Zero-Shot Video Question Answering via Frozen Bidirectional Language Models Supplementary Material

Antoine Yang[1,2], **Antoine Miech**[3], **Josef Sivic**[4], **Ivan Laptev**[1,2], **Cordelia Schmid**[1,2]
[1]Inria Paris [2]Département d'informatique de l'ENS, CNRS, PSL Research University
[3]DeepMind [4]CIIRC CTU Prague
https://antoyang.github.io/frozenbilm.html

In this Supplementary Material, we present the following items:

- *(i)* Additional qualitative examples of zero-shot VideoQA predictions (Section 1)
- *(ii)* A qualitative analysis of the *frozen* self-attention patterns in *FrozenBiLM* (Section 2)
- *(iii)* Additional information about our experimental setup (Section 3), including datasets (Section 3.1) and implementation details (Section 3.2)
- *(iv)* Additional experimental results (Section 4), including a comparison to BLIP [16] in their zero-shot VideoQA settings (Section 4.1), results on zero-shot image-VQA (Section 4.2), detailed zero-shot VideoQA results segmented per question type (Section 4.3), zero-shot results with different random seeds (Section 4.4), additional ablation studies in few-shot settings (Section 4.5), zero-shot settings (Sections 4.6 and 4.7) and fully-supervised settings (Section 4.8)

1 Qualitative examples for zero-shot VideoQA

To complement the qualitative examples shown in Figure 3 of the main paper, Figure 1 and the video *video_examples.mp4* illustrate additional qualitative results of zero-shot VideoQA for our *FrozenBiLM* model and compares them to Just Ask [30], as well as to variants of our approach that do not *freeze* the language model (*UnFrozenBiLM*) and use no visual modality (text-only), as evaluated in Section 4.2 of the main paper. Consistently with the analysis done in Section 4.4 in the main paper, we observe that the *unfrozen* variant can predict answers that lack text-only commonsense reasoning, *e.g.* in the first example of Figure 1b, the word *follow* is grammatically incorrect; in the second example of Figure 1b, it is unlikely that a singer *plays* a toad. The text-only variant does have strong language understanding, but makes visually-unrelated predictions. In contrast, consistently with our quantitative results (see Tables 1, 2 and 5 from the main paper), our model *FrozenBiLM* is able to correctly answer various questions in the diverse VideoQA paradigms (open-ended VideoQA, video-conditioned fill-in-the-blank, multiple-choice VideoQA), showing both a strong textual commonsense reasoning and a complex multi-modal understanding.

Our zero-shot model still underperforms compared to VideoQA-supervised models (see Table 7 of the main paper) and we analyze its failure cases in Figure 1a. Qualitatively, we find that the zero-shot model can fail on examples requiring complex temporal or spatial understanding *e.g.* in the third example of the second row, the model does not detect a toaster behind the woman; in the second example of the second row, it gets confused as the person browses through many different tabs from their phone. It can also be semantically inaccurate, as in the first example of the second row, the model confuses a restaurant with a bakery; in the fourth example of the second row, it confuses a chicken with another kind of bird.

[4]Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University in Prague.

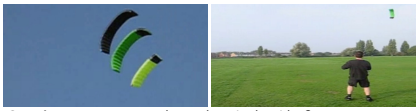
36th Conference on Neural Information Processing Systems (NeurIPS 2022).

(a) **Zero-Shot open-ended VideoQA.** The first row illustrates successful predictions on the iVQA dataset [29] (leftmost example) and the ActivityNet-QA dataset [31] (three rightmost examples). The second row illustrates incorrect predictions on the iVQA dataset.

Question: What are the men standing in front of?
GT Answer: fireplace
***Just Ask*:** cabinets
***UnFrozenBiLM*:** kitchen
***FrozenBiLM (text-only)*:** building
***FrozenBiLM (ours)*:** fireplace

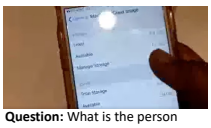
Question: Which category of sports does this sport belong to?
GT Answer: surfing
***Just Ask*:** second
***UnFrozenBiLM*:** swimming
***FrozenBiLM (text-only)*:** 1
***FrozenBiLM (ours)*:** surfing

Question: Is there green grass on the roof?
GT Answer: yes
***Just Ask*:** no
***UnFrozenBiLM*:** no
***FrozenBiLM (text-only)*:** no
***FrozenBiLM (ours)*:** yes

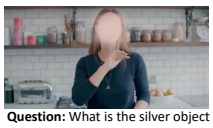


Question: How many people are there in the video?
GT Answer: 1
***Just Ask*:** 2
***UnFrozenBiLM*:** 4
***FrozenBiLM (text-only)*:** 2
***FrozenBiLM (ours)*:** 1

