Language Models with Image Descriptors are Strong Few-Shot Video-Language Learners

Anonymous Author(s)
Affiliation
Address
email

Abstract

The goal of this work is to build flexible video-language models that can generalize to various video-to-text tasks from few examples, such as domain-specific captioning, question answering, and future event prediction. Existing few-shot video-language learners focus exclusively on the encoder, resulting in the absence of a video-to-text decoder to handle generative tasks. Video captioners have been pretrained on large-scale video-language datasets, but they rely heavily on finetuning and lack the ability to generate text for unseen tasks in a few-shot setting. We propose **VidIL**, a few-shot **Vid**eo-language Learner via **I**mage and **L**anguage models, which demonstrates strong performance on few-shot video-to-text tasks without the necessity of pretraining or finetuning on any video datasets. We use the image-language models to translate the video content into frame captions, object, attribute, and event phrases, and compose them into a temporal structure template. We then instruct a language model, with a prompt containing a few in-context examples, to generate a target output from the composed content. The flexibility of prompting allows the model to capture any form of text input, such as automatic speech recognition (ASR) transcripts. Our experiments demonstrate the power of language models in understanding videos on a wide variety of video-language tasks, including video captioning, video question answering, video caption retrieval, and video future event prediction. Especially, on video future event prediction, our few-shot model significantly outperforms state-of-the-art supervised models trained on large-scale video datasets. Code and processed data will be publicly available for research purposes.

1 Introduction

One major gap between artificial intelligence and human intelligence lies in their abilities to generalize and perform well on new tasks with limited annotations. Recent advances in large-scale pre-trained generative language models [44, 6, 70, 24] have shown promising few-shot capabilities [71, 42, 62] in understanding natural language. However, few-shot video-language understanding is still in its infancy. A particular limitation of most recent video-language pretraining frameworks [28, 21, 60, 67, 66, 25, 63] is that they are encoder-only, which means they do not have the ability to generate text from videos for purposes such as captioning [61, 56], question answering [59], and future prediction [23]. Meanwhile, unified video-language models [34, 48] that are capable of language decoding still rely heavily on finetuning using a large number of manually annotated video-text pairs, therefore cannot adapt quickly to unseen tasks. Few-shot video-to-text decoding is challenging because the natural language supervision for learning video-language representation is typically based on subtitles and automatic speech recognition (ASR) transcripts [37, 67], which differ significantly from downstream tasks in terms of distribution and may have poor semantic alignment across vision and text modalities.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

We propose to address this problem by harnessing the few-shot power of large-scale language models, such as GPT-3 [6] and InstructGPT [39]. Our inspiration is derived from the fact that humans are excellent visual storytellers [15], with the ability to piece together a coherent story from a few isolated images. To mimic this, we propose **VidIL**, a few-shot **Vid**eo-language Learner via **I**mage and **L**anguage models, to use image models to provide information about the visual content in the video (as well as optionally use ASR to represent speech), and then we instruct language models to interpret a video-based summary, answer, or other target output for diverse video-language tasks.

Videos contain rich semantics and temporal content at multiple granularities. Similar to static images, videos depict objects, attributes, and events. However, the sequence of frames further conveys object state changes, actions, and events. For example, in Figure 1, the frame captions describe static visual features such as *"a person holding a green object in hand"* for the first frame. In contrast, the video clip could be correctly captioned as *"a woman makes realistic looking leaves and flowers for a cake"*, or represented as a collection of the objects and events that occur at different timestamps in the video clip, such as *cutting mat* and *flowered design*. Hence, to inform video-level description and queries, we need to represent all of this information and its temporal ordering.

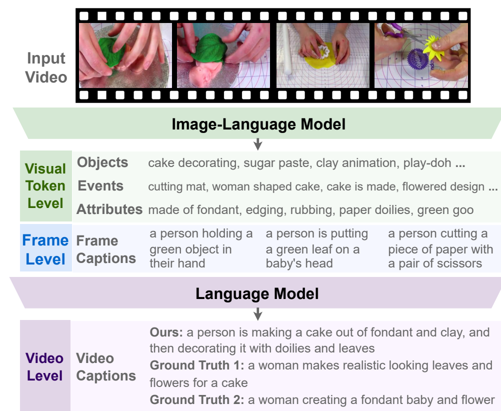

Figure 1: Multiple levels of information in videos.

To address these requirements, we propose to decompose a video into three levels: the video output, frame captions, and visual tokens (including object, event, attribute). One major benefit from this hierarchical video representation is that we can separate the visual and temporal dimensions of a video. We leverage well-trained image-language foundational models at lower levels to collect salient visual features from the sparsely sampled frames. Specifically, we leverage pretrained image-language contrastive model CLIP [43] to perform visual tokenization, based on the similarity score between frames and tokens of objects, events and attributes. The tokenization is done under the guidance of semantics role labeling [14], which provides us with candidate events with involved objects and related attributes. Next, in order to capture the overall semantics at the frame level, we employ the pretrained image captioner in image-language model BLIP [26] to obtain frame captions. We then instruct pretrained language models using in-context learning [39, 13, 50, 47] to interpret visual tokens and frame captions into the target text content. In detail, we temporally order visual tokens and frame captions using specially designed prompts such as *"First...Then...Finally"*, to instruct the pretrained language model to track the changes of objects, events, attributes and frame semantics along the temporal dimension.

Without pretraining or finetuning on any video datasets, we show that our approach outperforms both video-language and image-language state-of-the-art baselines on few-shot video captioning and question answering tasks. Moreover, on video-language event prediction, our approach significantly outperform fully-supervised models while using only 10 labeled examples. We further demonstrate that our generative model can benefit broader video-language understanding tasks, such as text-video retrieval, via pseudo label generation. Additionally, we show that our model is highly flexible in adding new modalities, such as ASR transcripts.

2 Related Work

2.1 Image-Language Models and Their Applications on Video-Language Tasks

Large-scale image-language pretraining models optimize image-text matching through contrastive learning such as CLIP [43] or by learning image-text alignments [64, 27, 57, 65, 33, 51, 8, 29, 72, 69, 17, 18, 16]. Recently, BLIP [26] proposes an image captioner pretrained with filtering in order to achieve optimal image-text alignment and has shown promising performance on various image-language tasks. However, video-language pretraining [25, 34, 28, 36, 3, 1, 41] is still hindered by noisy and domain-specific video datasets [22, 37]. Naturally, researchers start to explore transferring the rich knowledge from image models to videos. Different from the traditional way of representing videos by 3D dense features [12], recent work [21, 25] proves that sparse sampling is an effective way to represent videos, which facilitates applying pre-trained image-language models to video-language tasks [35, 11]. Specifically, the image-language model BLIP [26] sets new state-of-the-art on zero-shot retrieval-style video-language tasks, such as video retrieval and video question answering. However, for generation-style tasks such as domain-specific video captioning, video-language model UniVL [34] still leads the performance but highly rely on fine-tuning. In this work, we extend the idea of leveraging image-language models to a wide variety of video-to-text generation tasks. We further connect image-language models (ILM) with language models (LM) which empowers our model with strong generalization ability. We show that the knowledge from both image-language pretraining and language-only pretraining can benefit video-language understanding in various aspects.

2.2 Unifying Multi-Modal Tasks with Language Models

Connecting different modalities with a unified representation has been paid much attention to recently. Text-only generation models, such as T5 [45], has been extended to vision-language tasks by text generation conditioned on visual features [9, 52, 49, 74, 54]. In order to fully leverage the generalization power from pretained language models, [62] represents images using text in a fully symbolic way. [32] includes more modalities such as video and audio, but requires annotated video-text data to jointly training the language model with the video and audio tokenizer. In this work, we propose a temporal-aware hierarchical representation for describing a video textually. To our knowledge, we are the first work to leverage prompting a frozen language model for tackling few-shot video-language tasks with a unified textual representation. Concurrent work Socratic [68] uses a zero-shot language-based world-state history to represent long videos with given time stamps, while our model can quickly adapt to different video and text distributions with few examples. Furthermore, we show that by injecting *temporal markers* to the prompt we can make a pre-trained language model understand fine-grained temporal dynamics in video events. Compared with the concurrent work Flamingo [2], which requires dedicated vision-language post-pretraining, our framework does not require to pretrain on any videos. Our framework is simple and highly modulated where all the components are publicly available. Additionally, our framework is more flexible on adding new modalities, e.g., automatic speech recognition, without the need for complex redesigning.

3 Method

We propose a hierarchical video representation framework which decomposes a video into three levels, i.e., **visual token level**, **frame level** and **video level**. The motivation is to separate the spatial and temporal dimension of a video in order to leverage image-language and language-only foundation models, such as CLIP [43] and GPT-3 [6]. All three levels use a unified textual representation which enables us to leverage the powerful few-shot ability from pretrained language models.

