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: https://arxiv.org/abs/2206.08155
Zero-Shot VideoQA [1]

Cross-modal Training

Training data:

Web-scraped Video + Caption

FrozenBiLM

Pretrained BiLM

Little cute toy poodle dog running fast on the beach.

Zero-Shot VideoQA

Test data:

Video + Question

What is the dog doing?

Answer: Running

[1] Just Ask: Learning to Answer Questions from Millions of Narrated Videos, A. Yang et al, ICCV 2021.
FrozenBiLM idea

- **Background:** SoTA models for zero-shot VQA rely on frozen autoregressive language models [2].
- **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 (BiLM)!

| Autoregressive language models | Bidirectional language models (BiLM) |
|-------------------------------|-------------------------------------|
| [BOS] -> The                  | The dog is [MASK] in the snow -> running |
| 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 |                                 |

[2] Multi-modal Few-Shot Learning with Frozen Language Models, M. Tsipoukelli et al, NeurIPS 2021.
Multi-modal adaptation of a Frozen BiLM

Transformer Encoder (Language Model) with Adapters $\hat{f}_\theta$

Visual Backbone $h_\phi$

Visual Text Projection $P$

Add & Norm

Adapter $A$

Feed-Forward

Add & Norm

Adapter $A$

Self-Attention

Caption $\mathcal{C}$

Frame 1 $y_1$

Frame T $y_T$

The bride and the groom $[\text{MASK}]$ the $[\text{MASK}]$ at the $[\text{MASK}]$

MLM $m_6$

MLM $m_6$

MLM $m_6$

$t_1$ $t_2$ $t_3$ $t_4$ $t_5$ $t_6$ $t_7$ $t_8$ $t_9$ $t_{10}$ $t_{11}$

$t'_1$ $t'_2$ $t'_3$ $t'_4$ $t'_5$ $t'_6$ $t'_7$ $t'_8$ $t'_9$ $t'_{10}$ $t'_{11}$

$u_1$ $u_T$

$v_1$ $v_T$
Training data

- **Initialization of the frozen modules:** DeBERTa-V2-Xlarge (900M params) [3] for the BiLM, CLIP ViT-L/14 @224px [4] for the visual backbone.

- **Training data:** Videos with alt-text description from the WebVid10M. Such data are easy to obtain at scale and less noisy than narrated videos [5].

  - 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.

[3] DeBERTa: Decoding-enhanced BERT with Disentangled Attention, P. He et al, ICLR 2021.
[4] Learning transferable visual models from natural language supervision, A. Radford et al, arXiv 2021.
[5] Frozen in Time: A Joint Video and Text Encoder for End-to-End Retrieval, M. Bain et al, ICCV 2021.
Downstream task adaptation

**Answer prediction**: We map the masked token in the following prompts with an answer embedding module which is initialized from the frozen masked language modeling head.

- 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.

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.

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.
FrozenBiLM vs autoregressive LM

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

| Method          | Language Model | # LM params | Train time (GPUH) | iVQA | MSRVT-TQA | MSVD-QA | ActivityNet-QA | TGIF-QA |
|-----------------|----------------|-------------|-------------------|------|-----------|---------|----------------|---------|
| 1. GPT-Neo-1.3B| 1.3B           | 200         | 6.6               | 4.2  | 10.1      | 17.8    | 17.4           | 20.1    |
| Autoregressive  | 2. GPT-Neo-2.7B| 2.7B        | 9.1               | 7.7  | 17.8      | 24.5    | 21.4           | 37.3    |
| 3. GPT-J-6B     | 6B             | 820         | 21.4              | 9.6  | 26.7      | 24.5    | 21.4           | 37.3    |
| 4. BERT-Base    | 110M           | 24          | 12.4              | 6.4  | 11.7      | 16.7    | 23.1           |         |
| Bidirectional   | 5. BERT-Large  | 340M        | 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 video-conditioned fill-in-the-blank, open-ended VideoQA and multiple-choice VideoQA.