Question: What did the man with the backpack walk into?
GT Answer: bakery, bake shop
***Just Ask*:** stores
***UnFrozenBiLM*:** wall
***FrozenBiLM (text-only)*:** water
***FrozenBiLM (ours)*:** restaurant



Question: What is the person changing on the phone?
GT Answer: settings
***Just Ask*:** colors
***UnFrozenBiLM*:** camera
***FrozenBiLM (text-only)*:** phone
***FrozenBiLM (ours)*:** wallpaper

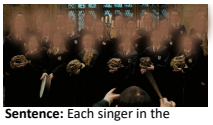


Question: What is the silver object behind the woman on counter?
GT Answer: toaster
***Just Ask*:** mirror
***UnFrozenBiLM*:** salt
***FrozenBiLM (text-only)*:** coin
***FrozenBiLM (ours)*:** spoon

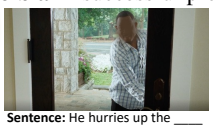
Question: What organism is shown at the end of the video?
GT Answer: bird
***Just Ask*:** worms
***UnFrozenBiLM*:** beef
***FrozenBiLM (text-only)*:** octopus
***FrozenBiLM (ours)*:** chicken

(b) **Zero-shot video-conditioned fill-in-the-blank** successful predictions on the LSMDC-FiB dataset [19].

Sentence: Someone _____ him to the truck and across the street.
GT Answer: chases
***UnFrozenBiLM*:** follow
***FrozenBiLM (text-only)*:** drags
***FrozenBiLM (ours)*:** chases



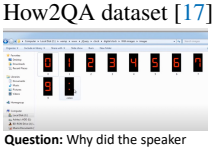
Sentence: Each singer in the front row ____ a huge toad.
GT Answer: holds
***UnFrozenBiLM*:** plays
***FrozenBiLM (text-only)*:** wears
***FrozenBiLM (ours)*:** holds



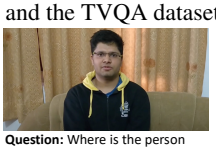
Sentence: He hurries up the ____ walkway to his house and enters.
GT Answer: front
***UnFrozenBiLM*:** screen
***FrozenBiLM (text-only)*:** wooden
***FrozenBiLM (ours)*:** front

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

(c) **Zero-shot multiple-choice VideoQA.** The first and second rows illustrate successful predictions on the How2QA dataset [17] and the TVQA dataset [15], respectively.



Question: Why did the speaker opened a folder on his computer?
A0: to show pictures of digital numbers
A1: to show photographs he has taken
A2: to show desktop wallpapers
A3: to show programs he downloaded
GT Answer: A0
***UnFrozenBiLM*:** A2
***FrozenBiLM (text-only)*:** A1
***FrozenBiLM (ours)*:** A0



Question: Where is the person in the clip most likely located?
A0: home
A1: corporate office
A2: sports stadium
A3: emergency room
GT Answer: A0
***UnFrozenBiLM*:** A3
***FrozenBiLM (text-only)*:** A2
***FrozenBiLM (ours)*:** A0

Question: What is the man doing to the branches?
A0: He is burning them.
A1: He is burying them.
A2: He is throwing them in water.
A3: He's painting them.
GT Answer: A0
***UnFrozenBiLM*:** A3
***FrozenBiLM (text-only)*:** A2
***FrozenBiLM (ours)*:** A0

Question: When did the chef flipped over the layer of rice and seaweed?
A0: after she sprinkled sesame
A1: after she added cucumber
A2: after she added fish
A3: after she cut the cucumbers
GT Answer: A0
***UnFrozenBiLM*:** A3
***FrozenBiLM (text-only)*:** A1
***FrozenBiLM (ours)*:** A0

Question: Where is the man with glasses after Dr Lisa Cuddy leaves the room?
A0: Leaning against the bookcase
A1: Sitting on a white chair
A2: Standing behind Dr House
A3: Laying on the floor next to the desk
A4: Sitting in a wheel chair
GT Answer: A1
***UnFrozenBiLM*:** A0
***FrozenBiLM (text-only)*:** A3
***FrozenBiLM (ours)*:** A1

Question: What adjustement does Beckett do before going to talk with Mr caraway?
A0: She puts on lipstick
A1: She puts on glasses
A2: She ties back her hair
A3: She changes into a skirt
A4: She zips up her jacket
GT Answer: A4
***UnFrozenBiLM*:** A2
***FrozenBiLM (text-only)*:** A2
***FrozenBiLM (ours)*:** A4

Question: What color was the bowl beside the stove when Robin was making crepes?
A0: Orange
A1: Red
A2: White
A3: Blue
A4: Green
GT Answer: A4
***UnFrozenBiLM*:** A0
***FrozenBiLM (text-only)*:** A3
***FrozenBiLM (ours)*:** A4

