Video ReCap: Recursive Captioning of Hour-Long Videos

Md Mohaiminul Islam[1] Ngan Ho[1] Xitong Yang[2] Tushar Nagarajan[2]
Lorenzo Torresani[2] Gedas Bertasius[1]
[1]UNC Chapel Hill [2]Meta AI
https://sites.google.com/view/vidrecap

Abstract

Most video captioning models are designed to process short video clips of few seconds and output text describing low-level visual concepts (e.g., objects, scenes, atomic actions). However, most real-world videos last for minutes or hours and have a complex hierarchical structure spanning different temporal granularities. We propose Video ReCap, a recursive video captioning model that can process video inputs of dramatically different lengths (from 1 second to 2 hours) and output video captions at multiple hierarchy levels. The recursive video-language architecture exploits the synergy between different video hierarchies and can process hour-long videos efficiently. We utilize a curriculum learning training scheme to learn the hierarchical structure of videos, starting from clip-level captions describing atomic actions, then focusing on segment-level descriptions, and concluding with generating summaries for hour-long videos. Furthermore, we introduce Ego4D-HCap dataset by augmenting Ego4D with 8,267 manually collected long-range video summaries. Our recursive model can flexibly generate captions at different hierarchy levels while also being useful for other complex video understanding tasks, such as VideoQA on EgoSchema. Data, code, and models are publicly available at `https://sites.google.com/view/vidrecap`.

1. Introduction

Many videos in the real world exhibit a hierarchical information structure that spans human behaviors at different temporal granularities (i.e., atomic actions, intermediate activity steps, long-term goals, etc.). However, most modern video captioning models ignore hierarchical video structure and are specifically tailored for short video inputs, typically limited to 5-15 seconds [3, 13, 20, 32, 35, 36, 40, 43, 47, 48, 54, 60]. These short-range captioning methods capture atomic actions and low-level visual details, such as objects and scenes. Moreover, these models are often prohibitively resource-intensive when applied to longer videos, making them ill-suited for understanding human activities occurring over long periods (e.g., several hours) [26, 43, 48, 60].

In this paper, we investigate a hierarchical video captioning task requiring generating captions at multiple hierarchy levels given a long video input (e.g., several minutes to several hours). Studies in psychology [8, 10, 15] and social cognitive theories [4] have shown the inherent hierarchical structures of human behavior, consisting of atomic actions at the lowest level, intermediate steps in the middle and overall goals/intents at the highest level of the hierarchy. Inspired by these prior studies, we also assume three levels of hierarchies for our video captioning task. At the most granular level, video captions describe individual frames or short video clips of several seconds, focusing on low-level visual elements such as objects, scenes, and atomic actions. As we move up the hierarchy, the short-term captions coalesce into medium-length video segment descriptions spanning activities extending beyond brief moments, such as the intermediate steps within broader activities (e.g., a single step in a cooking recipe) or short segments or sequences within a more extended storyline (e.g., a several minute-long scene within a movie). Lastly, the top level of the hierarchy encapsulates the long-term human goals in the video, intricate relationships between events and characters, and the overarching purpose behind the video, which can be captured via long-range video summaries (See Figure 1).

The task of hierarchical video captioning poses several technical challenges. Firstly, it necessitates models capable of handling vastly different input lengths, ranging from a few seconds to several hours. This contrasts with most existing methods, designed for fixed video durations of up to a few minutes. Secondly, long-range videos are highly redundant, requiring the model to aggregate only essential information while discarding unimportant visual cues. Thirdly, another critical challenge is comprehending the hierarchical structure in long videos and leveraging the synergy between distinct hierarchies.

To address these technical challenges, we propose Video ReCap, a model capable of processing videos of dramatically different lengths where input time spans may differ by up to three orders of magnitude (from a handful of seconds