3.1 Frame Level: Image Captioning

Following [21] we first perform sparse sampling to obtain several video frames. Unless otherwise specified, we sample 4 frames for frame level and 8 frames for visual token level. We then feed each frame into a pre-trained image-language model to obtain frame level captions. An example can be found in the blue part of Figure 2. In our experiments, we use BLIP [26], a recent image-language framework containing both image-grounded encoder and decoder, for generating frame captions. We follow [26] to do both captioning and filtering on each frame. However, as mentioned in Section 1, videos contain rich semantics and temporal contents at multiple granularities. It is not enough to generate video-level target text such as video captions solely based on frame captions. Thus, we further perform visual tokenization for each frame to capture features at a finer granularity.

3.2 Visual Token Level: Structure-Aware Visual Tokenization

At this level, we aim to extract the textual representations of salient visual token types, such as objects, events and attributes. We found that pre-defined classes for classification, such as those in ImageNet [10], are far from enough for covering the rich semantics in open-domain videos. Thus,

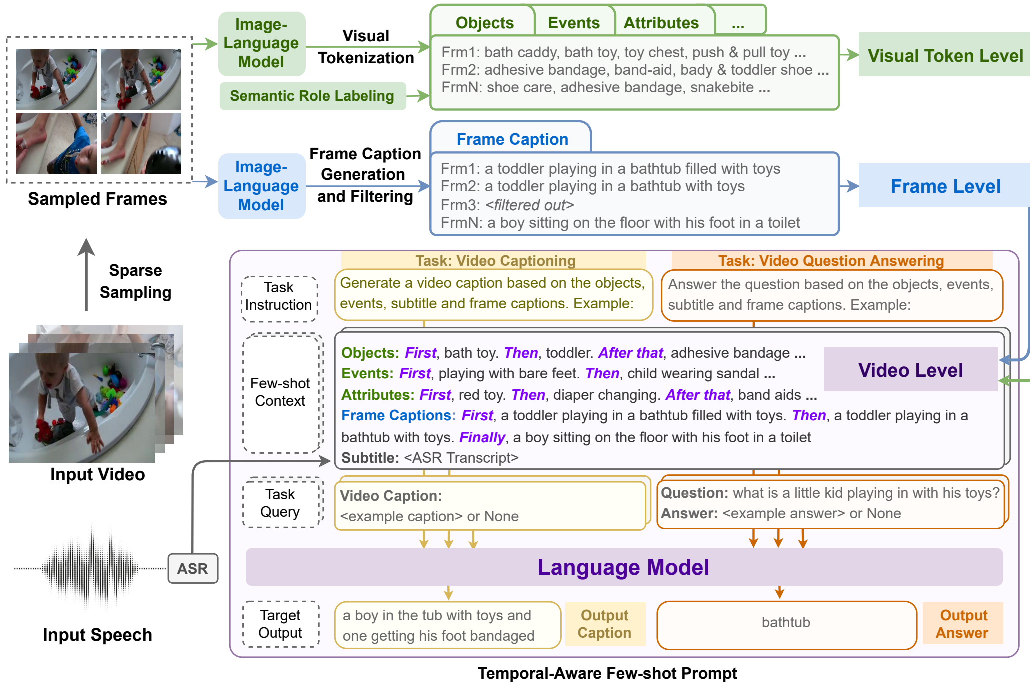

Figure 2: Overview of VidIL framework. We represent a video in a unified textural representation containing three semantic levels: **visual token level**, **frame level**, and **video level**. At visual token level, we extract salient objects, events, attributes for each sampled frame. At frame level, we perform image captioning and filtering. And at video level, we construct the video representation by aggregating the visual tokens, frame captions and other text modalities such as **ASR**, using a few-shot temporal-aware prompt. We then feed the prompt to a pre-trained language model together with task-specific instructions to generate target text for a variety of video-language tasks.

instead of using classification-based methods for visual tokenization as in previous work [32, 62], we adopt a retrieval-based visual tokenization approach by leveraging pre-trained contrastive image-language models. Given a visual token vocabulary which contains all candidate object, event, and attribute text phrases, we compute the image embedding of a frame and the text embeddings of the candidate visual tokens using a contrastive multi-modal encoder, CLIP [43]. We then select top 5 visual tokens per frame based on the cosine similarity of the image and text embeddings. An example of the extracted object tokens can be found in the green part of Figure 2.

Unlike in images where objects and attributes already cover most visual features, events are more informative in videos. In order to discover events from video frames, we construct our own event vocabulary by extracting event structures from Visual Genome [19] synsets[1] using **Semantic Role Labeling**. Specifically, we first select the phrases that contains at least one verb and one argument as **events**. Then we remove highly similar events based on their sentence similarity using Sentence-BERT [46] embeddings. For object vocabulary, we adopt OpenImage [20] full classes (∼20k), instead of using the visually groundable subset (∼600) as in concurrent work [68]. We found that using *large but noisy* vocabulary is more effective than using *small but clean* vocabulary in our retrieval-based setting with CLIP. For attribute vocabulary, we adopt visual genome attribute synset. In Section 4.6, we provide ablation study on the impact of different types of visual tokens. The statistics of visual token vocabulary can be found in Appendix.

3.3 Video Level: Temporal-Aware Few-shot Prompting

Once we obtain the textual representation from frame level and visual token level, the final step is to put the pieces together to generate a video level target text. The goal is to build a model that can be quickly adapted to any video-to-text generation task with only a few examples. To this

[1]We use the keys in Visual Genome [19] object synsets which contains frequent <verb,object> pairs.

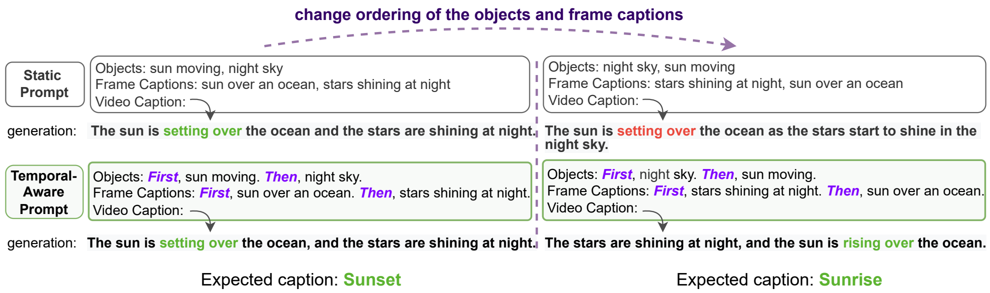

Figure 3: Temporal-aware prompt successfully distinguishes the **Sunset** and **Sunrise** scenario from the temporal ordering change of objects and frame captions, while the static prompt fails.

end, we propose to leverage large-scale pre-trained language models, such as GPT-3 [6], with a temporal-aware few-shot prompt. As shown in Figure 2, our framework can be readily applied to various video-to-text generation tasks, such as video captioning and video question answering, with a shared prompt template. The proposed prompting strategy enables a language model to attend to the lower level visual information as well as taking into account the temporal ordering.

Here, we use the video captioning task depicted in Figure 2 to illustrate the details. The few-shot prompt consists of three parts: **instruction**, **few-shot context**, and **task query**. The **instruction** is a concise description of the generation task, e.g., "Generate a video caption based on the objects, events, attributes and frame captions. Example:", which is proved to be effective in zero-shot and few-shot settings [6, 58]. The **few-shot context** contains the selected in-context examples as well as the test video instance. Each video instance is represented by the aggregated visual tokens[2], e.g., "Objects: First, bath toy. Then,...", the frame captions, e.g., "Frame Captions: First, a toddler playing in a bathtub filled with toys. Then,...", and the ASR inputs if available, e.g., "Subtitle:<ASR Transcript>". Finally, the **task query** is a task-specific suffix indicating the target text format, e.g. "Video Caption:". For in-context examples (omitted here for simplicity), the task query is followed by ground truth annotation, while for the test instance, the generation starts at the end of the task query.

Formally, we denote the instruction line as $\mathbf{t}$, few-shot context as $\mathbf{c}$, the task query as $\mathbf{q}$, and the target text as $\mathbf{y}$, where $\mathbf{y} = (y_1, y_2, ..., y_L)$. The generation of the next target token y_l can be modeled as:

$$y_l = \arg\max_y p(y|\mathbf{s}, \mathbf{c}, \mathbf{q}, y_{<l}) \tag{1}$$

In order to capture the temporal dynamics between frames and visual tokens, we further propose to inject **temporal markers** to the prompt. As shown in the few-shot context in Figure 2, each visual token and frame caption is prefixed with a natural language phrase indicating its temporal ordering, e.g., "First,","Then,", and "Finally,". We found adding the temporal marker can make the language model condition on not only the literal but also the temporal information of the context. We show an example in Figure 3, where we compare our *temporal-aware prompt* with a *static prompt* on video captioning using InstructGPT. Again, the in-context examples are omitted here, which can be found in Appendix. In this example, the only difference between the two context is the ordering of the visual tokens and the frame captions. For the context on the left, where "sun moving" appears before "night sky", we are expected to see a story talking about **sunset**, while for the context on the right, we are expected to see **sunrise**. We can see the static prompt generates captions about sunset for both contexts, while the temporal-aware prompt can capture the temporal ordering correctly and generate sunrise for the context on the right.

