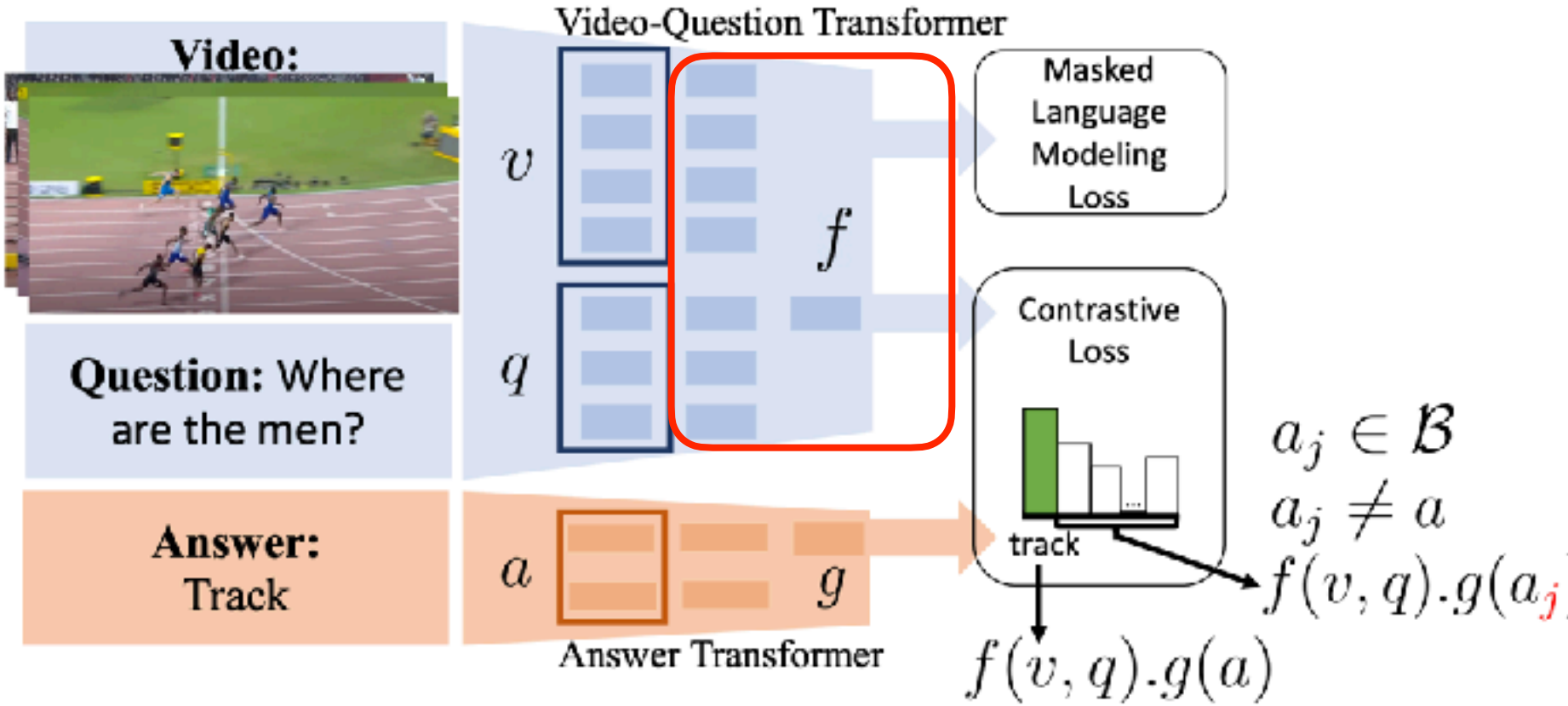
VideoQA model

- The model is composed of two modules: (i) a video-question module, and (ii) an answer embedding module.