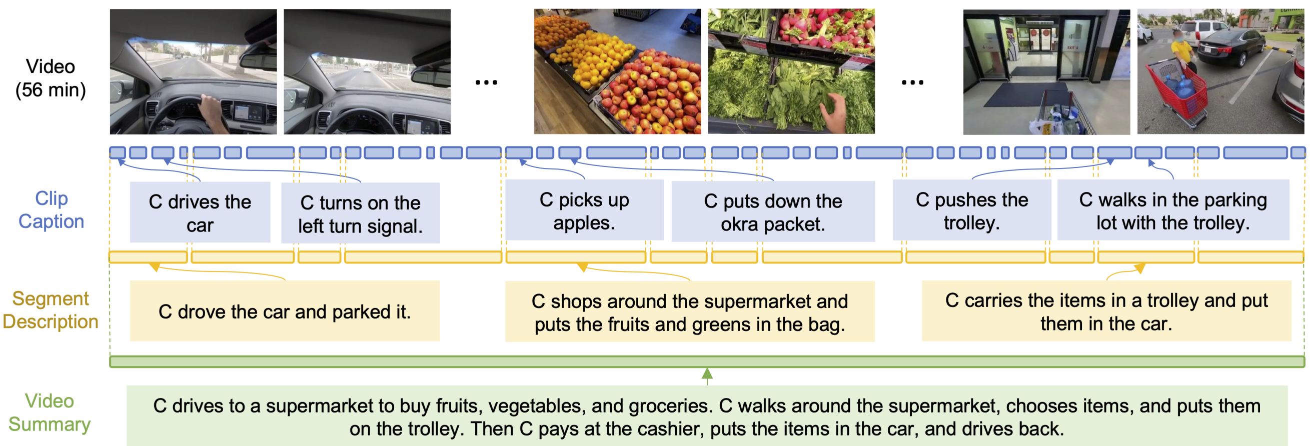

Figure 1. **Hierarchical Video Captioning.** We aim to generate hierarchical captions for a long-range video (e.g., several hours long) at three temporal granularities. First, we generate short clip captions for each few seconds of the video focusing on atomic human actions. Afterward, we produce medium-length segment descriptions for every few minutes of the video, capturing the intermediate steps within a longer activity or a video segment within an extended storyline. Finally, our method generates a summary for a long-range video depicting the overall intent and goals of the actors in the video.

to a few hours) and generating captions at multiple hierarchy levels. Our model encompasses three key attributes that empower its hierarchical video captioning capability. Firstly, Video ReCap adopts a recursive video-language architecture, allowing it to generate captions across distinct hierarchical tiers. At the first level, the model generates captions from features extracted from short video clips, typically lasting a few seconds. As we move up the hierarchy, the model uses sparsely sampled video features and captions generated at the previous hierarchy level as inputs to produce video captions for the current hierarchy level. Such a recursive design effectively leverages the synergy between different video hierarchies and allows us to handle very long video inputs (e.g., up to 2 hours) efficiently. Moreover, it facilitates our model to leverage the powerful reasoning abilities of modern LLMs. Secondly, we implement a curriculum learning scheme, commencing with training on short video clip captions and progressively incorporating data from higher-level hierarchies, namely medium-length segment descriptions and long-range video summaries. Such a hierarchical curriculum learning strategy allows the model to gradually learn the hierarchical structure of the video, starting from short low-level captions to long high-level video summaries. Thirdly, to mitigate the challenge of limited manually annotated hierarchical captioning data, we use LLMs to generate pseudo-summary data spanning different temporal lengths and then use these pseudo-annotations as additional data to train our model.

To evaluate Video ReCap, we introduce Ego4D-HCap dataset, a new hierarchical video captioning benchmark that contains long-range egocentric videos lasting up to several hours with manually annotated captions at multiple hierarchical levels. To build Ego4D-HCap benchmark, we utilize Ego4D [19], the largest publicly available long-range egocentric video dataset, which provides time-stamped captions and video-segment summaries of up to 5 minutes. We then augment the subset of Ego4D videos with manually annotated 8,267 long-range video summaries, where each video spans up to two hours. Consequently, the Ego4D-HCap becomes a rich resource with three levels of hierarchical captions for long untrimmed egocentric videos, encompassing captions for short clips, intermediate descriptions for few-minute video segments, and video-level summaries for long video sequences.

Our results show that Video ReCap outperforms strong prior video captioning baselines [27, 64] across all three temporal hierarchies by a large margin. We also demonstrate that Video ReCap can be effectively used for other complex video understanding tasks, such as long-form video question-answering on EgoSchema [33] where our approach outperforms the previous best method by a substantial margin (**+18.13%**).

2. Related Works

Video Captioning Methods. Early works in video captioning used template-based approaches [23, 25, 40, 46, 58]. Subsequently, these methods were replaced by deep learning methods built using CNN-RNN encoder-decoder architectures [7, 16, 34, 35, 44, 52, 53, 61]. The recent introduction of Transformer [17, 50] led to a plethora of transformer-based video captioning methods [7, 20, 26, 35, 36, 43, 44, 48, 53, 60]. Though these approaches have shown great success in short clip captioning, most are limited to short videos of a few seconds and cannot generate captions spanning multiple temporal hierarchies for hour-long videos.