[2]To obtain video level visual tokens, the visual tokens extracted from each frame are further ranked and ordered based on frequency and frame index. More details can be found in Appendix C.

4 Experiments

4.1 Experimental Setup

To comprehensively evaluate our model, we show results on four video-language understanding tasks in few-shot settings: video captioning, video question answering (QA), video-language event prediction, and text-video retrieval. We compare our approach with state-of-the-art approaches on five benchmarks, i.e, MSR-VTT [61], MSVD [7], VaTeX [56], YouCook2 [73], and VLEP [23].

Implementation Details. We use CLIP-L/14[4] as our default encoder for visual tokenization. We adopt BLIP captioning checkpoint[5] fine-tuned on COCO [31] for frame captioning. We use InstructGPT [39] as our default language model for generating text conditioned on the few-shot prompt. To construct event vocabulary, we use the semantic role labeling model from AllenNLP[6]. The experiments are conducted on 2 NVIDIA V100 (16GB) GPUs.

Table 1: Statistics of datasets in our experiments

Dataset	Task	Split Count # train / # eval
MSR-VTT [61]	Captioning; QA	6,513 / 2,990
MSR-VTT [61]	Retrieval	7,010 / 1,000
MSVD [7]	Question Answering	30,933 / 13,157
VaTeX v1.1[3] [56]	Captioning; Retrieval	25,991 / 6,000
YouCook2 [73]	Captioning	10,337 / 3,492
VLEP [23]	Event Prediction	20,142 / 4,192

In-context Example Selection. From our preliminary experiments, we found that the generation performance is sensitive to the quality of in-context examples. For example, for QA tasks such as MSVD-QA where the annotations are automatically generated, the <question, answer> pair in randomly selected in-context examples can be only weakly-correlated with the video context. Thus, instead of using a fixed prompt for each query, we dynamically filter out the irrelevant in-context examples. Specifically, given a randomly sampled M-shot support set from the training set, we select a subset of N-shots as in-context examples based on their SentenceBERT [46] similarities with text queries. Furthermore, we reorder the selected examples in ascending order based on the similarity score to account for the recency bias [71] in large language models. For QA tasks, we choose the most relevant in-context examples by comparing with questions. While for captioning task, we compare with frame captions. If not otherwise specified, we use M=10 and N=5, which we consider as 10-shot training.

4.2 Few-shot Video Captioning

We report BLEU-4 [40], ROUGE-L [30], METEOR [5], and CIDEr [53] scores on three video captioning benchmarks covering both open-domain (MSR-VTT, VaTeX) and domain-specific (YouCook2) videos. We compare with both state-of-the-art video captioner (UniVL [34]) and image captioner (BLIP [26]). In order to implement the BLIP baseline for few-shot video captioning, we extend the approach used for text-video retrieval evaluation in [26] to video-language training. Specifically, we concatenate the visual features of sampled frames and then feed them into the image-grounded text-encoder to compute the language modeling loss. This is equivalent to stitching the sampled frames into a large image and then feeding it to BLIP for image captioning. We found that this simple approach results in very strong baselines.
As shown in Table 2, existing methods have strong bias on certain datasets. For example, UniVL performs well on YouCook2 but fails on MSR-VTT and VaTeX, while BLIP performs the opposite. This is because UniVL is pretrained on HowTo100M which favors instructional videos, i.e., YouCook2, while BLIP is pre-training on image-caption pairs which favors description-style captions, i.e., MSR-VTT and VaTeX. On the contrary, our model performs competitively on both open-domain and instructional videos, and significantly outperforms the baselines in the average CIDEr score across all three benchmarks. This indicates that by leveraging language models, we can maintain strong few-shot ability regardless of the video domain and or the target caption distribution.

As discussed in Section 1, video captions describes the content in various semantic levels. The N-gram based metric may not fairly reflect the models' performance in capturing the video-caption

[4] https://huggingface.co/openai/clip-vit-large-patch14
[5] https://github.com/salesforce/BLIP#finetuned-checkpoints
[6] https://docs.allennlp.org/models/main/models/structured_prediction/predictors/srl/

Table 2: 10-shot video captioning results. ♠ indicates concurrent work. The reported *Flamingo* [2] results are using 16 shots. *#Video$_{PT}$* represents the number of videos used for pre-training. *B-4*, *R-L*, *M*, *C* represents *BLEU-4*, *ROUGE-L*, *METEOR* and *CIDEr*. *Avg C* represents the average CIDEr score across all available benchmarks. *ASR* indicates whether the model has access to the ASR subtitles. *BLIP* and *BLIP$_{cap}$* use the pretrained checkpoint and the finetuned checkpoint on COCO captioning. All results are averaged over three random seeds.

Method	#Video$_{PT}$	ASR	MSR-VTT Caption				YouCook2 Caption				VaTex Caption				Avg C
			B-4	R-L	M	C	B-4	R-L	M	C	B-4	R-L	M	C	
Few-shot															
UniVL	1.2M	No	2.1	22.5	9.5	3.6	**3.3**	**25.3**	**11.6**	**34.1**	1.7	15.7	8.0	2.1	13.3
BLIP	0	No	**27.7**	43.0	23.0	**39.5**	0.7	9.0	3.4	11.5	13.5	39.5	15.4	20.7	23.9
BLIP$_{cap}$	0	No	21.6	48.0	22.7	30.2	3.7	8.6	3.8	9.4	20.7	41.5	17.4	28.9	22.8
VidIL(ours)	0	No	26.0	**51.7**	**24.7**	36.3	2.6	22.9	9.5	27.0	**22.2**	**43.6**	**20.0**	**36.7**	**33.3**
UniVL	1.2M	Yes	-	-	-	-	4.3	26.4	12.2	48.6	2.7	17.7	10.2	3.4	26.0
VidIL(ours)	0	Yes	-	-	-	-	**10.7**	**35.9**	**19.4**	**111.6**	**23.2**	**44.2**	**20.6**	**38.9**	**75.3**
♠Flamingo-3B(16)	27M	No	-	-	-	-	-	-	-	73.2	-	-	-	57.1	-
♠Flamingo-80B(16)	27M	No	-	-	-	-	-	-	-	84.2	-	-	-	62.8	-
Fine-tuning															
UniVL	1.2M	No	42.0	61.0	29.0	50.1	11.2	40.1	17.6	127.0	22.8	38.6	22.3	33.4	70.2
UniVL	1.2M	Yes	-	-	-	-	16.6	45.7	21.6	176.8	23.7	39.3	22.7	35.6	106.2

alignment. We further verifies this hypothesis in Section 4.5. Thus, in addition to the automatic metrics, we include qualitative examples illustrated in Figure 4. More examples are in Appendix.

Additionally, for most existing methods and also concurrent work, e.g., Flamingo [2], adding a new modality often requires a dedicated model redesign and or retraining. However, the nature of our framework, where we use a unified textual representation for each level, making it highly flexible for incorporating new modalities. As shown in row 6 in Table, our model can effectively utilize extra information from ASR to obtain significantly better few-shot performance on certain datasets such as YouCook2.

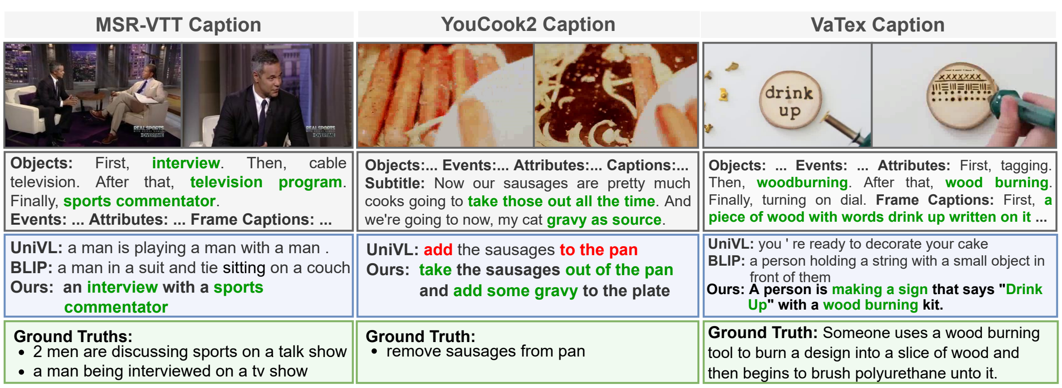

Figure 4: Qualitative examples on video captioning. Grey boxes contains part of the video representation from our model. Blue boxes contains caption generation from different models. Green boxes contains ground truth annotations. **Bold green text** highlights the correct information that is not captured in baseline outputs which can be reasoned from our visual tokens and frame captions.