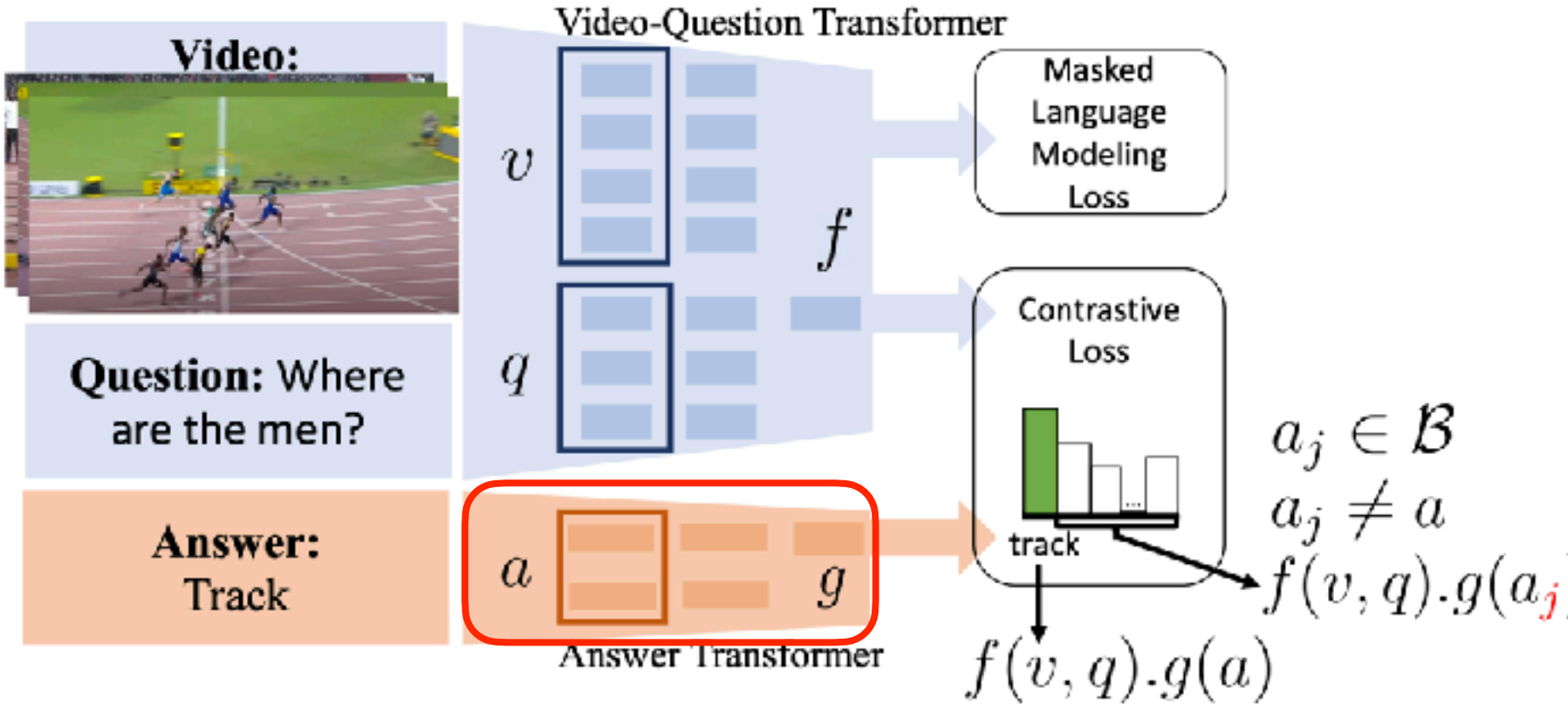
VideoQA model

- The model is composed of two modules: (i) a video-question module, and (ii) an answer embedding module.