Question: What did Raj do after he discovered the wine bottle was empty?
A0: Raj laughed out loud
A1: Raj called Howard on the phone
A2: Raj put the bottle down and got cake to eat from the refrigerator
A3: Raj ran in a circle
A4: Raj went to the bathroom
GT Answer: A2
***UnFrozenBiLM*:** A1
***FrozenBiLM (text-only)*:** A3
***FrozenBiLM (ours)*:** A2

Figure 1: **Zero-Shot VideoQA.** Qualitative comparison between Just Ask [30] (row 3 in Table 5 of the main paper), our model (row 4 in Table 5 of the main paper), its *unfrozen* variant (row 2 in Table 1 of the main paper) and its text-only variant (row 2 in Table 2 of the main paper), for zero-shot VideoQA. The last column of each row illustrates a single video example with two frames, while other columns illustrate each video example with one frame. We show more examples on our webpage [1].

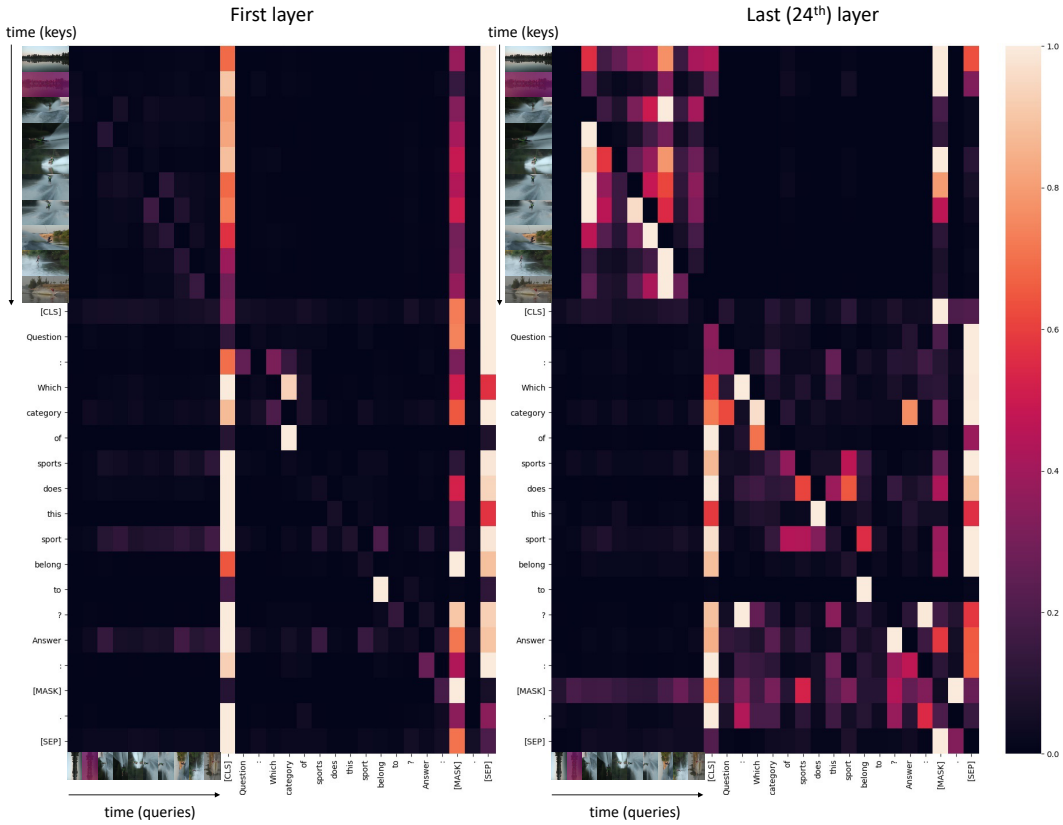


Figure 2: ***FrozenBiLM* self-attention visualization for zero-shot VideoQA.** Visualization of the attention weights between the different visual tokens from the video prompt and the textual tokens from the text embedder, for the second example of the first row in Figure 1. A column corresponds to the weights of the different visual and text tokens for the given token. These attention weights are averaged across all 24 heads, and renormalized by the maximum weight for each token (*i.e.* each column) for the purpose of visualization. Lighter colors correspond to higher attention weights (see the colorbar on the right). In the first layers (left), we observe that the multi-modal interactions mainly flow through the [CLS], [MASK] and [SEP] tokens, and that there is little interaction between the different visual tokens. In the last layers (right), we observe that visual tokens attend to each other and the [MASK] token attends to the visual tokens, while the [CLS] and [SEP] tokens mainly attend to text tokens. Note that the self-attention weights are *frozen* after text-only pretraining.