4.3 Few-shot Video Question Answering

We compare the test accuracy of our approach with few-shot pretrained BLIP, BLIP$_{VQA}$ [26], and concurrent work Flamingo [2] on two video question answering benchmarks, MSR-VTT_QA and MSVD_QA. BLIP$_{VQA}$ represents finetuned BLIP on VQA [4] dataset, which is the previous SOTA on zero/few-shot video question answering. In order to have fairer comparison with BLIP$_{VQA}$, we reduce the shot number to 5 and report the average accuracy on three sets of randomly selected

5-shot examples. As shown in Table 3, our method outperforms previous SOTA by a large margin. Comparing with concurrent work Flamingo, which is post-pretrained on a large number of video-text data, our model is training-free and did not observe any video data. However, with only image-language and language-only knowledge, our 5-shot model is able to beat 8-shot Flamingo-3B and achieve on-par performance with 4-shot Flamingo-80B.

Table 3: Video QA results. BLIP_{VQA} is finetuned on VQA [4]. ♠ indicates concurrent work. PT, FT indicates pretraining and finetuning.

Method	$\#\text{video}_{\text{PT}}$	$\#\text{video}_{\text{FT}}$	MSR-VTT	MSVD
BLIP	0	0-shot	0.55	0.45
BLIP	0	5-shot	0.84	0.53
BLIP_{VQA} [26]	0	0-shot	19.2	35.2
VidIL(ours)	0	5-shot	**21.2**	**39.1**
♠Flamingo-3B [2]	27M	4-shot	14.9	33.0
♠Flamingo-3B [2]	27M	8-shot	19.6	37.0
♠Flamingo-80B [2]	27M	4-shot	23.9	41.7
♠Flamingo-80B [2]	27M	8-shot	27.6	45.5
ALPRO [25]	2M	full-shot	42.1	45.9

Table 4: Accuracy (%) on VLEP hidden test set.

Method	$\#\text{video}_{\text{FT}}$	Acc
VLEP [23]	20142	67.5
MERLOT [67]	20142	68.4
VidIL(ours)	10-shot	**72.0**
Human	-	90.5

4.4 Few-shot Video-Language Event Prediction

In this section, we show that our model not only can answer questions about the video visual features but also answering "What is more likely to happen next?". Given a video with associated subtitle transcript as premise, the video-language event prediction (VLEP) task is to predict the most likely future event. The original VLEP [23] paper formulates the problem as a binary classification problem where the model will be chosen from two possible future event candidates. Instead, we formulate this problem as another video-to-text generation problem to fit into our framework. Figure 5 depicts an example with the same format as in Figure 2. Similar to the evaluation setting in QA, the generated free-form text will first be mapped to one of the two candidate answers using SentenceBert [46], and then calculate the accuracy. In Table 4, we report accuracy on the hidden test set of VLEP [23]. To our surprise, our 10-shot model beats state-of-the-art fully-supervised baseline, i.e., MERLOT [67], by a large margin ($\sim 4\%$). This is showing that our model has strong few-shot ability not only in video-language understanding but also in prediction. Since event prediction tasks rely heavily on temporal ordering, we show that with the proposed temporal-aware prompting, language models can be guided to capture temporal dynamics between historical and future events.

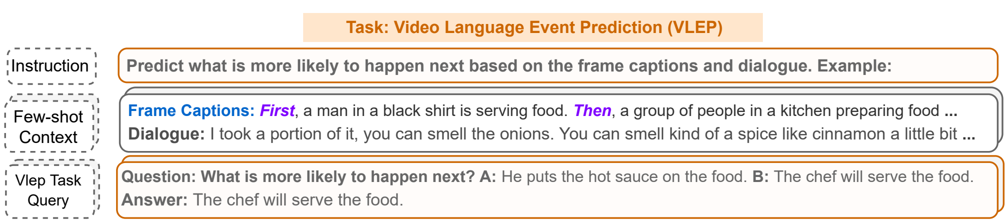

Figure 5: Prompt for VLEP task.

4.5 Semi-supervised Text-Video Retrieval

In addition to video-to-text generation tasks, we show that a broader range of video-language tasks can benefit from our few-shot video captioner from a data perspective. Here, we consider a low-budget semi-supervised setting where we only have a few labeled video-caption pairs and a large amount of unlabeled videos. The idea is to leverage our video captioner to generate **pseudo labels** for training any given vision-language models. As a case study, we evaluate on two text-video retrieval benchmarks, i.e., MSR-VTT and VaTeX. We use greedy decoding to generate pseudo caption for each video in the training set. We then train an identical base model, i.e., BLIP, using different pseudo labeled data as well as ground truth annotations. We report Recall @ 1 and 5 for both video-to-text and text-to-video retrieval. Table 5 shows that through training on our pseudo labels, we can achieve significant improvements compared with zero-shot BLIP. We also show that the performance gain

Table 5: Semi-supervised text-video retrieval with 10 labeled examples. V_{label} or $V_{unlabel}$ are the number of labeled and unlabeled videos, respectively. *t_R1* and *t_R* denote video-to-text Recall@1 and 5. *v_R1* and *v_R5* denote text-to-video Recall@1 and 5.

Model	Pseudo Label	MSR-VTT Retrieval					VaTex Retrieval				
		$V_{label}/V_{unlabel}$	t_R1	t_R5	v_R1	v_R5	$V_{label}/V_{unlabel}$	t_R1	t_R5	v_R1	v_R5
BLIP	-	-	33.2	57.2	40.5	62.8	-	28.2	53.4	**34.0**	58.6
BLIP	UniVL	10 / 7010	33.1	57.3	33.6	57.7	10 / 22685	25.5	47.7	26.1	49.1
BLIP	BLIP	10 / 7010	35.6	60.8	39.8	60.4	10 / 22685	26.3	50.5	29.3	53.6
BLIP	BLIP_{cap}	10 / 7010	35.3	58.0	39.1	63.3	10 / 22685	23.9	46.8	27.5	49.7
BLIP	VidIL(ours)	10 / 7010	**39.6**	**64.5**	**40.8**	**65.2**	10 / 22685	**33.3**	**59.1**	33.7	**59.5**
BLIP	Ground Truth	7010 / 0	43.6	66.2	43.1	67.2	22685 / 0	40.1	66.4	40.1	66.6
ALPRO [25]	Ground Truth	140200 / 0	32.0	60.6	33.9	60.7	-	-	-	-	-
DRL [55]	Ground Truth	180000 / 0	54.1	77.4	52.9	78.5	-	-	-	-	-

is not simply a result of training on more data, since finetuning on the pseudo labels generated by other baselines (UniVL, BLIP) is less effective and can even hurt the performance. Furthermore, on MSR-VTT Recall @ 5 we can even achieve comparable performance against BLIP model finetuned on full ground truth annotations.

Another interesting observation is that, compared with the video captioning results in Table 2, we found that the gain of our model over baselines on text-video retrieval is more visible than on captioning. A key factor in performing well on text-video retrieval tasks is to learn a good video-text multi-modal alignment. This result shows that our pseudo labels capture richer video-text alignment that can benefit the retrieval-style downstream task. The N-gram based generation metrics, e.g., BLEU, may not be able to fully reflect the alignment information, due to the variety of semantic levels in video captions. Furthermore, from a data perspective, our video captioner can be viewed as a data augmentation tool which is capable of generating or augmenting any open-domain video-language pretraining datasets with minimal human effort. As a result, we can potentially improving video-language pretraining by constructing a cleaner and more diverse video-text corpus.

Table 6: Impact of visual tokens and the temporal dimension

Video Representation		Avg↑	Std↓
Visual Token	Frame	39.6	3.7
	Frame+Object	40.3	2.9
	Frame+Object+Event	39.9	2.8
	Frame+Object+Attibute	**40.9**	2.9
	Frame+Object+Event+Attribute	40.8	**2.4**
Temporal	Reduce to one frame	38.5	2.4
	Reverse temporal order	40.7	1.7

Table 7: Impact of shot selection. *#ICE* indicates the number of in-context examples in the prompt.

