Distilling Vision-Language Models on Millions of Videos

Supplementary Material

7. Instruction-Following Templates

We provide the templates to generate the instruction-following data in Table 12. Specifically, each prompt starts with a brief instruction followed by two examples for few-shot prompting. The examples used for prompting temporal-reasoning, causal-reasoning, and short question-answer pairs are enumerated in Table 13, Table 14, and Table 15, respectively. We randomly sample two examples out of three at each time.

8. Examples of Video Captioning

We provide some more examples of captions before and after video-specific adaptation in Figure 6. We can see that our video-language model with visual adaptation generates short and accurate captions for videos. The outcome is comparable to the one that is achieved by frame-level image captioning following by LLM-summarization. Furthermore, our video-language model with both visual and language adaptation provides more details when describing the same video.

Additionally, we show some representative statistics from the natural language understanding perspective in Table 7. We analyze the sentence length and frequency of unique verbs/nouns using Spacy [25]. Our generated video captions (VideoCC+) are nearly 50% longer than alt-text. They also contain more unique verbs and nouns than the original alt-text, indicating higher diversity.

9. Dataset Details

In this section, we summarize the datasets that we used in §5 to evaluate the video-language model and the dual-encoder model. The datasets that we used to adapt the vision-language model from images to videos and distill the resultant video-language model for pseudo-captioning have already been summarized in §5.1.

9.1. Data for Evaluating the Dual-Encoder Model

MSR-VTT [68] consists of 10K video clips with video captioning, each of which has 20 captions. We follow the 1k-A splits in [61], namely 9K/1K for training/testing, and report text-to-video retrieval (TVR) Recall@{1,5,10} on the testing split.

Kinetics-600 [6] contains around 480K 10-second video clips from 600 action classes. We follow the standard splits, namely 390K/30K/ 60K for training/validation/testing, and report top-1 accuracy on the validation split.

	Average length	# of uniq. verbs	# of uniq. nouns
VideoCC	10.66	8,000	13,097
VideoCC$^+$	15.74	8,317	29,712

Table 7. **Statistics of original alt-text and pseudo-captions.**

VATEX [60] consists of around 41K videos sampled from the Kinetics-600 dataset, each of which has 10 English captions and 10 Chinese captions. Only the English annotations are used for evaluation following [61, 69]. We follow the splits in [61], namely 26K/1.5K/1.5K for training/validation/testing, and report text-to-video retrieval (TVR) Recall@{1,5,10} on the testing split.

9.2. Data for Evaluating the Video-Language Model

MSR-VTT Captions [68] consists of 10K video clips with video captioning, each of which has 20 captions. We follow the standard splits in [68], namely 6.5K/0.5K/3K for training/validation/testing, and report captioning results measured by CIDEr score on the testing split.

ActivityNet Captions consists of 100K temporally localized sentences for 20K videos. We follow the standard splits in [28], namely 10K/5K/5K videos for training/validation/testing, and assume ground truth temporal proposals is known at evaluation. We report captioning results measured by CIDEr score on `val_2` split.

MSR-VTT-QA [66] has the same amount of videos of MSR-VTT but is augmented with 243K question-answer pairs. We follow the standard splits in [66], namely 158K/12K/73K QA pairs for training/validation/testing. We report the accuracy (using exact string match as in PaLI [10]) on the testing split.

ActivityNet-QA [74] builds upon ActivityNet and contains 58K question-answer pairs. We follow the standard splits, namely 32K/18K/8K QA pairs for training/validation/testing. We report accuracy (using exact string match as in PaLI [8, 10]) on the testing split.

NExT-OE-QA [65] is the open-ended task for NExT-QA dataset. It contains 52,044 question-answer pairs for a total of 5,440 videos. Following [65], we report Wu-Palmer Similarity (WUPS) score on the test set, which has 9.2K QA pairs.