2 Qualitative analysis of the *frozen* self-attention patterns in *FrozenBiLM*

We show in Section 4.2 of the main paper that the visual modality is crucial for the zero-shot VideoQA performance. Here we further analyze qualitatively *how*, for zero-shot VideoQA, our model makes use of the visual modality through self-attention layers which are *frozen* after text-only pretraining. Figure 2 illustrates the self-attention patterns in *FrozenBiLM* for the second example in the first row of Figure 1. Despite the freezing, we observe that these layers actually enable visual-linguistic interactions. Indeed, in the first layer (Figure 1, left), the [CLS], [MASK] and [SEP] tokens significantly attend to the visual tokens. Moreover, we observe substantially different patterns in the last layer (Figure 1, right): while the [MASK] token still attends to visual tokens, the different visual tokens at different timesteps attend between each other and the [CLS] and [SEP] tokens mainly attend to other text tokens. Consistently with results presented in Section 4.2 of the main paper, this qualitative analysis suggests that the *frozen* self-attention layers in *FrozenBiLM* do enable visual-linguistic interactions.

3 Experimental setup

In this section we first present additional information on the used datasets (Section 3.1) and then describe implementation details (Section 3.2).

3.1 Datasets

In this section, we give further details about the downstream datasets we use. Their licenses are mentioned in our code in the separate folder *code*.

LSMDC-FiB [19] is an open-ended video-conditioned fill-in-the-blank task which consists in predicting masked words in sentences that describe short movie clips [22, 23]. It contains 119K video clips and 349K sentences, split into 297K/22K/30K for training/validation/testing.

iVQA [29] is a recently introduced open-ended VideoQA dataset, focused on objects, scenes and people in instructional videos [20]. It excludes non-visual questions, and contains 5 possible correct answers for each question for a detailed evaluation. It contains 10K video clips and 10K questions, split into 6K/2K/2K for training/validation/testing.

MSRVTT-QA [27], **MSVD-QA** [27] and **TGIF-FrameQA** [9] are popular open-ended VideoQA benchmarks automatically generated from video descriptions [4, 18, 28]. Questions are of five types for MSRVTT-QA and MSVD-QA: what, who, how, when and where; and four types for TGIF-QA: object, number, color and location. MSRVTT-QA contains 10K video clips and 243K question-answer pairs, split into 158K/12K/73K for training/validation/testing. MSVD-QA contains 1.8K video clips and 51K question-answer pairs, split into 32K/6K/13K for training/validation/testing. TGIF-QA contains 46K GIFs and 53K question-answer pairs, split into 39K/13K for training/testing.

ActivityNet-QA [31] is an open-ended VideoQA dataset consisting of long videos [3] (3 minutes long on average), and covering 9 question types (motion, spatial, temporal, yes-no, color, object, location, number and other). It contains 5.8K videos and 58K question-answer pairs, split into 32K/18K/8K for training/validation/testing.

How2QA [17] is a multiple-choice VideoQA dataset focused on instructional videos [20]. Each question is associated with one correct and three incorrect answers. It contains 28K video clips and 38K questions, split into 35K/3K for training/validation.

TVQA [15] is a multiple-choice VideoQA dataset focused on popular TV shows. Each question is associated with one correct and four incorrect answers. It contains 22K video clips and 153K questions, split into 122K/15K/15K for training/validation/testing. The test set is hidden and only accessible a limited number of times via an online leaderboard.

3.2 Implementation details

Architecture hyperparameters. We truncate text sequences up to $L = 256$ tokens. Video features are extracted by sampling $T = 10$ frames, each resized at 224×224 pixels, from the video. These frames are sampled at temporally equal distance, with a minimum distance of 1 second. For videos shorter than T seconds, we pad the video prompt up to T tokens. The dimension of the visual features from ViT-L/14 [6] is $D_f = 768$. The transformer encoder from DeBERTa-V2-XLarge [7] has 24 layers, 24 attention heads, a hidden dimension of $D = 1536$ and an intermediate dimension in the feed-forward layers of 6144. For the adapters [8], we use a bottleneck dimension of $D_h = \frac{D}{8} = 192$.

Training. For all training experiments, we use the Adam optimizer [12] with $\beta = (0.9, 0.95)$ and no weight decay. We use Dropout [25] with probability 0.1 in the adapters and in the transformer encoder. When finetuning the language model weights, we divide the batch size by a factor 2 so to accommodate with the GPU memory constraints.

Cross-modal training. To train on WebVid10M, we use a total batch size of 128 video-caption pairs split in 8 NVIDIA Tesla V100 GPUs. We use a fixed learning rate of $3e^{-5}$ for the variant with adapters. We find that the variant without adapters that freezes the language model weights prefers a higher learning rate of $3e^{-4}$, and that the variant *UnfrozenBiLM* that finetunes the language model weights prefers a lower one of $1e^{-5}$.