#shot	w/o selection			w/ selection		
	#ICE	Avg↑	Std↓	#ICE	Avg↑	Std↓
5	5	38.4	2.1	5	40.4	1.2
10	10	41.3	3.6	5	40.8	2.4
20	20	42.6	3.3	5	42.2	2.0
30	30	40.0	2.9	5	41.1	1.9

4.6 Ablation Studies

We perform comprehensive ablation studies on our few-shot prompt including the impact of different video representation, the number of shots and the in-context selection. All the ablation results are evaluated on MSVD_QA validation set, we report the mean and standard deviation of each setting on three sets of randomly sampled shots. For the cases **with** in-context example selection, we further select 5 examples as in-context examples from the sampled shots, while for the cases **without** in-context selection, all shots will be feed into the prompt. In Table 6, we show adding visual tokens consistently improves not only the model accuracy but also the model variance. A lower standard deviation indicates that the model is less sensitive to the few-shot sampling.

To further demonstrate the impact of the additional temporal dimension of videos, we perform two ablation on the "`Frame+Object+Event+Attribute`" setting. First, we reduce the number of frame captions and visual tokens to be one[7] for each video. We found that the performance drops significantly compared with using the default four frames, which indicates the model's ability to incorporate information from multiple timestamps. Further, we found that fine-grained temporal

[7] we use the frame caption and visual tokens from the middle frame

modeling is rarely required for performing well on current video-language benchmarks. As shown in the ablation result where we reverse the order of all visual tokens and frame captions, the performance decreased only marginally, which indicates that current benchmarks may not be sufficient in reflecting the benefits from better temporal ordering.

In Table 7, we first show that, with the same context length, namely, 5 in-context examples, in-context example selection significantly increases the performance as well as the robustness. At 10-shot, and 20-shot, directly fitting more shots into the prompt results in better performance. In-context selection achieves slightly lower performance but with significantly better efficiency due to shorter context. Interestingly, at 30-shot, in-context selection with 5 examples outperforms directly adding all 30 shots into the prompt. This is showing that in-context selection can help the model utilize a larger number noisy video examples. Nevertheless, we still observe that the benefit of adding more shots saturated at around 20 to 30 shots, even if with in-context selection. we view this as a remaining challenging on how to make language models benefit from longer contexts.

5 Conclusions, Limitations and Future Work

This paper proposes VidIL, a few-shot **Vid**eo-language Learner via **I**mage and **L**anguage models. It demonstrates the strong ability of large-scale language models on performing video-to-text tasks when frame features are provided as unified text representations using image-language models. We propose a temporal order aware prompt by decomposing videos into a hierarchical structure, which is able to plug in multiple levels of frame features, along with speech transcripts. Without pretraining on videos, our model outperforms vision-language models learned from large-scale video datasets on a variety of few-shot tasks, such as domain-specific captioning, question answering, and future event prediction. One limitation of using unified textual representation is that we might lose low-level visual features which can be essential for some specific tasks, such as fine-grained spatial visual question answering. We also observe that current video-language benchmarks rarely require explicit temporal tracking on the frames and visual tokens. Future work will focus on leveraging large-scale language models for learning script knowledge from long videos where temporal dynamics are better emphasized. For broader impact on the society, please refer to the supplemental material.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *ArXiv preprint*, abs/2204.14198, 2022. 3, 7, 8

[3] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society, 2015. 7, 8

[5] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. 6

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina

Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1, 2, 3, 5

[7] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, 2011. Association for Computational Linguistics. 6

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[9] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR, 2021. 3

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 3

[11] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *ArXiv preprint*, abs/2106.11097, 2021. 3

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019. 3

[13] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online, 2021. Association for Computational Linguistics. 2

[14] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia, 2018. Association for Computational Linguistics. 2

[15] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California, 2016. Association for Computational Linguistics. 2

[16] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12976–12985, 2021. 2

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 2021. 2

[18] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR, 2021. 2

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 4, 16

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 4, 16

[21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learningvia sparse sampling. In *CVPR*, 2021. 1, 3

[22] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium, 2018. Association for Computational Linguistics. 3

[23] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784, Online, 2020. Association for Computational Linguistics. 1, 6, 8

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, 2020. Association for Computational Linguistics. 1

[25] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *arxiv*, 2021. 1, 2, 3, 8, 9

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *ArXiv preprint*, abs/2201.12086, 2022. 2, 3, 6, 7, 8, 17

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[28] 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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online, 2020. Association for Computational Linguistics. 1, 2

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[30] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain, 2004. 6

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[32] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7001–7011. IEEE, 2021. 3, 4

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23, 2019. 2

[34] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *ArXiv preprint*, abs/2002.06353, 2020. 1, 2, 3, 6, 17

[35] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *ArXiv preprint*, abs/2104.08860, 2021. 3

[36] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9876–9886. IEEE, 2020. 2

[37] 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 *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE, 2019. 1, 3

[38] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 16

[39] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *ArXiv preprint*, abs/2203.02155, 2022. 2, 6

[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 6

[41] Mandela Patrick, Po-Yao Huang, Yuki Markus Asano, Florian Metze, Alexander G. Hauptmann, João F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2

[42] Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182, Online, 2020. Association for Computational Linguistics. 1

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 16

[44] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv preprint*, abs/1910.10683, 2019. 1

[45] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020. 3

[46] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. 4, 6, 8, 16

[47] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, 2021. Association for Computational Linguistics. 2

[48] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. *ArXiv preprint*, abs/2201.08264, 2022. 1

[49] Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. Language models can see: Plugging visual controls in text generation. *ArXiv preprint*, abs/2205.02655, 2022. 3

[50] Derek Tam, Rakesh R. Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. Improving and simplifying pattern exploiting training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4980–4991, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2

[51] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. 2

[52] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 3

[53] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society, 2015. 6

[54] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *ArXiv preprint*, abs/2202.03052, 2022. 3

[55] Qiang Wang, Yanhao Zhang, Yun Zheng, Pan Pan, and Xian-Sheng Hua. Disentangled representation learning for text-video retrieval. *ArXiv preprint*, abs/2203.07111, 2022. 9

[56] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE, 2019. 1, 6

[57] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv preprint*, abs/2108.10904, 2021. 2

[58] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *ArXiv preprint*, abs/2109.01652, 2021. 5

[59] 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 *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653, 2017. 1

[60] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 1

[61] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society, 2016. 1, 6

[62] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. *ArXiv preprint*, abs/2109.05014, 2021. 1, 3, 4

[63] Ziyi Yang, Yuwei Fang, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, et al. i-code: An integrative and composable multimodal learning framework. *ArXiv preprint*, abs/2205.01818, 2022. 1

[64] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216, 2021. 2

[65] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *ArXiv preprint*, abs/2111.11432, 2021. 2

[66] 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. *ArXiv preprint*, abs/2201.02639, 2022. 1

[67] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 8

[68] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. *ArXiv preprint*, abs/2204.00598, 2022. 3, 4

[69] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 2

[70] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *ArXiv preprint*, abs/2205.01068, 2022. 1

[71] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 2021. 1, 6

[72] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13041–13049. AAAI Press, 2020. 2

[73] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598. AAAI Press, 2018. 6

[74] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Xiaogang Wang, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. *ArXiv preprint*, abs/2112.01522, 2021. 3

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See supplemental material.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1 and supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Section 4.6.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] see Section 4.6.
   (b) Did you mention the license of the assets? [Yes] see the citations in Section 4.6.
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] see the citations in Section 4.6.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] see the citations in Section 4.6.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Additional Qualitative Examples

Additional qualitative examples on MSR-VTT, YouCook2 and VaTex captioning can be found in Figure 6,7. We show that our framework can capture important video semantics (shown in **bold green text**), such as objects, events and attributes, that are missing in the captions generated by baselines.

B Few-shot Prompt Examples

We show a full view of the few-shot prompts used in video captioning (Figure 8, 9), video question answering (Figure 10) and video-language event prediction (Figure 11). Additionally, in Figure 12, we show the omitted in-context examples for Figure 3 in the main body.

C Additional Implementation Details

Table 8: Statistics of visual token vocabulary.

Visual Token	Source	Original Size	Final Size
Objects	OpenImage v6 Classe Names	19,975	19,965
Events	Visual Genome Object Synset (keys)	40,154	7,414
Attributes	Visual Genome Attribute Synset (keys)	18,720	16,693

Statistics of Visual Token Vocabulary We construct our visual token vocabulary based on OpenImage [20] v6 class names[8], visual genome [19] object synsets[9] and visual genome attribute synsets[10]. The statistics can be found in Table 8. Visual genome synsets are <key, value> pairs, where the keys are noisy natural language phrases and the values are the mapped WordNet synsets [38]. For object vocabulary, we perform minimum cleaning by removing fictional character names such as `"robin (fictional character)"`, which we found are highly biased by the CLIP [43] model on video frames. For attribute vocabulary, we clean up attribute synset keys by removing phrases with a cosine similarity larger than 0.9 using SentenceBert [46] embedding, such as `"facing upward"` and `"facing upwards"`. For event vocabulary, we select phrases containing <verb,object> structures from the object synset keys by running semantic role labeling[11]. We then remove semantically similar entries with a threshold of 0.9 based on SentenceBert embeddings.