10. Implementation Details

10.1. Adapting the Vision-Language Model

We inherit the training recipe of PaLI-3 when adapting the vision-language model from images to videos. Specifically,

Method	Pre-training Dataset	MSR-VTT TVR			VATEX TVR			Kinetics-600	
		R@1	R@5	R@10	R@1	R@5	R@10	Top-1	Top-5
InternVideo [61]	WIT→Mixed (12M)	40.0	65.3	74.1	49.5	79.7	87.0	-	
ViCLIP [63]	WIT→WebVid (10M)	35.6	-	-	-	-	-	58.7	81.0
ViCLIP [63]	WIT→InternVid (10M)	42.4	-	-	-	-	-	62.2	84.9
CLIP (ViT-*st*-L)	WIT→S-MiT	45.2	70.8	80.5	**66.7**	**92.0**	**96.2**	**64.2**	**88.8**
	WIT→VideoCC$^+$ (**Ours**)	48.2	72.2	80.8	64.2	90.2	95.1	61.1	85.6
	WIT→InternVid$^+$ (**Ours**)	46.3	71.5	80.3	65.2	91.3	95.5	62.7	86.2
	WIT→VideoCC$^+$+InternVid$^+$ (**Ours**)	**48.4**	**73.5**	**81.9**	65.6	91.7	95.8	62.8	86.4

Table 8. **Comparison of zero-shot text-to-video retrieval performance on MSR-VTT & VATEX and video recognition performance on Kinetics-600 between human-labeled and pseudo-captioned videos.** $\mathcal{D}^+$ means that the captions in the video dataset $\mathcal{D}$ are generated by our proposed pipeline. $\mathcal{D} \in \{\text{VideoCC}, \text{InternVid}\}$ in our experiments.

we use AdaFactor with $\beta_1 = 0$ and $\beta_2 = 0.8$. For the learning rate schedule, we use a linear warmup at the first 1K iteration, followed by inverse square-root decay. The peak learning rate is 10^{-4}. During visual adaptation, we use a batch size of 64 and train the model for 40K iteration, which is equivalent to $\sim$5 epochs, on 128 TPU-v5e chips. During language adaption, we use a batch size of 256 and train the model for 10K iteration, which is equivalent to $\sim$2.6 epochs, on 128 TPU-v5e chips.

10.2. Training the Dual-Encoder Model

We use SGD with momentum $\beta = 0.9$ as the optimizer by default. For the learning rate schedule, we use a linear warmup at the first 5K iterations followed by cosine decay. The peak learning rate is 10^{-3}. We use a batch size of 1,024 and train the model for 100 epochs. Besides, we observe that using the AdamW optimizer gives a faster convergence speed and a better final zero-shot performance when training the dual-encoder model with pseudo-captions. Therefore, we use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$ and weight decay of 0.01 and train the model for 20 epochs when reporting the main result in Table 2. We use the default SGD-optimizer recipe in Table 4. For data augmentation, we apply standard scale jittering augmentation with a scale range of $(0.9, 1.33)$ and take a 224×224 crop.

11. More ablations

11.1. Justifying the Two-Stage Adaptation Design

We choose the adaptation order based on the information flow of the system, *i.e*. the Vision-Language Model needs to first perceive well in order to speak correctly. Thus we first adapt visual encoder to videos and then instruct-tune language model for details and completeness. We conduct experiments of reversing the order in Table 9 and see a noticeable drop of WUPS (29.5→17.9).

We also compare prompt-tuning (PT) [30], an adapter-based fine-tuning method during language adaptation. Though the captioning result is close, the zero-shot QA accuracy is much worse. This indicates that adapter methods are not aimed at general-purpose multi-modal models.

	S-MiT (CIDEr)	NExT-QA (WUPS)
V → L	42.3	29.5
L → V	41.9	17.9
V → L (PT)	40.9	4.88

Table 9. **Effect of different adaptation orders.**

% of S-MiT videos	0	1%	10%	100%
S-MiT (CIDEr)	11.5	37.6	39.6	42.3

Table 10. **Effect of data size in visual adaptation.**

% of VidLN videos	0	1%	10%	100%
NExT-QA (WUPS)	1.2	8.8	16.0	29.5

Table 11. **Effect of data size in language adaptation.**

11.2. Effect of Data Scale for Adaptation

We first describe the criteria of choose training datasets in our adaptation. The goal of visual adaptation is to align video features with language, especially actions that hardly exist in images. This calls for unambiguous and clean video-text pairs. The goal of language adaptation is to learn a general-purpose video-language model from videos with detailed text annotations, *i.e*. narratives. We thus choose the best available datasets to serve these two goals: S-MiT is the largest human-annotated video captioning dataset; VidLN is the largest and most diverse video dataset with detailed human narratives.