Method	Pretraining Data	Finetuning Data	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA
BLIP [16]	129M image-text pairs	VQA	—	19.2	35.2	—	—
FrozenBiLM (no image-VQA training)	WebVid10M	∅	26.8	16.7	33.8	25.9	41.9
FrozenBiLM (no cross-modal training)	∅	VQA	14.6	6.9	12.6	22.6	33.3
FrozenBiLM (Ours)	WebVid10M	VQA	**34.6**	**22.2**	**39.0**	**33.1**	**43.4**

Table 1: Results of our model after cross-modal training, finetuning on the open-ended image-VQA dataset [2] and directly evaluating on open-ended VideoQA without using any VideoQA supervision, as in BLIP [16].

Method	Motion	Spatial	Temporal	Yes-No	Color	Object	Location	Number	Other
Just Ask [29]	2.3	1.1	0.3	36.3	11.3	4.1	6.5	0.2	4.7
FrozenBiLM	**12.7**	**6.8**	**1.6**	**53.2**	**16.5**	**17.9**	**18.1**	**26.2**	**25.8**

Table 2: Zero-shot VideoQA results segmented per question type on the ActivityNet-QA dataset, compared with Just Ask [29].

Method	MSRVTT-QA						MSVD-QA					
	What	Who	Number	Color	When	Where	What	Who	Number	Color	When	Where
Just Ask [29]	1.8	0.7	**66.3**	0.6	0.6	4.5	7.8	1.7	**74.3**	18.8	3.5	0.0
FrozenBiLM	**10.7**	**28.7**	55.0	**11.4**	**9.2**	**9.3**	**26.0**	**45.0**	69.9	**56.3**	**5.2**	**17.9**

Table 3: Zero-shot VideoQA results segmented per question type on the MSRVTT-QA dataset (left) and the MSVD-QA dataset (right), compared with Just Ask [29].

Downstream task finetuning. To finetune our model on downstream datasets, we use a total batch size of 32 video-question-answer triplets (respectively 32 video-sentence pairs) split in 4 NVIDIA Tesla V100 GPUs for open-ended VideoQA datasets (respectively video-conditioned fill-in-the-blank datasets) and 16 video-question pairs split in 8 NVIDIA Tesla V100 GPUs for multiple-choice VideoQA datasets. We train for 20 epochs for all downstream datasets except LSMDC-FiB for which we find that training for 5 epochs leads to similar validation results. We warm up the learning rate linearly for the first 10% of iterations, followed by a linear decay of the learning rate (down to 0) for the remaining 90%. On each dataset, we run a random search and select the learning rate based on the best validation results. We search over 10 learning rates in the range $[1e^{-5}, 1e^{-4}]$ for variants that freeze the language model weights, and $[5e^{-6}, 5e^{-5}]$ for the variant *UnfrozenBiLM* that finetunes the language model weights.

Answer vocabulary for open-ended VideoQA. In the zero-shot setting, we use an answer vocabulary composed of the top $1,000$ answers in the corresponding training dataset, following [32]. In the fully-supervised setting, we experiment both with the vocabulary composed of the top $1,000$ answers and the vocabulary composed of all answers appearing at least twice in the corresponding training dataset and choose the one leading to best validation results. Following [32], questions with out-of-vocabulary answer are not used for finetuning, and are automatically considered as incorrect during evaluation.

4 Experiments

In this section, we complement the experiments presented in Section 4 of the main paper. We first present a comparison with BLIP [16] in their zero-shot settings in Section 4.1. In Section 4.3 we show detailed zero-shot VideoQA results segmented per question category and compare our method with Just Ask [29]. Next we analyze the impact of the random seed used in the cross-modal training on the zero-shot VideoQA results in Section 4.4. We also show the importance of freezing the language model in few-shot settings in Section 4.5. We present additional ablation studies in the zero-shot setting in Section 4.7. Finally we show the benefit of cross-modal training and adapter training in fully-supervised settings in Section 4.8.

4.1 Comparison with BLIP

In addition to the zero-shot results presented in the main paper, we here investigate a different but related *zero-shot* setting defined in BLIP [16], where a network trained on manually annotated image-VQA annotations is evaluated directly on open-ended VideoQA datasets. In detail, BLIP uses

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 [21]	400M image-texts	1.2	9.2	2.1	7.2	1.2	3.6	47.7	26.3
Just Ask [30]	HowToVQA69M + WebVidVQA3M	—	13.3	5.6	13.5	12.3	—	53.1	—
Reserve [33]	YT-Temporal-1B	31.0	—	5.8	—	—	—	—	—
FrozenBiLM (Ours)	WebVid10M	**51.5±0.1**	**28.3±0.9**	**14.4±1.4**	**30.0±2.2**	**25.4±0.7**	**39.7±2.1**	**57.9±0.6**	**57.9±1.2**

Table 4: Comparison with the state of the art for zero-shot VideoQA, reporting mean and standard deviation over 5 cross-modal training runs with different random seeds. Results on TVQA are reported on the validation set given that the hidden test set can only be accessed a limited number of times.