Implementation Details for Visual Token Aggregation Once we obtained top 5 visual tokens for each frame, we further aggregate them to construct the video-level visual tokens which will be part of the few-shot prompt. We first rank the visual tokens based on their single frame ranking score with

[8] https://storage.googleapis.com/openimages/v6/oidv6-class-descriptions.csv

[9] https://visualgenome.org/static/data/dataset/object_synsets.json.zip

[10] https://visualgenome.org/static/data/dataset/attribute_synsets.json.zip

[11] https://docs.allennlp.org/models/main/models/structured_prediction/predictors/srl/

the appearance frequency across all frames as tie breaker. In our implementation, we consider up to top 4 video-level visual tokens, we then filter out any visual token that has not been ranked within top 2 in any frames. To identify the ordering of the obtained video-level visual tokens, we consider the frame index from which they are extracted from as their temporal indicator. If a visual token occurs in multiple frames, we use the averaged frame index as its temporal indicator. Finally, in order to apply temporal prompt template to variable number of visual tokens, we use a dynamic template which changes according to the number of tokens. For example, if we have three visual tokens, we remove `"After that"` and only use `"First", "Then", "Finally"`. If we have more then four visual tokens, we repeat `"Then"` or `"After that"` for tokens in the middle.

Implementation Details for Few-shot Video Captioning Baselines In order to finetune the pretrained baselines (UniVL [34], BLIP [26], BLIP$_{cap}$ [26]) with few annotated examples on video captioning, we set the learning rate to be small and the warm-up steps to be high. Specifically, for UniVL, we set the number of epoches to be 50 and the linear warmup steps to be 40. We use a learning rate of $1e$-6 for captioning task without ASR input and $3e$-6 with ASR input. For BLIP and BLIP$_{cap}$, we set the number of epoches to be 5 with a learning rate of $5e$-7. For each video, we sample 4 frames (each with a size of 224) at training time and 8 frames at test time. We set all batch size to be the same as the few-shot number, i.e., 10.

Implementation Details for Semi-supervised Text-Video Retrieval We use pretrained BLIP with Vit-B/16[12] as our base model for training on different pseudo labeled datasets as well as ground truth annotations for text-video retrieval. We train the model for one epoch using a batch size of 16 and a learning rate of $5e$-6. For each video, we sample 4 frames (each with a size of 224) at training time and 8 frames at test time. We follow [26] to first select k candidates based on the video-text feature similarity, where the video features are represented by concatenated frame features. We then rerank the selected candidates based on their pairwise Image-Text Matching (ITM) score. We set $k = 64$ for both MSR-VTT and VaTex retrieval.

D Broader Impact

An open-domain few-shot video-language learner has a wide range of beneficial applications for society, such as automatically detecting violent or mature content in videos and helping people with vision impairment understand videos. However, since the language model is pretrained on massive internet-scale text data, there might be unexpected output that can have potential negative impact on the society, such as bias against people of a certain gender, race or sexuality. Future work and dedicated collaboration from the community are needed to alleviate the potential negative societal impact of large language models.

[12] https://storage.googleapis.com/sfr-vision-language-research/BLIP/models/model_base.pth

MSR-VTT Caption

Objects: First, cannelloni. Then, enchilada. After that, competitive eating. Finally, **tex-mex food**.
Events: First, rolled up sleeve. Then, wrapped items. After that, man eats with hands. Finally, men eating.
Attributes: First, black with red sauce. Then, meatfilled. After that, feasting. Finally, holding left overs.
Frame Captions: First, a couple of men sitting at a table with bowls of food. Then, **a table topped with lots of food and condiments**. After that, a couple of men sitting at a table with food...

UniVL: it ' s got a lot of flavor in it it ' s got a lot of flavor..
BLIP: a man and a woman eating food
BLIPcap: a couple of men standing in front of a table filled with food
Ours: two men eat a hearty meal of tex-mex food

Ground Truths:
- two men discuss mexican street food
- two people are sitting in front of a lot of food and talking about it

Objects: First, **autograph**. Then, afro. After that, **fan convention**. Finally, band-aid.
Events: First, hat the girl is wear. Then, child touching. After that, surfers blonde hair. Finally, fans sitting.
Attributes: First, **piece signing**. Then, fans. After that, acoustical. Finally, **sing it loud**.
Frame Captions: First, a group of people standing around a blue table. Then, a group of people on a stage with microphones. Finally, a group of people sitting on stools on top of a stage.

UniVL: a woman ' s wedding ceremony
BLIP: a group of people on stage
BLIPcap: a group of people standing on top of a stage
Ours: a band is signing autographs for their fans

Ground Truths:
- a boy band performs and signs autographs
- a band meeting fans and then performing

Objects: First, **thor**. Then, **avengers**. After that, **ultron**. Finally, **iron man**.
Events: First, cap is red. Then, cap is. After that, there is a statue. Finally, cap is black.
Attributes: First, wearing red shirt. Then, bending his head. After that, cap. Finally, iron.
Frame Captions: First, a man standing in front of a tall building. Then, a scene from the **movie** iron man. Finally, a man with long hair standing in front of a window.

UniVL: a man is playing a man is playing a man ' s game
BLIP: a man with long hair standing in front of a building
BLIPcap: a man standing in front of a tall building
Ours: a man is watching a movie

Ground Truths:
- a man is being interviewed about a movie
- a video about avengers
- chris helmsworth discusses avengers age of ultron

Objects: First, step cutting. Then, laminate flooring. After that, wood flooring. Finally, plywood.
Events: First, floor shows. Then, **floor trim**. After that, **leveled floors**. Finally, man wearing knee pad.
Attributes: First, **covering floor**. Then, **sanding**. After that, push to walk. Finally, stabilizing.
Frame Captions: First, a person standing on a hard wood floor. Then, a person sitting on a couch in front of a sliding glass door. Finally, a man standing on top of a hard wood floor.

UniVL: the first step is to remove the flooring from the floor joist...
BLIP: a person standing on a wooden floor
BLIPcap: a person standing in front of a wooden door
Ours: a man is refinishing a hardwood floor

Ground Truths:
- a man is installing new flooring
- a carpenter places down some wood floring
- a man is fixing the floor

Objects: First, storage chest. Then, ranged weapon. After that, **minecraft**. Finally, meat chop.
Events: First, block the light. Then, mirrored doors. After that, there is a kitchen. Finally, there are 4 vanilla.
Attributes: First, forgotten. Then, dont cross. After that, breaking on left. Finally, clipped in.
Frame Captions: First, **a computer generated image of a room in minecraft**. Then, a computer generated image of a stair case. Finally, **a computer screen shot of a room in minecraft**.

UniVL: a man is playing a game
BLIP: a room in minecraft
BLIPcap: a computer generated image of a bathroom with a toilet
Ours: a man is playing a video game

Ground Truths:
- a man is playing a video game
- a boy eats chicken in minecraft
- gameplay footage of minecraft

Figure 6: Additional qualitative examples on MSR-VTT Captioning.

YouCook2 Caption

Objects: First, **pork loin**. Then, vacuum sealer. After that, plastic wrap. Finally, cooking show.
Events: First, person cutting. Then, grey serving tray. After that, cutting board. Finally, dish he is preparing.
Attributes: First, cutting food. Then, farmers. After that, removing food. Finally, chopped.
Frame Captions: First, a woman in a pink shirt talking to someone ... Finally, a man and a woman preparing food in a field.
Subtitle: You going to **pound** it? We want to **give it a nice whack**. It be like I'm beating that hit it hit it Laura, OK? I think we can feel it now.

UniVL: pound the chicken breast in the pan
Ours: **pound** the **pork loin**

Ground Truths:
add a piece of pork in a ziplock bag and pound it

Objects: First, deep frying. Then, fried food. After that, food warmer. Finally, fried noodles.
Events: First, fried rice. Then, people are eating. After that, fried food. Finally, frying rack.
Attributes: First, **stirfried**. Then, non stick. After that, being cooked. Finally, full of food.
Frame Captions: First, a **pan** of food on a stove top. Then, a **pan** filled with food on top of a stove. Finally, a person putting food into a **pan** on top of a stove.
Subtitle: And **already cooked**. My Udham noodles. Not like this. Oh, I wish I had smellovision This is smelling so good.