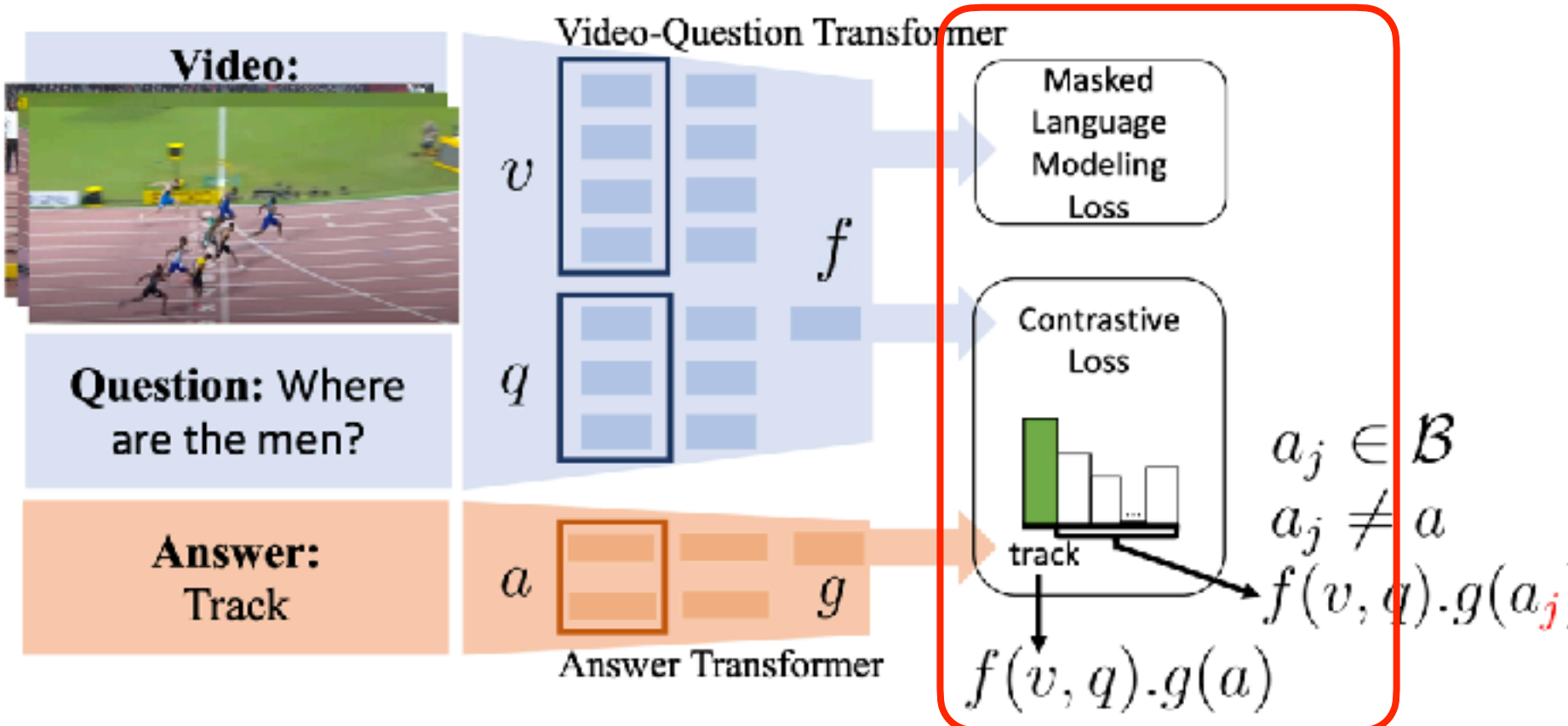
VideoQA model

- The model is composed of two modules: (i) a video-question module, and (ii) an answer embedding module.



VideoQA model

- The model is composed of two modules: (i) a video-question module, and (ii) an answer embedding module.