Variant	Supervision	Fill-in-the-blank	Open-ended					Multiple-choice	
		LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
1. *UnFrozenBiLM*	0% (zero-shot)	37.1	21.0	**17.6**	31.9	20.7	30.7	45.7	45.6
2. *FrozenBiLM*	0% (zero-shot)	**51.5**	**26.8**	16.7	**33.8**	**25.9**	**41.9**	**58.4**	**59.2**
3. *UnFrozenBiLM*	1% (few-shot)	46.2	23.5	33.4	43.7	31.6	51.7	68.0	68.6
4. *FrozenBiLM*	1% (few-shot)	**56.9**	**31.1**	**36.0**	**46.5**	**33.2**	**55.1**	**71.7**	**71.8**
5. *UnFrozenBiLM*	10% (few-shot)	52.6	29.5	38.9	49.8	36.5	57.8	73.2	74.8
6. *FrozenBiLM*	10% (few-shot)	**59.9**	**35.3**	**41.7**	**51.0**	**37.4**	**61.2**	**75.8**	**77.3**
7. *UnFrozenBiLM*	100% (fully-supervised)	58.9	37.7	45.0	53.9	**43.2**	66.9	**87.5**	79.1
8. *FrozenBiLM*	100% (fully-supervised)	**63.5**	**39.6**	**47.0**	**54.8**	**43.2**	**68.6**	86.7	**82.4**

Table 5: Few-shot results, by finetuning *FrozenBiLM* using a small fraction of the downstream training dataset, compared with the variant *UnFrozenBiLM* which does not freeze the language model weights. Results on TVQA are reported on the validation set given that the hidden test set can only be accessed a limited number of times.

the open-ended image-VQA dataset [2] for finetuning after pretraining on 129M image-text pairs, including COCO [5] and Visual Genome [13] which are manually annotated. To adapt our model to this setting, we finetune our model *FrozenBiLM* pretrained on WebVid10M on the image-VQA dataset using the same procedure as for finetuning on VideoQA datasets (see Section 3.3 in the main paper), *i.e.* notably with a *frozen* language model. In particular, we finetune on VQA for 10 epochs with an initial learning rate of $1e^{-5}$ which is warmed up for the first 10% iterations, and linearly decayed to 0 for the remaining 90% iterations. Table 1 shows that the resulting model not only improves over our model without image-VQA finetuning (*i.e.* in zero-shot mode as defined in the main paper) or our model trained on VQA only (*i.e.* without cross-modal training), but also substantially outperforms BLIP on both MSRVTT-QA and MSVD-QA. These results further demonstrate the strong capabilities of *FrozenBiLM* in settings where no VideoQA annotation is available.

4.2 Results on zero-shot image-VQA

We next evaluate our pretrained model on the VQAv2 [2] validation set in the zero-shot setting, *i.e.*, without any supervision of visual questions and answers. Frozen [26] achieves 29.5% accuracy in this setting using an autoregressive language model. In comparison, our *FrozenBiLM* model is 7 times smaller than Frozen and achieves 45.0% accuracy. We conclude that our model can perform competitively on the image-VQA tasks despite being tailored for videos.

4.3 Detailed zero-shot VideoQA results segmented per question category

We complement the comparison to the state of the art for zero-shot VideoQA given in Section 4.4 of the main paper with results segmented per question type for ActivityNet-QA in Table 2, and for MSRVTT-QA and MSVD-QA in Table 3. Compared to Just Ask [29], we observe large and consistent improvements over all question categories, except for the *number* category on MSRVTT-QA and MSVD-QA. These results show that our approach is efficient in the diverse question categories of zero-shot VideoQA.

4.4 Impact of the random seed on zero-shot VideoQA

To verify the robustness of our approach with respect to the random seed, we run cross-modal training for *FrozenBiLM* with 5 different random seeds. We report the mean and standard deviation of zero-shot accuracy in Table 4, compared with state-of-the-art approaches that only report their

	Inference Strategy	Fill-in-the-blank	Open-ended				
		LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA
1.	Average token embeddings	**51.5**	26.8	16.7	33.8	25.9	41.9
2.	Multiple mask tokens	51.0	**27.0**	**17.1**	**34.4**	**26.1**	**42.0**

Table 6: Impact of the inference strategy on the zero-shot open-ended VideoQA performance.