Video Captioning Datasets. Most existing video captioning datasets contain short video clip inputs (5-30 seconds) [12, 39, 55, 57]. There exist several datasets with longer videos of 1-5 minutes [21, 24, 65], but the captions of these datasets still focus on short-term visual concepts (e.g., atomic actions, presence of objects, etc.). Instead, our work aims to develop models and datasets for hierarchical video captioning that spans multiple temporal granularity levels ranging from short clip captions to long-range video summaries. To do this, we introduce Ego4D-HCap dataset by augmenting Ego4D with long-range video summaries of hour-long videos. This leads to a hierarchical video captioning dataset consisting of short clip captions, medium-range segment descriptions, and long-range video summaries.

Hierarchical Video Understanding. Several recent datasets include hierarchical activity annotations for procedural videos [6, 42, 45, 49, 66]. However, these datasets define a fixed taxonomy for the activity labels of each hierarchy and focus on procedural activity recognition. In contrast, we assume free-form natural language descriptions for multiple levels to capture inherent hierarchical structure in real-world videos (not limited to only instructional videos). Aside from the datasets, several methods [2, 28, 63] learn hierarchical feature embeddings for several-minute-long videos (e.g., 5 minutes). In contrast, our work focuses on generating free-form hierarchical captions for hour-long videos at multiple temporal scales.

3. Technical Approach

3.1. Problem Overview

Given a long, untrimmed video input, we aim to generate textual captions at multiple hierarchy levels of the video. Formally, as our inputs, we consider a long-range video sequence $V_i = [I_i^{(t)}]_{t=1,\ldots,T}$ comprised of T RGB frames, denoted by $I_i^{(t)}$. Our goal is then to generate captions at three distinct hierarchical levels: $Y_i^{(\ell)} = [y_{i,j}^{(\ell)}]_{j=1,\ldots,|Y_i^{(\ell)}|}$ for $\ell = 1, 2, 3$, where $y_{i,j}^{(\ell)}$ depicts a j^{th} word in a caption i for the hierarchy level l. Each hierarchy of captions is generated sequentially starting with the short-term video clip captions, $Y_i^{(1)}$, describing fine-grained actions and objects occurring within few seconds intervals throughout the video (e.g., a person picks up an apple in Figure 1). Afterward, the model outputs medium-length segment descriptions $Y_i^{(2)}$, which capture intermediate steps or summaries unfolding over a few minutes of the video (e.g., a person driving a car and parking it in Figure 1). Finally, the model finishes its generation with long-range video summaries $Y_i^{(3)}$ representing video content for the entire video input.

3.2. Recursive Video-Language Model

We now describe the Video ReCap model, which contains three high-level components: a Video Encoder, Video-Language Alignment, and a Recursive Text Decoder. We illustrate our approach in Figure 2 and describe each component below.

Video Encoder. First, we utilize an off-the-shelf video encoder (e.g., TimeSformer [9]) to extract features from a long-range video. Given a short video clip, the video encoder outputs dense spacetime features. We divide the entire video uniformly and extract a sequence of features $X_i = [x_{i,j}]_{j=1,\ldots,|C|}$, where $|C|$ is the number of video clips, $x \in \mathbb{R}^{F\times H\times W\times D}$ is the spatiotemporal features of a particular clip, F is the number of frames, H is the height, W is the width, and D is the feature dimension. We use dense spacetime features for short-clip captions so that the model can identify low-level visual cues (i.e., objects and atomic actions); for higher-level captions (e.g., segment descriptions and video summaries), we use global features (e.g., CLS features) to reduce the computational cost and capture the global properties of long video inputs.