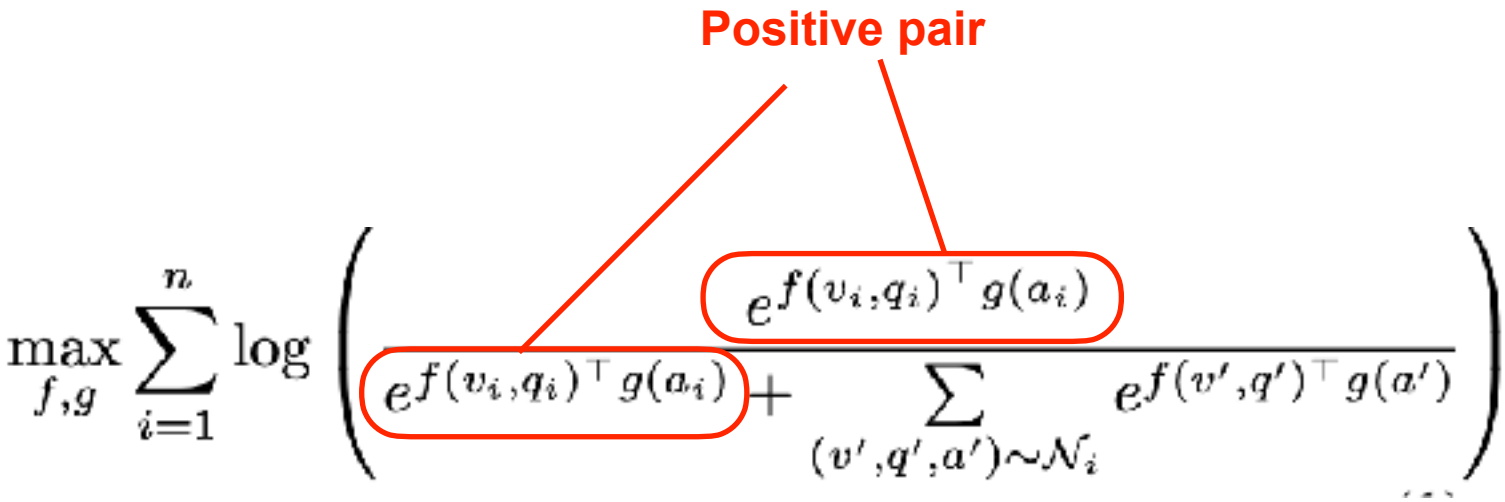
Training Procedure

- The authors use a standard contrastive loss function to train their model.

$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{e^{f(v_i,q_i)^\top g(a_i)}}{e^{f(v_i,q_i)^\top g(a_i)} + \sum_{(v',q',a')\sim\mathcal{N}_i} e^{f(v',q')^\top g(a')}} \right)$$

Training Procedure

- The authors use a standard contrastive loss function to train their model.



Training Procedure

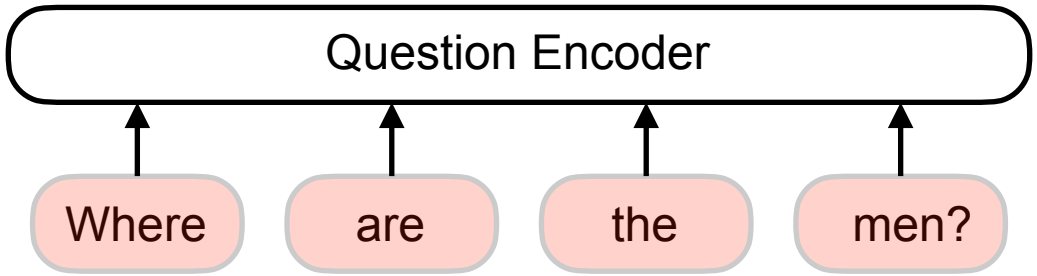
- The authors use a standard contrastive loss function to train their model.

$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{e^{f(v_i,q_i)^\top g(a_i)}}{e^{f(v_i,q_i)^\top g(a_i)} + \sum_{(v',q',a')\sim\mathcal{N}_i} e^{f(v',q')^\top g(a')}} \right)$$

Negative pairs

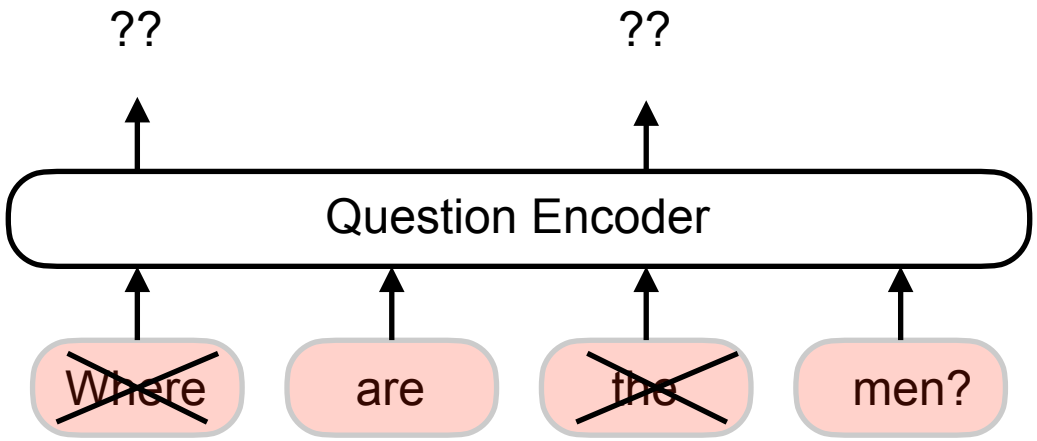
Masked Language Modeling

- Masking loss is applied to questions tokens during training.



