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

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]

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

Answer extractor
Transformer
\(T_a\)

Question generator
Transformer
\(T_q\)

Manually annotated
QA text corpus

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Generating VideoQA data

Raw narration $S$

- "to dry before you stick him on a kick!"
- "put up some pictures of him with another monkey."
- "monkey as well so you can make many"
- "as you like thank you for watching"

Extracted sentence $p(s)$

- "I put up some pictures of him with another monkey."

Sentence extractor $p$

Answer extractor $T_a$

- "Monkey"

Extracted answer $a$

Generated question $q$

- "What animal did I put up pictures of him with?"

Outputs $U$

Sentence-aligned video $U$

Generating VideoQA data

Raw narration $S$

“to dry before you stick him on a kick I”

“put up some pictures of him with another”

“monkey as well so you can make many”

“as you like thank you for watching”

Extracted sentence $p(s)$

“I put up some pictures of him with another monkey.”

Answer extractor $T_a$

“Monkey”

Extracted answer $a$

“What animal did I put up pictures of him with?”

Generated question $q$

Outputs

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

Approximately 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

Approximately 31%

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

Approximately 39%

**QA Generation error**

**QA unrelated to video**
VideoQA model and training procedure

Video: Where are the men?

Answer: Track

Video-Question Transformer

Masked Language Modeling Loss

Contrastive Loss

\[ a_j \in B \]

\[ a_j \neq a \]

\[ f(v, q) \cdot g(a_j) \]

\[ f(v, q) \cdot g(a) \]
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:*

<table>
<thead>
<tr>
<th>Method</th>
<th>Pretraining Data</th>
<th>iVQA</th>
<th>MSRVTT-QA</th>
<th>MSVD-QA</th>
<th>ActivityNet-QA</th>
<th>How2QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>Ø</td>
<td>0.09</td>
<td>0.02</td>
<td>0.05</td>
<td>0.05</td>
<td>25.0</td>
</tr>
<tr>
<td>(i) QA-T</td>
<td>HowToVQA69M</td>
<td>4.4</td>
<td>2.5</td>
<td>4.8</td>
<td>11.6</td>
<td>38.4</td>
</tr>
<tr>
<td>(ii) VQA-T</td>
<td>HowTo100M</td>
<td>1.9</td>
<td>0.3</td>
<td>1.4</td>
<td>0.3</td>
<td>46.2</td>
</tr>
<tr>
<td>(iii) VQA-T</td>
<td>HowToVQA69M</td>
<td><strong>12.2</strong></td>
<td><strong>2.9</strong></td>
<td><strong>7.5</strong></td>
<td><strong>12.2</strong></td>
<td><strong>51.1</strong></td>
</tr>
</tbody>
</table>
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

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

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:

<table>
<thead>
<tr>
<th>Pretraining Data</th>
<th>Finetuning</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓ HowToVQA69M</td>
<td>✓</td>
<td>47.9</td>
<td>28.1</td>
<td>15.6</td>
<td>8.5</td>
</tr>
<tr>
<td>✓ HowTo100M</td>
<td>✓</td>
<td>46.7</td>
<td>22.0</td>
<td>8.6</td>
<td>3.6</td>
</tr>
<tr>
<td>✓ Ø</td>
<td></td>
<td>38.4</td>
<td>16.7</td>
<td>5.9</td>
<td>2.6</td>
</tr>
<tr>
<td>✗ HowToVQA69M</td>
<td></td>
<td>9.0</td>
<td>8.0</td>
<td>9.5</td>
<td>7.7</td>
</tr>
</tbody>
</table>

Most frequent answers

Least frequent answers
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.

<table>
<thead>
<tr>
<th>Generation method</th>
<th>Zero-Shot</th>
<th>Finetuning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>iVQA</td>
<td>ActivityNet-QA</td>
</tr>
<tr>
<td>[Heilman 2010]</td>
<td>7.4</td>
<td>1.1</td>
</tr>
<tr>
<td>Ours</td>
<td>12.2</td>
<td>12.2</td>
</tr>
</tbody>
</table>

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

## Ablation: pretraining losses

<table>
<thead>
<tr>
<th>MLM</th>
<th>Sampling without answer repetition</th>
<th>Zero-Shot</th>
<th>Finetuning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>iVQA</td>
<td>MSVD-QA</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>11.1</td>
<td>6.1</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>12.1</td>
<td>7.0</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>10.9</td>
<td>6.4</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td><strong>12.2</strong></td>
<td><strong>7.5</strong></td>
</tr>
</tbody>
</table>

=> Best results with MLM and our contrastive loss
Ablation: scale

<table>
<thead>
<tr>
<th>Pretraining data size</th>
<th>Zero-Shot</th>
<th>Finetuning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>iVQA</td>
<td>MSVD-QA</td>
</tr>
<tr>
<td>0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1%</td>
<td>4.5</td>
<td>3.6</td>
</tr>
<tr>
<td>10%</td>
<td>9.1</td>
<td>6.2</td>
</tr>
<tr>
<td>20%</td>
<td>9.5</td>
<td>6.8</td>
</tr>
<tr>
<td>50%</td>
<td>11.3</td>
<td>7.3</td>
</tr>
<tr>
<td>100%</td>
<td><strong>12.2</strong></td>
<td><strong>7.5</strong></td>
</tr>
</tbody>
</table>

=> 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.