

# Zero-Shot VideoQA

Answer questions about videos without training on manual annotation of visual data.

## Motivation

> A strong approach relies on frozen autoregressive language models [1] but requires large models.

Recent work in natural language [2] shows strong zero-shot results with lighter bidirectional masked language models (BiLM).

### Contributions

**FrozenBiLM:** a framework that handles multimodal inputs using frozen BiLM and enables zeroshot VideoQA through masked language modeling.

Extensive ablation studies and improvements over previous autoregressive models.

SoTA on 8 zero-shot benchmarks, competitive performance on fully-supervised benchmarks and promising few-shot results.

### **Code and models:**

https://github.com/antoyang/FrozenBiLM



## References

M. Tsimpoukelli et al., Multimodal few-shot learning with frozen language models. In NeurIPS 2021. [2] T. Shick, et. al., It's not just size that matters: small language models are also few-shot learners. In NAACL 2021.

- [3] A. Radford, et. al., Learning transferable visual models from natural language supervision. In arXiv, 2021. [4] A. Yang, et. al., Learning to answer visual questions from Web videos. In TPAMI, 2022.
- [5] R. Zellers, et. al., Merlot reserve: neural script knowledge through vision, language and sound. In CVPR, 2022. [6] TJ. Fu et al., Violet: end-to-end video-language transformers with masked visual token modeling. In arXiv 2021.
- [7] AJ. Wang et al, All in one: exploring unified video-language pretraining. In arXiv 2022. [8] R. Zellers et al. Merlot: multi-modal neural script knowledge models. In NeurIPS 2021.

# Zero-Shot Video Question Answering via Frozen Bidirectional Language Models

Antoine Yang<sup>1,2</sup>, Antoine Miech<sup>3</sup>, Josef Sivic<sup>4</sup>, Ivan Laptev<sup>1,2</sup>, Cordelia Schmid<sup>1,2</sup> <sup>2</sup>Ecole Normale Supérieure, PSL <sup>1</sup>Inria Paris <sup>3</sup>DeepMind <sup>4</sup>CIIRC, CTU in Prague



- **Frozen parameters:** pretrained BiLM and visual backbone.
- > Trained parameters: visual-to-text projection, adapters, layer

# **Cross-modal Training**

- **Loss:** visually-conditioned masked language modeling.
- **Data:** videos with alt-text description from WebVid10M.

## Downstream Task Adaptation —

### Input prompts:

- **Open-ended VideoQA:** "[CLS] Question: <Question>? Answe [MASK]. Subtitles: <Subtitles> [SEP]"
- Multiple-choice VideoQA: "[CLS] Question: <Question>? Is '<Answer Candidate>'? [MASK]. Subtitles: <Subtitles> [Si
- Video-conditioned fill-in-the-blank: "[CLS] <Sentence with a [MASK] token>. Subtitles: <Subtitles> [SEP]"
- Answer embedding module: maps a masked token to an ans using the frozen masked language modeling head.
- Fully-supervised finetuning: standard cross-entropy loss wh keeping the same weights frozen.