Video-Language Alignment. Next, we utilize a Video-Language (VL) Alignment module which takes the video features, X_i and the captions generated in the previous hierarchy $Y_i^{(\ell-1)}$ as input and outputs a fixed number of embeddings $Z_i = [z_{i,j}]_{j=1,\ldots,|Z|}$, where $z \in \mathbb{R}^{D_z}$, $|Z|$ is the number of embeddings, and D_z is the hidden dimension. The objective of the alignment module is to map the video and text features to the joint feature space so that the subsequent text decoder can jointly process both features as in [27]. Moreover, this scheme enables us to compress a large number of video and text features (e.g., several thousand) into a small set of embeddings (e.g., 256), dramatically reducing the computational cost. In particular, we use a frozen pretrained language model (e.g., DistilBERT [41]) to learn a fixed number of video embeddings from the video features X_i by injecting trainable cross-attention layer inside each transformer block of the LM. We also learn a fixed number of text embeddings from the captions generated at the previous hierarchy $Y_i^{(\ell-1)}$ by using a similar frozen LM with trainable cross-attention layers. Finally, we concatenate the video and text embeddings to get the joint embeddings Z_i, which is used by the subsequent text decoder for generating captions $Y_i^{(\ell)}$. Note that the first hierarchy level (i.e., clip caption) has no text features and uses only video embeddings as Z_i.

Recursive Text Decoder. We use a pretrained language model (e.g., GPT2 [38]) as our recursive text decoder for generating captions at multiple hierarchy levels. The decoder takes the video-text embeddings Z_i produced by the video-language alignment module (described above) and then generates captions Y_i^{ℓ} for the hierarchy ℓ. Note that we use captions generated at the previous hierarchy level $Y_i^{\ell-1}$ as one of the inputs (along with video features X_i), which enables a recursive caption generation pipeline. Note that for short-term caption generation (i.e., Y_i^1), the textual

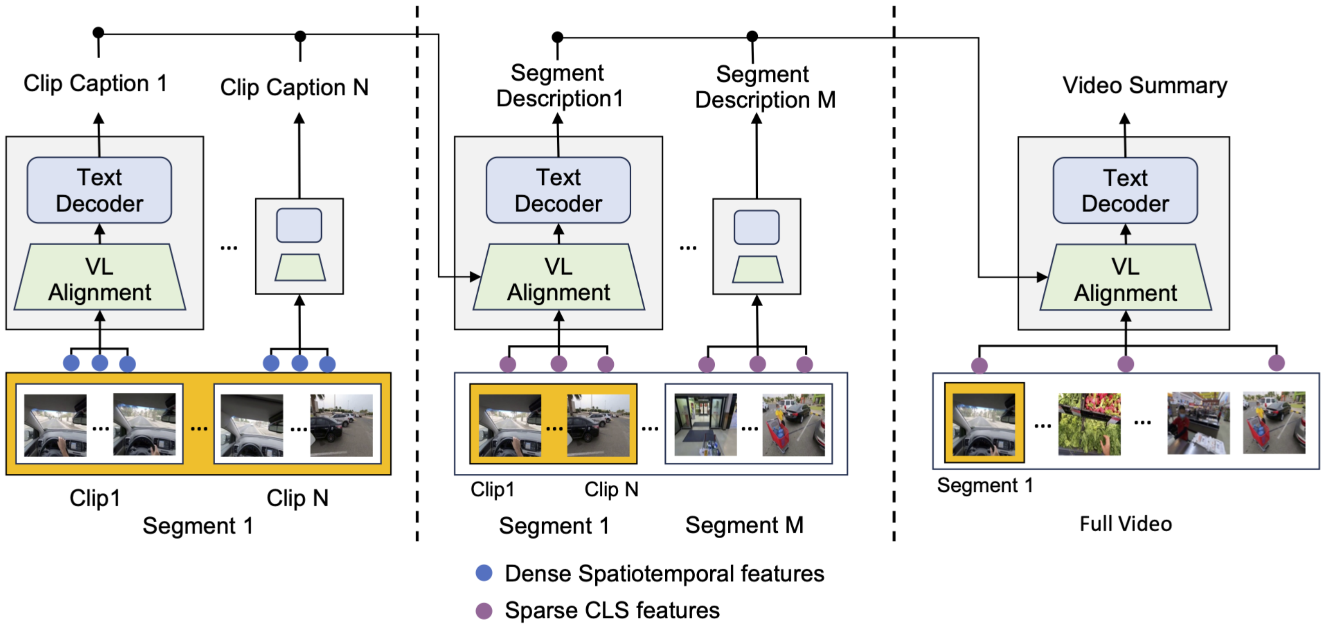

Figure 2. **The Video ReCap model**. **(Left)** First, we generate captions for each short clip (e.g., a few seconds long) of the video using the dense spatiotemporal features extracted by a pretrained video encoder (not shown in the figure). **(Middle)** Then Video ReCap produces segment descriptions for every few minutes of the video using sparsely sampled features (e.g., CLS features) and the previously generated clip captions belonging to a particular segment. **(Right)** Finally, Video ReCap generates the full video summary by utilizing sparsely sampled CLS features from the entire video and the previously generated segment descriptions. The Video-Language (VL) Alignment module maps the video and text features to a joint space so that the subsequent text decoder can jointly process them. Note: the yellow box represents the first segment of the video in each of the three panels, zooming in from right to left.