We study the effect of the dataset size in Table 10 and Table 11. Following the motivations above, we measure the visual adaptation performance by supervised video captioning and the language adaptation performance by zero-shot QA. For both stages, our model benefits from more data. This justifies the scaling ability of our model.

11.3. Self-training with Pseudo-captioned Videos

The generated captions along with the videos can be used to further improve the VLM via self-training. We do this in the stage of visual adaptation because the language adaptation stage is mainly fueled by instruction-following data and adding pseudo-captioning leads to potential model drifting. Let $\mathcal{D}_l = \{(\mathbf{x}, \mathbf{c})\}$ and $\mathcal{D}_u = \{(\mathbf{x}, \tilde{\mathbf{c}})\}$ denote the set of human-captioned videos and VLM-captioned videos respectively. In each step, we construct a training batch by sampling a few samples from both sets, namely $\mathcal{B} = \mathcal{B}_u \cup \mathcal{B}_l$, where $\mathcal{B}_l \subset \mathcal{D}_l$ and $\mathcal{B}_u \subset \mathcal{D}_u$. Compared to self-training with "pseudo-labels", *i.e*. either manually assigned one-hot targets after filtering [16, 18] or output logits [3, 54], pseudo-captioning provides richer supervision and naturally handles the long-tail issue.

11.4. Comparing with human-labeled data

In this section, we compare the performance of the dual-encoder model trained on the human-labeled data and pseudo-captions. We train a dual-encoder model on the human-labeled S-MiT because (1) it is the largest human-labeled video-caption dataset to date and (2) our video-language model that is used to generate pseudo-captions for unlabeled videos is trained on S-MiT first. The zero-shot retrieval and recognition performance is reported in Table 8 in comparison with the result on VideoCC$^+$ and InternVid$^+$. We can see that the dual-encoder model trained on both VideoCC$^+$ and InternVid$^+$ clearly outperforms the one trained on S-MiT in terms of MSR-VTT zero-shot text-to-video retrieval recall. This indicates that our adapted video-language model not only distills human-labeled video dataset, but also generalizes to unseen videos that are within the same domain. When looking at the retrieval result on VATEX and classification result on Kinetics-600, the dual-encoder model trained on either VideoCC$^+$ or InternVid$^+$, however, is slightly inferior to that on S-MiT. We ascribe this to the semantic correlation between S-MiT and Kinetics/VATEX: S-MiT is built on top of Moments-in-Times (MiT) whose videos are all tagged with one action or activity label similar to the way Kinetics and VATEX is constructed and the action concepts between MiT and Kinetics are closely related.

Temporal reasoning

You are an AI visual assistant that can analyze a single video. You receive a few sentences, each describing the same video you are observing. The task is to use the provided caption, create a plausible question about the video, and provide the answer in detail.

Create questions that requires reasoning about temporal relationships between actions, determined by order of occurrence. The questions can also cover interactions between different persons or objects.

To answer such questions, one should require first understanding the visual content, then based on the background knowledge or reasoning, either explain why the things are happening that way, or provide guides and help to user's request. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first.

Always answer as if you are directly looking at the video.

(1) Captions:\n{}\n
Generated QA:\n{}\n
(2) Captions:\n{}\n
Generated QA:\n{}\n
(3) Captions:\n{}\n
Generated QA:\n

Causal reasoning

You are an AI visual assistant that can analyze a single video. You receive a few sentences, each describing the same video you are observing. The task is to use the provided caption, create a plausible question about the video, and provide the answer in detail.

Create questions that explain actions, either uncovering the intentions of the previously occurring actions or stating causes for subsequent actions.