# **BiLM vs Autoregressive LM**

Method		Params	iVQA	MSRVTT-C	QA MSVD-C	QA Activ	vityNet-QA	TGIF-QA
Autoregressive (GF	⊃T-J)	6B	21.4	9.6	26.7		24.5	37.3
Bidirectional (BER	T-Base)	110M	12.4	6.4	11.7		16.7	23.1
Bidirectional (DeBE	ERTa-V2-XLarge)	890M	27.3	16.8	32.2		24.7	41.0
		- Cor	mparis	son to S	Sota —			
Zero-shot result	S.		•					
Method	LSMDC i\	VQA N	ISRVTT-QA	MSVD-QA	ActivityNet-QA	A TGIF-Q	A How20	QA TVC
SoTA	31.0 [5] 13	3.3 [4]	5.8 [5]	13.5 [4]	12.3 [4]	3.6 [3]	53.1 [	4] 26.1
FrozenBiLM	51.5 2	26.8	16.7	33.8	25.9	41.9	58.4	<b>59</b> .
Question: W holding at th GT Answer: g	JAKCOFACUS       JAKCOFACUS         J hat is the man       JAKCOFACUS         I hat is the man <th>on: What item har I features a tree? wer: quilt</th> <th>nging on Question GT Answ</th> <th>n: What is the sitting ng? ver: knit sweater</th> <th>Question: Where is the work sitting on? GT Answer: camel</th> <th>oman Question: cabinet do GT Answer</th> <th>What is the color of th or in the video? r: red</th> <th>he</th>	on: What item har I features a tree? wer: quilt	nging on Question GT Answ	n: What is the sitting ng? ver: knit sweater	Question: Where is the work sitting on? GT Answer: camel	oman Question: cabinet do GT Answer	What is the color of th or in the video? r: red	he
<ul> <li>Cuestion: We holding at the fortune of the fortune of</li></ul>	What is the man he start of the video? guitar, electric guitar bewriter LM: beer (text-only): scissors (ours): guitar text-only: scissors(ours): guitar	on: What item hai Il features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): 1 BiLM (ours): quilt <b>5. Note</b>	nging on Question man doir GT Answ Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi	n: What is the sitting ng? ver: knit sweater : tie cow mBiLM: swimming LM (text-only): eating LM (ours): knit sweater : tenBiLM or	Question: Where is the work sitting on? GT Answer: camel Just Ask: horse yard UnFrozenBiLM: desert FrozenBiLM (text-only): ch FrozenBiLM (ours): camel	oman Question: cabinet do GT Answer Just Ask: d UnFrozenBiLM FrozenBiLM FrozenBiLM	What is the color of th or in the video? r: red resser BiLM: blue M (text-only): black M (ours): red	ne
<ul> <li>Guestion: W holding at th GT Answer: Just Ask: typ UnFrozenBil FrozenBil KrozenBil Mothod</li> </ul>	Anatis the man he start of the video? guitar, electric guitar pewriter LM: beer (text-only): scissors (ours): guitar FrozenE TrozenE	on: What item hai Il features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): T BiLM (ours): quilt S. Note	nging on Question man doir GT Answ Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi	n: What is the sitting ng? ver: knit sweater : tie cow mBiLM: swimming LM (text-only): eating LM (ours): knit sweater CenBiLM or MSVD-QA	Question: Where is the work sitting on? GT Answer: camel Just Ask: horse yard UnFrozenBiLM: desert FrozenBiLM (text-only): ch FrozenBiLM (ours): camel	oman Question: cabinet do GT Answer Just Ask: d UnFrozenBiLM FrozenBiLM FrozenBiLM	What is the color of th or in the video? r: red resser BiLM: blue V (text-only): black V (ours): red eters. How2QA	ne
<ul> <li>Guestion: W holding at th GT Answer: Just Ask: typ UnFrozenBil FrozenBil FrozenBil Mothod</li> <li>Method</li> <li>SoTA</li> </ul>	Just Asi During Participation Particip	on: What item hai Il features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): 1 BiLM (ours): quilt S. Note 'QA MS 4 [4]	nging on Question man doir GT Answ Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi Additional State of the state o	n: What is the sitting ng? ver: knit sweater tie cow mBiLM: swimming LM (text-only): eating LM (ours): knit sweater CONBILMON ABSVD-QA 48.3 [7]	Question: Where is the work sitting on? GT Answer: camel Just Ask: horse yard UnFrozenBiLM: desert FrozenBiLM (text-only): chr FrozenBiLM (ours): camel <b>Ny trains 30N</b> ActivityNet-QA 41.4 [8]	air Alignment District of the second secon	What is the color of the or in the video? r: red resser BiLM: blue V (text-only): black V (ours): red eters. How2QA 85.3 [4] 8	ne TVQA 36.1 [5]