UniVL: add some seasoning spice udham noodles and mix in
Ours: add in the **already cooked** noodles to the **pan**

Ground Truths:
add udon noodles to the pan and stir

Vatex Caption

Objects: First, bubble blowing toy. Then, human head. After that, **balloon**. Finally, dog toy.
Events: First, **floating balloon**. Then, head turned. After that, head looking. Finally, **blow up ornament**.
Attributes: First, **helium filled**. Then, head on. After that, head. Finally, rubbery.
Frame Captions: First, **a pink flamingo balloon sitting on top of a table**. Then, a clock tower lit up in the dark. Finally, a blurry image of a person in a body suit.

UniVL: a child ' s head is a child ' s head .
BLIP: a person holding a pink object
BLIPcap: a close up of a person holding a pink object
Ours: A person **blows up a balloon** and ties it off.

Ground Truths:
- A person is twisting blown up balloons into a head figure and then into an animal figure
- A person shows how to make a balloon dog using balloons

Objects: First, spelling bee. Then, **applause**. After that, talent show. Finally, helping hand.
Events: First, to shake hands. Then, students are looking. After that, people are standing. Finally, band is white.
Attributes: First, handing something. Then, highfiving. After that, moving up. Finally, **clapping her hands**.
Frame Captions: First, a group of people standing in front of a crowd of people. Then, a group of people standing in front of a crowd. Finally, a group of people standing in front of a podium.

UniVL: you ' re going to have a great day .
BLIP: a group of people standing in front of a crowd
BLIPcap: a group of people that are standing in front of a microphone
Ours: **A group of people standing in front of a podium, clapping their hands.**

Ground Truths:
- A woman reads off names in front of crowd and young people get a piece of paper, the crowd applauds.

Figure 7: Additional qualitative examples on YouCook2 and VaTex Captioning.

Generate a video caption based on the objects, events, attributes and frame captions. Example:

Objects: First, kaval. Then, special agent. After that, detective. Finally, saw u.
Events: First, to board the plane. Then, wears blue shirt. After that, bus is hyundai. Finally, tour bus has sup.
Attributes: First, coupled. Then, sas. After that, driving away. Finally, expandable.
Frame Captions: First, a woman in a car looking out the window. Then, a car with a lot of fire coming out of it. After that, a man laying on top of a bed next to a pair of glasses. Finally, a woman laying in the grass with her eyes closed.
Video Caption: it is the clips of a movie

Objects: First, animal sports. Then, liger. After that, predation. Finally, lion.
Events: First, zebras are playing. Then, zebra is bent over. After that, zebras are eating. Finally, giraffe's find food.
Attributes: First, next to zebra. Then, trying to eat. After that, looking for meal. Finally, stalking.
Frame Captions: First, a blurry photo of a zebra in a field. Then, a zebra laying on its back in a field. After that, a dog chasing a zebra on the ground. Finally, a lion running through a field with rocks.
Video Caption: a lion is catching an zebra

Objects: First, western screech owl. Then, eastern screech owl. After that, screech owl. Finally, otter.
Events: First, bird has a beak. Then, animated bird. After that, its pupils are green. Finally, eyes are open.
Attributes: First, startled. Then, owl shaped. After that, tawny. Finally, used as number.
Frame Captions: First, a close up of an owl with an open mouth. Then, a close up of an owl with its mouth open. After that, a close up of an owl with a blurry background. Finally, a blurry picture of a dog laying on the floor.
Video Caption: an owl making a weird sound

Objects: First, chopper. Then, billfish. After that, cartoon. Finally, colt.
Events: First, people look surprize. Then, two kids crouched. After that, animated kid. Finally, man watching blond.
Attributes: First, covered with cartoon. Then, gotee. After that, cartoon image. Finally, cartoon.
Frame Captions: First, a cartoon picture of a man and a woman talking to each other. Then, a cartoon picture of a person walking next to a cartoon character. After that, a cartoon of a man with a hat on. Finally, a blurry photo of a room with a trash can.
Video Caption: cartoon children are confronted by bullies

Objects: First, action-adventure game. Then, strategy video game. After that, stage combat. Finally, ac ace.
Events: First, player is running. Then, player leans back. After that, two players move. Finally, screen shows game.
Attributes: First, suspended on corner. Then, remake. After that, suspended in air. Finally, making the ledge.
Frame Captions: First, a screenshot of a video game with a man in a red shirt. Then, a screenshot of a video game with a man on a ledge. After that, a screenshot of a video game with a person on a skateboard. Finally, a screenshot of a video game with a bird in the foreground.
Video Caption: this is a video game being played

Objects: <query objects>
Events: <query events>
Attributes: <query attributes>
Frame Captions: <query frame captions>
Video Caption:

Figure 8: An example of few-shot prompts for MSR-VTT captioning.

Generate a video caption based on the objects, events, attributes, frame captions and subtitle. Example:

Objects: First, velouté sauce. Then, béchamel sauce. After that, stock pot. Finally, chocolate milk.
Events: First, cooking product. Then, canned food. After that, batter wearing. Finally, cleaning liquid.
Attributes: First, greycream. Then, boiled. After that, stirring food. Finally, serving ball.
Frame Captions: First, a piece of cake sitting on top of a stove. Then, a blender filled with liquid on top of a stove. After that, a dirty pan sitting on top of a stove. Finally, a man with a bald head with a small animal on it.
Subtitle: We're going to start adding in some chicken stock, so just add that straight into the pan.
Video Caption: add in some chicken stock in the pan and whisk

Objects: First, egg decorating. Then, egg slicer. After that, baking mold. Finally, hollandaise sauce.
Events: First, measuring cups. Then, scambled eggs. After that, baking cups. Finally, spilled egg yolk.
Attributes: First, spreading butter. Then, purple fondant. After that, yellow frosting. Finally, made of fondant.
Frame Captions: First, a person is mixing a mixture in a bowl. Then, a person using a spoon to mix ingredients in a bowl. After that, a blue tray filled with muffins on top of a counter. Finally, a close up of a person making a muffin.
Subtitle: And cover the hot dog with the rest of the batter. Bake these.
Video Caption: add batter to cover the hot dog

Objects: First, cabbage soup diet. Then, moqueca. After that, west african cuisine. Finally, minestrone.
Events: First, cooked food. Then, stewed vegetables. After that, meal cooked. Finally, cooking product.
Attributes: First, simmering. Then, stirring food.
Frame Captions: First, a pot of food sitting on top of a stove.
Subtitle: Recipe going to put in about a tablespoon of some parsley. You can add a little bit of olive oil at the end. I like the taste of that myself, and I'm going to put some salt in now. Put any salt in yet and I like to wait till the very end because it does make the. Being a little tough, but in about a table.
Video Caption: add parsley olive oil and salt to the pan

Objects: First, tabbouleh. Then, ful medames. After that, fattoush. Finally, israeli salad.
Events: First, veggie topping. Then, cilantro is green. After that, eaten salad. Finally, people are eating.
Attributes: First, mixed into salad. Then, containing salad. After that, removing food. Finally, mixing food.
Frame Captions: First, a person is mixing a salad in a bowl. Then, a person pouring a glass of water into a bowl of food. Finally, a close up of a person mixing a salad in a bowl.
Subtitle: To add the future bread. Just put it on top. Read it. Like today. By the way, this is a salad that can be combined with maybe 5 Dishel Thought.
Video Caption: add pita bread and mix in

Objects: First, mexican food. Then, korean taco. After that, sandwich wrap. Finally, piadina.
Events: First, food is served. Then, hand holding food. After that, quesadilla being cut. Finally, some wrapped food.
Attributes: First, taco shells. Then, preparing food. After that, wrap ad. Finally, wrap.
Frame Captions: First, a woman standing in a kitchen preparing food. Then, a person is putting toppings on a tortilla. After that, a person is putting a tortilla wrap on a plate. Finally, a close up of a person preparing food on a plate.
Subtitle: Easy and then just roll it up. I need to get my Gwacham.
Video Caption: roll the burrito up

Objects: <query objects>
Events: <query events>
Attributes: <query attributes>
Frame Captions: <query frame captions>
Subtitle: <ASR transcript>
Video Caption:

Figure 9: An example of few-shot prompts with ASR input for YouCook2 captioning.