	T	D_h	Visual	Fill-in-the-blank	Open-ended					Multiple-choice	
			Backbone	LSMDC	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	How2QA	TVQA
1.	1	192	ViT-L/14 (CLIP)	50.4	24.8	12.4	28.3	24.9	41.5	54.3	54.6
2.	10	96	ViT-L/14 (CLIP)	**52.4**	**28.6**	13.7	29.0	25.1	**42.3**	**59.3**	58.0
3.	10	384	ViT-L/14 (CLIP)	51.4	27.5	15.6	31.2	23.9	41.8	58.0	57.8
4.	10	192	ViT-B/16 (ImageNet)	49.4	23.8	13.3	25.7	25.1	36.8	56.5	57.2
5.	10	192	ViT-B/16 (CLIP)	50.8	25.5	14.6	30.3	25.6	41.0	57.6	58.2
6.	10	192	ViT-L/14 (CLIP)	51.5	26.8	**16.7**	**33.8**	**25.9**	41.9	58.4	**59.2**

Table 7: Impact of the number of frames T used by the model, the hidden dimension D_h in the adapters and the visual backbone on the zero-shot VideoQA results. All models are trained on WebVid10M and use multi-modal inputs (video, speech and question) at inference.

results based on a single run. We observe that the random seed does not affect the comparison to prior work done in Section 4.4 in the main paper, as our model improves over previous work for zero-shot VideoQA [21, 30, 33] by significant margins.

4.5 Freezing the language model is also beneficial in few-shot settings

Sections 4.2 and 4.5 in the main paper demonstrate that freezing the language model combined with training adapters outperforms finetuning the language model in the zero-shot and fully-supervised settings. In Table 5, we further show that freezing the language model combined with training adapters outperforms finetuning the language model in the few-shot setting as defined in Section 4.5 of the main paper (compare rows 3 and 4, or rows 5 and 6). Interestingly, the difference is larger when using 1% of the downstream training dataset (rows 3 and 4) compared to using 10% (rows 5 and 6) or 100% (rows 7 and 8). These results demonstrate that our approach is particularly efficient in settings where VideoQA annotations are scarce.

4.6 Ablation of the multi-token inference strategy

For multi-token answers in the open ended VideoQA setting, our *FrozenBiLM* simply averages the weights of different answer tokens. However, such simple scheme does not preserve the semantic structure of the answer. Hence we here investigate and compare another possible inference strategy in the zero-shot setting and discuss potential sources of improvement. We take inspiration from [10] and performs zero-shot VideoQA inference by using multiple mask tokens decoded in parallel. Then, for each video-question pair, we do one forward pass through the model per possible number of mask tokens (typically, 1 to 5) in order to score all possible answers in vocabulary $\mathcal{A}$. The score of a given answer is then obtained by multiplying the probability of its individual tokens, possibly normalized by its number of tokens. As shown in Table 6, we observe that such a decoding strategy (row 2) does not significantly improve the accuracy of our model over the one used in *FrozenBiLM* (row 1). We hypothesize that this is due to the fact that the current open-ended VideoQA datasets [9, 27, 29, 31] contain a great majority of short answers, e.g. 99% of the answers in the MSRVTT-QA test set are one-token long with our tokenizer [14]. Additionally, a possible solution to further improve the decoding in this alternative scheme is to increase the length of the masked spans at pretraining, as in [11]. [24] provides another potential solution to score multi-token answers in our framework, by masking tokens one by one and computing pseudo-likelihood scores.

4.7 Additional ablation studies in the zero-shot setting

We here complement zero-shot ablation studies reported in Section 4.2 of the main paper. We analyze the impact of the number of frames T used by the model, the hidden dimension in the adapters D_h and the size and pretraining of the visual backbone in Table 7. All models use the same setting as described in Section 4.2 in the main paper and detailed in Section 3. We first observe that using 10

	Template	iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA
1.	‘‘[CLS] Question: <Question>? Answer: [MASK]. Subtitles: <Subtitles> [SEP]‘‘	26.8	**16.7**	**33.8**	**25.9**	**41.9**
2.	‘‘[CLS] Q: <Question>? A: [MASK]. S: <Subtitles> [SEP]‘‘	**27.4**	16.2	32.5	25.5	**41.9**
3.	‘‘[CLS] <Question>? [MASK]. <Subtitles> [SEP]‘‘	23.1	13.6	28.0	21.6	25.2

Table 8: Impact of the prompt on the zero-shot open-ended VideoQA performance.