Ruestion: W holding at th GTAnswer: Just Ask: typ UnFrozenBil FrozenBilM FozenBilM SoTA FrozenBilM	Just Asi Current Start of the video? guitar, electric guitar Dewriter LM: beer (text-only): scissors (ours): guitarQuestic the wal GT Ansy Just Asi UnFroze Frozeni Frozeni Frozeni Stored Stored Frozeni Stored Stored Frozeni Stored <b< th=""><th>on: What item hai Il features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): BiLM (ours): quilt S. Note 'QA MS 4 [4] 4 9.6</th><th>ree Cuestion man doir GT Answ Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi 46.8 [7] 47.0</th><th>n: What is the sitting ng? ver: knit sweater tie cow mBiLM: swimming LM (text-only): eating LM (ours): knit sweater CONBILMON ABSVD-QA 48.3 [7] 54.8</th><th>Question: Where is the work sitting on? GT Answer: camel Just Ask: horse yard UnFrozenBiLM: desert FrozenBiLM (text-only): ch FrozenBiLM (ours): camel ActivityNet-QA 41.4 [8] 43.2</br></th><th>air Alparame Alpa. 5 [8] 68.6</br></br></br></th><th>What is the color of the or in the video? r: red resser BiLM: blue V (text-only): black V (ours): red eters. How2QA 85.3 [4] 8 86.7</br></br></br></br></br></br></br></br></br></br></th><th>ne TVQA 36.1 [5] 82.0</br></br></br></th></b<>	on: What item hai Il features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): BiLM (ours): quilt S. Note 'QA MS 4 [4] 4 9.6	ree Cuestion man doir GT Answ Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi 46.8 [7] 47.0	n: What is the sitting ng? ver: knit sweater tie cow mBiLM: swimming LM (text-only): eating LM (ours): knit sweater CONBILMON ABSVD-QA 48.3 [7] 54.8	Question: Where is the work sitting on? GT Answer: camel Just Ask: horse yard UnFrozenBiLM: desert FrozenBiLM (text-only): ch FrozenBiLM (ours): camel ActivityNet-QA 41.4 [8] 	air 	What is the color of the 	ne 
Cuestion: We holding at the GT Answer: Just Ask: type UnFrozenBilt FrozenBiltM FrozenBiltM     SoTA FrozenBiltM     SoTA     FrozenBiltM	Analis the man he start of the video? guitar, electric guitar pewriter LM: beer (text-only): scissors (ours): guitar Cours):	on: What item har Il features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): BiLM (ours): quilt S. NOte 'QA MS 4 [4] 4 9.6 F	anging on Ausstion man doir GT Answ Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi AT.0	n: What is the sitting ng? ver: knit sweater tie cow mBiLM: swimming LM (text-only): eating LM (ours): knit sweater CONBILM ON ABSVD-QA 48.3 [7] 54.8 Solution ABSVD-QA ABSS [7] 54.8	ActivityNet-QA 41.4 [8] 43.2	air Alparame Alpasion: Cabinet do GT Answer Just Ask: d UnFrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM	What is the color of the or in the video? r: red resser BiLM: blue V (text-only): black V (ours): red eters. How2QA 85.3 [4] & 86.7	ne TVQA 36.1 [5] 82.0
Cuestion: Wholding at the Andrews: Bust Ask: type UnFrozenBil Method SoTA FrozenBilLM FrozenBilLM	Vhat is the man he start of the video? guitar, electric guitar bewriter LM: beer (text-only): scissors (ours): guitar <b>d benchmarks</b> <b>d benchmarks</b> <b>d benchmarks</b> <b>53</b> .7 [6] 35. 63.5 3? 63.5 3?	on: What item hai Il features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): BiLM (ours): quilt S. Note 'QA MS 4 [4] 4 9.6 G. F	anging on Auguestion man doir GT Answ Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi AT.0 AT.0 COMPASI Compassion	A svd-QA	Question: Where is the work sitting on? GT Answer: came! Just Ask: horse yard UnFrozenBiLM: desert FrozenBiLM (text-only): ch FrozenBiLM (ours): came! ActivityNet-QA 41.4 [8] 43.2 A3.2 A3.2	air Cuestion: Cabinet do GT Answer Just Ask: d UnFrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM	What is the color of th or in the video? r: red resser BiLM: blue V (text-only): black V (ours): red eters. How2QA 85.3 [4] & 86.7	ne TVQA 36.1 [5] 82.0