Answer the question based on the objects, events, attributes and frame captions. Example:

Objects: First, jheri curl. Then, orgasm. After that, making out. Finally, special effects.
Events: First, man kissing. Then, kiss the frog. After that, lips are together. Finally, couple embracing.
Attributes: First, bitten off. Then, pursing lips. After that, one of a couple. Finally, 70s.
Frame Captions: First, a man kissing a woman on the cheek. Then, a woman talking on a cell phone in a room.
Question: what are kissing each other?
Answer: couple

Objects: First, boiled egg. Then, pickled egg. After that, egg slicer. Finally, egg shaker.
Events: First, cutting pizza. Then, chopped onions. After that, prepairing food. Finally, scambled eggs.
Attributes: First, soft boiled. Then, egg shaped. Finally, cutting food.
Frame Captions: First, an egg being sliced on a cutting board with a knife. Then, a person cutting an egg on a cutting board. After that, a person peeling an egg on a cutting board. Finally, a person is peeling an egg on a cutting board.
Question: what is a woman chopping?
Answer: egg

Objects: First, roar. Then, lion. After that, brazilian terrier. Finally, akbash dog.
Events: First, child is playing. Then, moving her tail. After that, running dog. Finally, dog running.
Attributes: First, playing catch. Then, cub. After that, eager to play. Finally, jumping up.
Frame Captions: First, a baby lion playing with a toy in the grass. Then, a small lion cub playing with a ball. Finally, a baby lion standing on top of a lush green field.
Question: what does a lion try to climb over?
Answer: wall

Objects: First, vibraslap. Then, indian musical instruments. After that, sound engineer. Finally, saxophonist.
Events: First, grooves for racks. Then, man performing. After that, man is playing. Finally, pipes connected.
Attributes: First, blue sunny. Then, hifi. After that, playing music. Finally, sax.
Frame Captions: First, a man playing a saxophone in a living room. Then, a man playing a musical instrument in a living room.
Question: what did the man play the sax in?
Answer: room

Objects: First, shoe care. Then, hairstyling product. After that, waxing kit. Finally, printer maintenance kit.
Events: First, wearing shin guards. Then, guy dipping. After that, man is preparing. Finally, man prepares food.
Attributes: First, recycling symbol. Then, dry in background. After that, cleaning supplies. Finally, odor diffuser.
Frame Captions: First, a man sitting at a table with a lot of bottles on it. Then, a man sitting at a table with a bunch of bottles on it.
Question: what is the man doing?
Answer: use

Objects: <query objects>
Events: <query events>
Attributes: <query attributes>
Frame Captions: <query frame captions>
Question: <query question>
Answer:

Figure 10: An example of few-shot prompts for MSVD question answering.

Predict what is more likely to happen next based on the frame captions and dialogue. Example:

Frame Captions: First, a man in a black shirt is serving food. Then, a group of people in a kitchen preparing food. After that, a man standing in front of a large pan filled with food. Finally, a frying pan filled with corn and a spoon.
Dialogue: I took a portion of it, you can smell the onions. You can smell kind of a spice like cinnamon a little bit. He's adding some oil, he's adding in some more onions. Oh, man. speaking in foreign language. - [Mark] Ah, wow, it smells so good. And he's actually gonna make it into a sandwich.
Question: What is more likely to happen next? A: He puts the hot sauce on the food. **B:**The chef will serve the food.
Answer: The chef will serve the food.

Frame Captions: First, a close up of a person wearing a suit and tie. Then, a close up of a person with long hair. Finally, a man sitting at a table with a chess board in front of him.
Dialogue: Beckett : Central Park, Washington Square Park. Beckett : Those would be great places to meet up with someone. Beckett : without drawing attention. Castle : Exactly. Now what if each piece stood for the first letter of a word? Bishop for "B." Pawn for "P"? Okay, "B" and then seven spaces. That could be Brooklyn. And Blakely made his phone call from Brooklyn. So, Brooklyn, B-B-P, Brooklyn Bridge Park? That meeting is at 5 : 00. That's in half an hour. Castle : If Blakely shows, we can find out what Pandora is and we can find Gage. Castle : What do you say? Beckett : Blakely should have been here by now. Beckett : Maybe he knows that Tracy's dead. Beckett : or maybe Gage already killed him. Castle : Choose the audacity of hope. I say he'll be here. Then shouldn't you call Sophia? Castle : And look like an a ss if I'm wrong? Castle : You know, I have to admit, Beckett : I'm actually kind of surprised that you've never mentioned her before.
Question: What is more likely to happen next? A: Beckett plays a move in the game. **B:**Beckett pulls out a knife out of his shoe
Answer: Beckett plays a move in the game.

Frame Captions: First, a man standing at a podium in front of a class. Then, a woman sitting in a chair in front of a group of people. After that, a woman with long blonde hair sitting in front of a man. Finally, a man standing at a podium in front of a monitor.
Dialogue: Ted : Thank you! Ted 2030 : Now... Professor Mosby had arrived. Of course, if I had taken that girl's question... who, by the way,was not your mom. your mom was sitting... Wait, let me finish this story real quick. Here's what that girl would have said. Blond girl : I'm sorry to bother you, Professor Mosby, Blond girl : but this isn't Architecture 101. Blond girl : This is Economics 305. Blond girl : You're in the wrong classroom. Yes, I was in the wrong classroom. And thus began. the most humiliating seven minutes of my life. Ted : Here's your think-about-it for the day. Ted : Every single person in this room. Ted : is already an architect. A girl : Architect?
Question: What is more likely to happen next? A: Marshall reads a letter that brings him to tears. **B:**Marshall reads the note to Lily.
Answer: Marshall reads the note to Lily.

Frame Captions: First, two men standing in a kitchen preparing food. Then, a couple of men standing next to each other in a kitchen. After that, two men standing in front of a large pan of food. Finally, a large pot of food on a table.
Dialogue: The ultimate mutton karahi oh look at that.
Question: What is more likely to happen next? A: The hosts tell the viewers how good the lobster is **B:**The host tells the camera "I'm ready to try it out".
Answer: The host tells the camera "I'm ready to try it out".

Frame Captions: First, a man standing next to a desk in a room. Then, a woman in a red jacket talking to a man. Finally, a woman talking to a man in a room.
Dialogue: Stuart : - Hey, Leonard. - Oh, hi. - How's it going? - Good, good. Leonard : - You? - Fine. - Oh, yeah, hey, can I ask you something? - Sure. Penny : You know your friend Stuart? Sheldon : Yes. Penny : Well, he asked me out again and I said yes. Penny : And then I started thinking maybe I should talk to you first. - About what? - Well, does it bother you? Penny : Me going out with one of your friends?
Question: What is more likely to happen next? A: The girl will agree with Leonard and ask good follow up questions. **B:**Leonard says no it doesn't bother him in an awkward way.
Answer: Leonard says no it doesn't bother him in an awkward way.

<Omit five examples here>

Frame Captions: First, a woman sitting on a couch holding a bouquet of flowers. Then, a woman sitting in a chair talking to a man. After that, a woman in a gold dress sitting on a couch. Finally, a woman sitting on a couch holding a remote control.
Dialogue: Phoebe : So you two were married, huh? Phoebe : What happened, you just drift apart? Do you remember our wedding day? Did you know I slept with the best man? Yes, he told me. At least I think that was what he said. It was difficult to understand with his legs wrapped around my head. Mrs. Geller : Here comes the bride. Phoebe : Oh, my God, Monica!
Question: What is more likely to happen next? A: People in the room will tell Monica that she is pretty. **B:**Monica will claim this is the best day ever.
Answer:

Figure 11: An example of few-shot prompts for video-language event prediction (VLEP) task.

Temporal-aware Prompt

Generate a video caption based on the objects and frame captions. Example:

Objects: First, closed door. Then, door handle. Then, opened door. Finally, room.
Frame Captions: First, man standing next to a wooden door. Then, a close view of a door handle. Then, man in a room.
Video Caption: A man is standing in front of a closed door, he reaches for the handle and opens it, and then he walks into a room

Objects: First, room. Then, opened door. Then, door handle. Finally, closed door.
Frame Captions: First, man in a room. Then, a close view of a door handle. Then, man standing next to a wooden door.
Video Caption: A man in a room walking towards the opened door, he reaches for the handle and closes it.

Objects: First, night sky. Then, sun moving.
Frame Captions: First, stars shining at night. Then, sun over an ocean.
Video Caption: **The stars are shining at night, and the sun is rising over the ocean.**

Static Prompt

Generate a video caption based on the objects and frame captions. Example:

Objects: closed door, door handle, opened door, room.
Frame Captions: man standing next to a wooden door. a close view of a door handle. man in a room.
Video Caption: A man is standing in front of a closed door, he reaches for the handle and opens it, and then he walks into a room

Objects: room, opened door, door handle, closed door.
Frame Captions: man in a room. a close view of a door handle. man standing next to a wooden door.
Video Caption: A man in a room walking towards the opened door, he reaches for the handle and closes it.

Objects: night sky, sun moving.
Frame Captions: sun over an ocean, stars shining at night.
Video Caption: **The sun is setting over the ocean as the stars start to shine in the night sky.**

Figure 12: Full prompt of the "Sunrise" scenario for the example shown in Figure 3 in the main body. We show the impact of temporal-aware prompt on capturing temporal dynamics in videos. The sentences in blue following "Video Caption:" are generated by GPT-3. Text marked in green indicates the generated caption is semantically coherent with the given objects and frame captions, while text marked in red indicates incorrectness.