	Template	How2QA	TVQA
1.	‘‘[CLS] Question: <Question>? Is it ’’<Answer Candidate>’’? [MASK]. Subtitles: <Subtitles> [SEP]‘‘	**58.4**	**59.7**
2.	‘‘[CLS] Q: <Question>? Is it ’’<Answer Candidate>’’? [MASK]. S: <Subtitles> [SEP]‘‘	57.7	58.2
3.	‘‘[CLS] <Question>? <Answer Candidate>? [MASK]. <Subtitles> [SEP]‘‘	47.6	55.0

Table 9: Impact of the prompt on the zero-shot multiple-choice VideoQA performance.

	Cross-modal Training	Frozen LM	Adapters	# Trained Params	Fill-in-the-blank LSMDC	Open-ended iVQA	MSRVTT-QA	MSVD-QA	ActivityNet-QA	TGIF-QA	Multiple-choice How2QA	TVQA
1.	✓	✗	✗	890M	58.9	37.7	45.0	53.9	**43.2**	66.9	**87.5**	79.1
2.	✓	✓	✗	**1M**	60.4	38.2	43.2	51.7	38.3	66.5	79.3	78.2
3.	✗	✓	✓	30M	57.1	34.3	46.2	51.9	41.8	67.4	75.8	70.8
4.	✓	✓	✓	30M	**63.5**	**39.6**	**47.0**	**54.8**	**43.2**	**68.6**	86.7	**82.4**

Table 10: Importance of cross-modal training and training various parameters for fully-supervised VideoQA. All models are finetuned on downstream VideoQA datasets, and use multi-modal inputs (video, speech and question) at inference.

frames significantly improves over using a single frame (compare rows 1 and 5). Next we note that using a hidden dimension of 96 or 384 in the adapters instead of 192 does not change the results significantly (see rows 2, 3 and 6). Moreover, we find that scaling up the size of the visual backbone is beneficial, as using ViT-L/14 instead of ViT-B/16, both being trained on CLIP [21], slightly improves the results (compare rows 4 and 6). Furthermore, we observe that the pretraining of the visual backbone is crucial, as using ViT-B/16 pretrained on 400M image-text pairs from CLIP significantly improves over using ViT-B/16 pretrained on ImageNet-21K, *i.e.* 22M image-label pairs (compare rows 4 and 5).

Finally, we ablate the importance of the prompt design on the zero-shot VideoQA performance. We report results with alternative prompts in Tables 8 and 9. We find that replacing the words “Question”, “Answer” and “Subtitles” by “Q”, “A” and “S”, respectively, in the templates described in Section 3.3 does not impact the zero-shot VideoQA accuracy (compare rows 2 and 1 in Tables 8 and 9). However, completely removing “Question”, “Answer”, “Subtitles” and “is it” in the templates results in a significant drop of performance (compare rows 3 and 1 in Tables 8 and 9). We conclude that it is important to have tokens that link the different textual inputs.

4.8 Cross-modal training and adapters are crucial for fully-supervised performance

We have examined the impact of cross-modal training and training various parameters of our architecture on the zero-shot VideoQA performance in Section 4.2 of the main paper. In Table 10, we complement these ablation studies by analyzing the importance of cross-modal training and training various parameters for the fully-supervised VideoQA performance. For this, we train on downstream datasets a variant with no adapters, and a variant without cross-modal training, following the same procedure as explained in Section 3.3 of the main paper and detailed in Section 3. We find that cross-modal training is significantly beneficial for the fully-supervised setting (compare rows 3 and 4). Similar to conclusions made in Section 4.5 of the main paper, training adapters while freezing the language model outperforms finetuning the language model in fully-supervised settings (see rows 1 and 4). Finally, we note that training adapters has a considerable importance on the performance in fully-supervised settings (compare rows 2 and 4). These results further demonstrate the strength of our approach in the fully-supervised setup.

References

[1] FrozenBiLM project webpage. https://antoyang.github.io/frozenbilm.html.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual question answering. In *ICCV*, 2015.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[4] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011.

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *ICLR*, 2021.

[8] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *ICML*, 2019.

[9] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

[10] Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. X-factr: Multilingual factual knowledge retrieval from pretrained language models. In *EMNLP*, 2020.

[11] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. In *TACL*, 2020.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[13] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016.

[14] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *ACL*, 2018.

[15] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, 2018.

[16] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[17] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*, 2020.

[18] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A New Dataset and Benchmark on Animated GIF Description. In *CVPR*, 2016.

[19] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *CVPR*, 2017.

[20] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[22] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *CVPR*, 2015.

[23] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *IJCV*, 2017.

[24] Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. Masked language model scoring. In *ACL*, 2020.

[25] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[26] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021.

[27] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM international conference on Multimedia*, 2017.

[28] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016.

[29] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021.

[30] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Learning to answer visual questions from web videos. *IEEE TPAMI*, 2022.

[31] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *AAAI*, 2019.

[32] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. In *NeurIPS*, 2021.

[33] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. MERLOT Reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 2022.