| Method       | Training Data          | Speech     | 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             | 25             |
| CLIP ViT-L/14 [71] 400M image-texts | ×          | 1.2        | 9.2               | 2.1                      | 7.2             | 3.6            | 47.7 | 26.1 |
| Just Ask [102] | HowToVQA69M + WebVidVQA3M | ×          | —                 | 13.3                     | 5.6             | 13.5           | —               | 53.1 | — |
| Reserve [110] | YT-Temporal-1B        | ×          | 31.0              | —                        | 5.8             | —              | —               | —     | — |
| FrozenBiLM (Ours) WebVid10M | ×          | 50.9       | 26.2              | 16.9                     | 33.7            | 25.9           | 41.9            | —     | — |
| 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 [1]:** typewriter
**UnFrozenBiLM:** beer
**FrozenBiLM (text-only):** scissors
**FrozenBiLM:** guitar

**Question:** What item hanging on the wall features a tree?
**GT answer:** quilt
**Just Ask [1]:** 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 [1]:** 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:** A0

- **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
**Fully-supervised results**

- Freezing the BiLM also helps in the fully-supervised setting.
- Competitive performance + high parameter efficiency.

| Method            | # Trained Params | Fill-in-the-blank | Open-ended | Multiple-choice |
|-------------------|------------------|-------------------|------------|-----------------|
|                   |                  | LSMDC             | iVQA MSRVT-QA | MSV-QA ActivityNet-QA | TVQQA |
| HCRN [45]         | 44M              | —                 | 35.4       | 36.8            | 57.9  | 71.4* |
| HERO [54]         | 119M             | —                 | —          | —               | —     | 74.1* | 73.6* |
| ClipBERT [48]     | 114M             | —                 | 37.4       | —               | —     | —     | —     |
| Just Ask [102]    | 157M             | 35.4              | 41.8       | 47.5            | 39.0  | —     | —     |
| SiaSamRea [107]   | —                | —                 | 41.6       | 45.5            | 39.8  | 60.2  | 84.1  |
| MERLOT [109]      | 223M             | 52.9              | —          | 41.4            | —     | 69.5  | —     | 78.7* |
| Reserve [110]     | 644M             | —                 | —          | —               | —     | —     | 86.1* |
| VIOLET [21]       | 198M             | 53.7              | 43.9       | 47.9            | —     | 68.9  | —     | —     |
| All-in-one [93]   | 110M             | —                 | 46.8       | 48.3            | —     | 66.3  | —     | —     |
| *UnFrozenBiLM* (Ours) | 890M           | 58.9*             | 45.0*      | 53.9*           | 43.2* | 66.9  | 87.5* | 79.6* |
| *FrozenBiLM w/o speech* (Ours) | 30M         | 58.6              | 39.7       | 47.0            | 54.4  | 43.2  | 68.6  | 81.5  | 57.5 |
| *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. * denotes results obtained with speech input.
Few-shot results

- Significant improvements over zero-shot when using 1% of the downstream training data for finetuning.

| Supervision          | Fill-in-the-blank | Open-ended | Multiple-choice |
|----------------------|-------------------|------------|-----------------|
|                      | LSMDC | iVQA | MSRVTT-QA | MSVD-QA | ActivityNet-QA | TGIF-QA | How2QA | TVQA |
| 1. 0% (zero-shot)    | 51.5  | 26.8 | 16.7      | 33.8    | 25.9        | 41.9    | 58.4   | 59.7 |
| 2. 1% (few-shot)     | 56.9  | 31.1 | 36.0      | 46.5    | 33.2        | 55.1    | 71.7   | 72.5 |
| 3. 10% (few-shot)    | 59.9  | 35.3 | 41.7      | 51.0    | 37.4        | 61.2    | 75.8   | 77.6 |
| 4. 100% (fully-supervised) | 63.5  | 39.6 | 47.0      | 54.8    | 43.2        | 68.6    | 86.7   | 82.0 |

Table 7: Few-shot results, by finetuning FrozenBiLM using a small fraction of the downstream training dataset.