feature set is initialized as empty (i.e., the base case of our model's recursion). Following prior works [1, 64], we insert trainable cross-attention blocks inside each transformer layer of our textual decoder and freeze the remaining layers. The cross-attention layer attends to video-text embeddings of the alignment module. Therefore, the proposed Video ReCap models the likelihood of caption $Y^{(\ell)}$ conditioned on the video X and the captions generated at lower-level hierarchy $Y^{(\ell-1)}$ using the following training objective:

$$p(Y^{(\ell)}|X) = \prod_{k=1}^{K} p(y_k^{(\ell)}|y_{<k}^{(\ell)}, X, Y^{(\ell-1)}) \tag{1}$$

Here, $y_k^{(\ell)}$ denotes the language token of the caption, $y_{<k}^{(\ell)}$ is the set of preceding tokens, and $Y^{(0)} = \emptyset$.

3.3. Hierarchical Curriculum Learning

Training a recursive video-language model is challenging for several reasons. First, the model must process videos of dramatically different input lengths (i.e., from a few seconds to several hours). Second, there is a significant data imbalance where short-term clip captions vastly outnumber the number of video segment descriptions and long-range summaries. Finally, exploiting the synergy between different hierarchy levels is crucial for generating meaningful and contextually relevant captions. To overcome these challenges, we draw motivation from classic studies of psychology [4, 8, 10, 15], which show a hierarchical organization of human perception of actions. Just as humans first perceive atomic actions before grasping mid-level actions and then infer goals from mid-level activities, our training strategy unfolds in a similar hierarchical fashion. Specifically, our training begins with samples from the lowest hierarchy level, namely clip captions. Subsequently, we train our model with higher-level captions, e.g., medium-length segment descriptions and long-range video summaries. This strategic progression allows the model to gradually understand the intricate hierarchical structure inherent in videos and maximize the synergy between all hierarchies. Moreover, this strategy effectively handles highly imbalanced training data across different hierarchies. Figure 3 shows

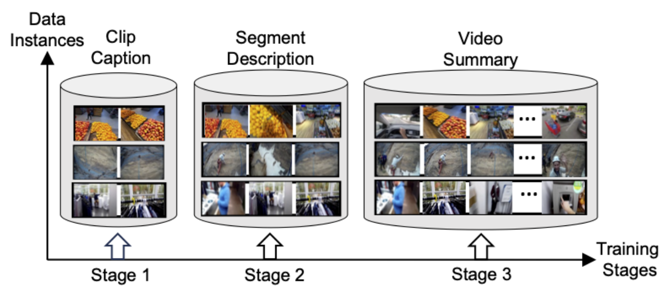

Figure 3. **Hierarchical Curriculum Learning.** We gradually learn the hierarchical structure of the video, starting from short low-level captions to long high-level video summaries.

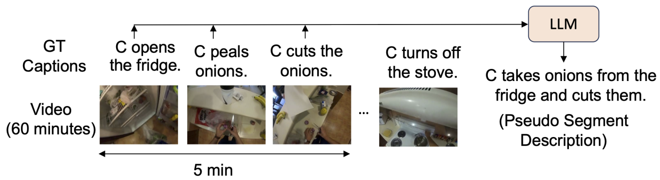

Figure 4. **Large Language Model Supervision.** Given short-term ground truth captions, we use an LLM to generate pseudo-ground truth annotations for medium-length segment descriptions and long-range video summaries to augment our training data.

an overview of the proposed curriculum learning strategy.

3.4. Additional Supervision using Language Models

Collecting captioning annotations for hour-long videos is time-consuming and costly. Thus, another critical challenge associated with hierarchical video captioning is the scarcity of manually annotated hierarchical captioning data, particularly for medium-length segment descriptions and long-range video summaries. We leverage Large Language Models (LLMs) to mitigate this issue. LLMs can effectively incorporate information from text inputs of varying lengths, which aligns perfectly with our objective of guiding the video model to generate captions across multiple hierarchies. Motivated by these insights, we use LLMs to generate a large number of pseudo-caption annotations for medium-length and long-range videos (i.e., our last two hierarchies). The process involves two main steps. First, given manually annotated hierarchical captions, we finetune an LLM teacher to generate medium-length segment descriptions and long-range video summaries from short-term clip captions concatenated across varying temporal durations. Afterward, we use such LLM-generated pseudo ground truth caption data as additional training samples to train Video ReCap (see Figure 4). Our experiments indicate that such pseudo ground truth data generated by LLMs effectively complements manually annotated data and significantly improves our model's captioning ability.