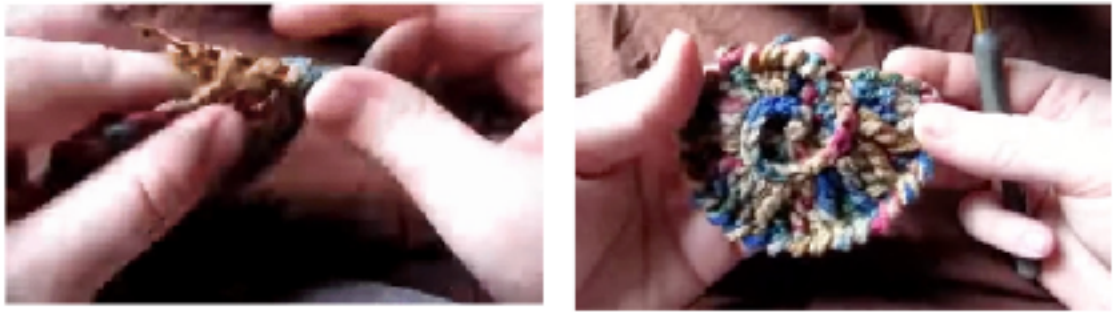
Masked Language Modeling

- Masking loss is applied to questions tokens during training.



iVQA Benchmark

- 10K manually annotated videos and question pairs from HowTo100M.
- Each clip is manually annotated with one question and 5 answers on Amazon Mechanical Turk.



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

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

Zero-shot VideoQA

- No manual supervision of visual data is used.
- The performance is evaluated on five videoQA datasets.
- No finetuning is done on any of these 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

- No manual supervision of visual data is used.
- The performance is evaluated on five videoQA datasets.
- No finetuning is done on any of these 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**

The performance of the random baseline is very low.

Zero-shot VideoQA

- No manual supervision of visual data is used.
- The performance is evaluated on five videoQA datasets.
- No finetuning is done on any of these datasets.

	Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
	Random	ø	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**

The proposed method significantly outperforms text-only question answering baseline.

Zero-shot VideoQA

- No manual supervision of visual data is used.
- The performance is evaluated on five videoQA datasets.
- No finetuning is done on any of these datasets.

	Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
	Random	ø	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**

Pretraining on HowToVQA69M is much more beneficial than pretraining on HowTo100M.

Zero-shot VideoQA

- The values next to the ground truth (GT) answers indicate the number of annotators that gave the answer.

Question: What design are they making?
GT Answer: rose (4), rose flower (1)
QA-T (HowToVQA69M): pinwheel
VQA-T (HowTo100M): piping bag
VQA-T (HowToVQA69M): rose

Question: What is in the man's hand?
GT Answer: shovel (3), spade (2)
QA-T (HowToVQA69M): coin
VQA-T (HowTo100M): planting
VQA-T (HowToVQA69M): shovel

Comparison to State-of-the-Art

- The model is first pretrained on HowToVQA69M.
- The authors then finetune it on the downstream dataset.

Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HCRN [Le 2020]	Ø	-	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 (Ø)	Ø	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**

Comparison to State-of-the-Art

- The model is first pretrained on HowToVQA69M.
- The authors then finetune it on the downstream dataset.

Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HCRN [Le 2020]	Ø	-	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 (Ø)	Ø	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**

The baseline model without any pretraining already achieves results close to the state-of-the-art.

Comparison to State-of-the-Art

- The model is first pretrained on HowToVQA69M.
- The authors then finetune it on the downstream dataset.

Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HCRN [Le 2020]	Ø	-	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 (Ø)	Ø	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**

HowToVQA69M pretraining brings substantial performance boost across all 5 datasets.

Comparison to State-of-the-Art

- The model is first pretrained on HowToVQA69M.
- The authors then finetune it on the downstream dataset.

Method	Pretraining Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	How2QA
HCRN [Le 2020]	Ø	-	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 (Ø)	Ø	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**

HowTo100M pretraining is less useful.

Importance of Dataset 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**

Contributions

- Automatically generated large-scale VideoQA dataset, using text-only supervision.
- Manually collected iVQA benchmark with redundant annotations and reduced language bias.
- Demonstrates the importance of large-scale pretraining, and sets the new state-of-the-art on multiple VideoQA benchmarks.

Discussion Questions

- Limitations of the proposed framework / dataset?

Discussion Questions

- Limitations of the proposed framework / dataset?
- Could this be applied to different domains beyond instructional videos?