To answer such questions, one should require first understanding the visual content, then based on the background knowledge or reasoning, either explain why the things are happening that way, or provide guides and help to user's request. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first.

Always answer as if you are directly looking at the video.

(1) Captions:\n{}\n
Generated QA:\n{}\n
(2) Captions:\n{}\n
Generated QA:\n{}\n
(3) Captions:\n{}\n
Generated QA:\n

Short QAs

You are an AI visual assistant that can analyze a single video. You receive a few sentences, each describing the same video you are observing. The task is to use the provided caption, create a plausible question about the video, and **provide a short answer with less than three words**.

To answer such questions, one should require first understanding the visual content, then based on the background knowledge or reasoning, either explain why the things are happening that way, or provide guides and help to user's request. Make the question challenging by not including the visual content details in the question so that the user needs to reason about that first.

Always answer as if you are directly looking at the video.

(1) Captions:\n{}\n
Generated QA:\n{}\n
(2) Captions:\n{}\n
Generated QA:\n{}\n
(3) Captions:\n{}\n
Generated QA:\n

Table 12. **The prompt template to create instruction-following data for temporal reasoning, causal reasoning, and short QAs.**

Captions: A baby girl on the left side wearing a grey t-shirt is carrying an egg then she throws the egg at the head of the man, then the egg falls on the ground and it breaks on a grey surface.
A man wearing a red t-shirt sitting on his knees is talking with the baby girl on a grey surface. In the background, there is a grey car, a grey surface, a brown mat, and people speaking and crying sounds are audible.
Question: What did the baby girl on the left side wearing a grey t-shirt do with the egg after she is carrying it?
Answer: The girl throws the egg at the head of the man.

Captions: A boy wearing black shorts is standing on the side of the swimming pool over small rocks and then he performs a backflip and injured himself. In the background, there is a swimming pool, rocks, trees, and people's voices and water splashing sound is audible.
Question: What was the boy wearing black shorts doing before he performing a backflip?
Answer: The boy is standing on the side of the swimming pool over small rocks. Then he performs a backflip and injured himself.

Captions: A boy wearing a black t-shirt rides a black bicycle in a backward direction and falls on a gray surface. A girl wearing a black cloth is moving on a gray surface, stops, and looks back toward the boy. In the background, there are gray surfaces, buildings, metallic barrier poles, trees, boats, and the sky is visible, and sounds of people speaking and the wind are audible.
Question: What did the girl do after the boy wearing a black t-shirt rides a black bicycle in a backward direction and falls on a gray surface?
Answer: The girl wearing a black cloth is moving on a gray surface. After the boy falls on a gray surface, she stops and looks back toward the boy.

Table 13. **Few-shot examples for prompting temporal-reasoning question-answer pairs.**

Captions: A baby girl on the left side wearing a grey t-shirt is carrying an egg then she throws the egg at the head of the man, then the egg falls on the ground and it breaks on a grey surface. A man wearing a red t-shirt sitting on his knees is talking with the baby girl on a grey surface. In the background, there is a grey car, a grey surface, a brown mat, and people speaking and crying sounds are audible.
Question: Why did the egg fall on the ground and breaks?
Answer: The egg was thrown by the girl at the head of the man sitting on his knees. However, it did not hit the man. Therefore, the egg falls on the ground and breaks on a grey surface.

Captions: A boy wearing black shorts is standing on the side of the swimming pool over small rocks and then he performs a backflip and injured himself. In the background, there is a swimming pool, rocks, trees, and people's voices and water splashing sound is audible.
Question: How did the boy standing on the side of the swimming pool over small rocks injure himself?
Answer: The boy standing on the side of the swimming pool over small rocks perform a backflip. Instead of falling into the water, he hit on the small rocks. That is why he injured himself.

Captions: A boy wearing a black t-shirt rides a black bicycle in a backward direction and falls on a gray surface. A girl wearing a black cloth is moving on a gray surface, stops, and looks back toward the boy. In the background, there are gray surfaces, buildings, metallic barrier poles, trees, boats, and the sky is visible, and sounds of people speaking and the wind are audible.
Question: How did the boy wearing a black t-shirt fall on a gray surface?
Answer: The boy wearing a black t-shirt rides a black bicycle in a backward direction. Riding a black bicycle backward is slow and hard to balance. Also, the boy is not good at riding a black bicycle backward. As a result, he fell on a gray surface.