Cuestion: W holding at th GT Answer: Just Ask: typ UnFrozenBiLM FrozenBiLM FrozenBiLM SoTA FrozenBiLM • New setting usin Supervision 0% (zero-shot)	Vhat is the man he start of the video? guitar, electric guitar pewriter LM: beer (text-only): scissors (ours): guitar <b>d benchmarks</b> <b>d benchmarks</b> <b>5</b> 3.7 [6] 35. 63.5 3! <b>ng a small fra</b> LSMDC iV 51.5 2	on: What item hai Il features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): BiLM (ours): quilt S. NOte 'QA MS 4 [4] 4 9.6 9.6 F ction of /QA MS	anging on Auguestion Auguestion Man doin GT Answ Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi AT.0 AT.0 CON-SI C	n: What is the sitting ng? ver: knit sweater tie cow nBiLM: swimming LM (text-only): eating LM (ours): knit sweater CenBiLM or A8.3 [7] 48.3 [7] 54.8 Stas S	ActivityNet-QA 41.4 [8] 43.2	air Cuestion: Cabinet do GTAnswer Just Ask: d UnFrozenBiLM FrozenBiLM	What is the color of the or in the video? r: red resser BiLM: blue V (text-only): black V (ours): red eters. How2QA 85.3 [4] & 86.7 86.7 How2QA B5.3 [4] &	ne TVQA 36.1 [5] 82.0 82.0
Cuestion: W holding at th GT Answer: Just Ask: typ UnFrozenBiLM FrozenBiLM FrozenBiLM SoTA FrozenBiLM • New setting usin Supervision 0% (zero-shot) 1% (few-shot)	Vhat is the man he start of the video? guitar, electric guitar bewriter LM: beer (text-only): scissors (ours): guitar <b>d benchmarks</b> <b>d benchmarks</b> <b>53</b> .7 [6] 35. 63.5 3! <b>ng a small fra</b> LSMDC iV 55.7 2 56.9 3	on: What item hail l features a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): BiLM (ours): quilt S. NOte S. Note 'QA MS 4 [4] 4 9.6 9.6 F Ction of /QA MS :6.8 :1.1	anging on anging on and oir <b>Cuestion</b> man doir <b>GT</b> Answe Just Ask: UnFrozen FrozenBi FrozenBi FrozenBi AT.O AT.O COV-SI	n: What is the sitting ng? ver: knit sweater tie cow mBiLM: swimming LM (text-only): eating LM (ours): knit sweater CenBiLM or A8.3 [7] 54.8 COT CESU A8.3 SA.8 SA.8	ActivityNet-QA 41.4 [8] 43.2 Just Ask: horse yard ActivityNet-QA ActivityNet-QA A3.2	air Cuestion: cabinet do GT Answer Just Ask: d UnFrozenBiLM FrozenBiLM FrozenBiLM FrozenBiLM A DATA 69.5 [8] 68.6 68.6 68.6	What is the color of th or in the video? r: red resser BILM: blue Y (text-only): black Y (ours): red ters. How2QA 85.3 [4] & 86.7 B6.7 How2QA 58.4 58.4 71.7	ne TVQA 36.1 [5] 82.0 XVQA 59.7 59.7 72.5
Cuestion: W holding at th GT Answer: Just Ask: typ UnFrozenBiLM FrozenBiLM FrozenBiLM SoTA FrozenBiLM New setting usin Supervision 0% (zero-shot) 1% (few-shot) 10% (few-shot)	Vhat is the man re start of the video? guitar, electric guitar pewriter LM: beer (text-only): scissors (ours): guitar d benchmarks LSMDC iV 53.7 [6] 35. 63.5 3 ng a small fra LSMDC iV 551.5 2 56.9 3 59.9 3	on: What item hail lifeatures a tree? wer: quilt k: christmas sock enBiLM: fabric BiLM (text-only): BiLM (ours): quilt S. NOte 'QA MS 4 [4] 4 9.6 9.6 f ction of /QA MS 26.8 31.1 5.3	anging on reging on <b>Question</b> <b>Construction</b> <b>Question</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b> <b>Construction</b>	A stream da MSVD-QA A standa A stream da MSVD-QA A stream da MSVD-QA A stream da MSVD-QA A stream da A stream da	ActivityNet-QA 41.4 [8] 43.2 ActivityNet-QA ActivityNet-QA 43.2 ActivityNet-QA 33.2 37.4	etuning. All parame All parame FrozenBiLM	What is the color of the or in the video? r: red resser BILM: blue M (text-only): black M (ours): red eters. How2QA 85.3 [4] & 86.7 B6.7 How2QA 58.4 71.7 58.4 71.7 75.8	ne TVQA 36.1 [5] 82.0 TVQA 59.7 59.7 72.5 77.6