Table 14. **Few-shot examples for prompting causal-reasoning question-answer pairs.**

Captions: A baby girl on the left side wearing a grey t-shirt is carrying an egg then she throws the egg at the head of the man, then the egg falls on the ground and it breaks on a grey surface. A man wearing a red t-shirt sitting on his knees is talking with the baby girl on a grey surface. In the background, there is a grey car, a grey surface, a brown mat, and people speaking and crying sounds are audible.
Question: who throws the egg at the man
Answer: baby girl

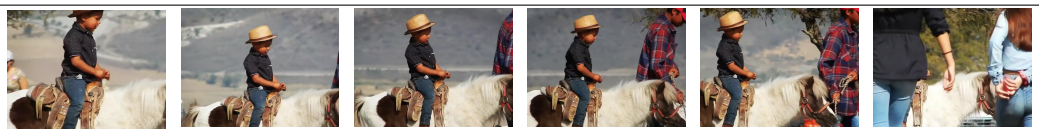
Captions: A boy wearing black shorts is standing on the side of the swimming pool over small rocks and then he performs a backflip and injured himself. In the background, there is a swimming pool, rocks, trees, and people's voices and water splashing sound is audible.
Question: what kind of pool is in the background
Answer: swimming

Captions: A boy wearing a black t-shirt rides a black bicycle in a backward direction and falls on a gray surface. A girl wearing a black cloth is moving on a gray surface, stops, and looks back toward the boy. In the background, there are gray surfaces, buildings, metallic barrier poles, trees, boats, and the sky is visible, and sounds of people speaking and the wind are audible.
Question: what happens when the man loses control
Answer: falls down

Table 15. **Few-shot examples for prompting short question-answer pairs.**

Raw alt-text:	person: view from the balcony
Image Captioner:	a bedroom with a bed and a chair / a bedroom with a bed and a chair / a balcony with a view of a baseball field / a balcony with a view of a baseball field / a balcony with a view of a baseball field / a bathroom with a bathtub and a glass shower door
Image Captioner + LLM summarization:	A bathroom with a bathtub and a glass shower door, a bedroom with a bed and a chair, and a balcony with a view of a baseball field.
Our method + Visual adaptation:	video shows the bedroom in a luxury home with a large bed and a large bathtub
Our method + V&L adaptation:	this is a video of a bedroom you can see a bed with a white mattress and there's a large glass door with a white railing going down to the patio which is covered in a white carpet with trees behind it

Raw alt-text:	football player heads a goal to make it ## as soccer player fails to jump high enough
Image Captioner:	soccer player kicking the ball during a game / soccer players fighting for the ball / soccer player scores a goal during a soccer game / soccer players are playing a game on a field / soccer players fighting for the ball / soccer player kicking the ball during a game
Image Captioner + LLM summarization:	Soccer players are playing a game of soccer on a field. A player scores a goal and celebrates.
Our method + Visual adaptation:	soccer players are playing on a soccer field
Our method + V&L adaptation:	soccer players playing soccer on the field. another player kicks the ball into the goal



Raw alt-text:	image may contain : people , smiling , on stage , playing a musical instrument , concert and outdoor
Image Captioner:	little boy riding a pony / little boy riding a pony / little boy riding a pony / two young boys riding a pony in a field / two young boys riding a pony in a field / two women walking a pony
Image Captioner + LLM summarization:	Two young boys are riding a pony in a field.
Our method + Visual adaptation:	there's a young boy riding on a miniature horse
Our method + V&L adaptation:	there's a young boy riding on a white horse a man wearing a plaid shirt is holding a harness on the horse and he's walking behind the horse while a woman in a blue and red shirt is walking behind them

Figure 6. **More examples of video captions by PaLI-3 before and after video-specific adaptation.** We show the keyframes on top for illustration purposes and the generated captions in the following blocks. Different details in text are highlighted. Best viewed in color.