3.5. Implementation Details

We use TimeSformer [9] as our video encoder to extract features that take an input clip of 4 RGB frames of 224×224. We use GPT2 [38] as our default text-decoder, with a hidden dimension of 768 and 12 transformer blocks. We use Adam optimizer [22] with a learning rate of 3^{-5} and a weight decay of 0.01. Our training pipeline also utilized cosine scheduling strategy [31]. Please refer to supplementary materials for additional implementation details.

4. Ego4D-HCap Dataset

We now describe our introduced Ego4D-HCap dataset, a hierarchical video captioning dataset comprised of a three-tier hierarchy of captions: short clip-level captions, medium-length video segment descriptions, and long-range video-level summaries. To construct Ego4D-HCap, we leverage Ego4D [19], the largest publicly available egocentric video dataset. Ego4D videos have several unique features, making them ideal for the hierarchical video captioning task. First, most videos in Ego4D are orders of magnitude longer (e.g., several hours) than the traditional video captioning datasets. Second, egocentric videos typically contain goal-driven and human activities at different hierarchy levels. Third, Ego4D videos capture human behaviors from various scenarios such as cooking, gardening, assembly, etc.

While Ego4D comes with time-stamped atomic captions and video-segment descriptions spanning up to 5 minutes, it lacks video-level summaries for longer video durations. To address this issue, we annotate a subset of 8,267 Ego4D videos with long-range video summaries, each spanning up to two hours. This enhancement provides a three-level hierarchy of captions, making it a perfect resource for validating the effectiveness of our model on the hierarchical video captioning task. In Table 1, we provide a detailed summary of our introduced Ego4D-HCap subset. Please refer to our supplementary materials for a more detailed analysis of the Ego4D-HCap dataset.

Hierarchy Level	# Samples	Avg. Duration
Clip Caption	5.27M	0.96 sec
Segment Description	17.5K	2.87 min
Video Summary	8.3K	28.46 min

Table 1. **Summary of Ego4D-HCap dataset.**

5. Experimental Setup

5.1. Hierarchical Video Captioning Baselines

Hierarchical video captioning is a relatively unexplored task, so there are no well-established baselines for comparing our work. Thus, we introduce the following video-language baselines, which we extend for this task.

- **Zero-Shot Baselines:**
  1. **BLIP2** [27]. A zero-shot baseline for *short-term clip captioning* that utilizes a state-of-the-art image captioning model.
  2. **BLIP2 + GPT3.5** [11, 27]. A zero-shot text-based baseline for *video segment descriptions* and *long-range video summaries*. Given BLIP2-generated captions, it uses GPT3.5 to generate video segment descriptions and long-range video summaries.
  3. **LaViLa + GPT3.5** [11, 64]. Similar to the above, a zero-shot baseline for *video segment* and *summary* generation using LaViLa captions fed into GPT3.5.
- **Finetuned Baselines:**
  1. **LaViLa + GPT2** [38, 64]. A fully-finetuned text-based baseline that takes LaViLa-generated clip cap-

Model	Visual Encoder	Text Decoder	Train Params	Clip Caption		
				CIDEr	ROUGE-L	METEOR
Zero-Shot						
BLIP2 [27]	VIT-G	FT5-XL	0	8.1	7.4	12.7
Finetuned						
LaViLa [64]	TSF-B	GPT2	258M	88.56	47.64	28.03
HierVidCap	TSF-B	GPT2	339M	**98.35**	**48.77**	**28.28**
HierVidCap-U	TSF-B	GPT2	113M	92.67	47.90	28.08

(a) Results for short-range clip captioning.

Model	Video Encoder	Text Decoder	Train Params	Pseudo Ann.	Segment Description			Video Summary		
					C	R	M	C	R	M
Zero-Shot										
BLIP2 [27] + GPT3.5 [11]	VIT-G	FT5-XL	0	✗	5.68	16.87	13.47	11.13	22.41	12.10
LaVila [64] + GPT3.5 [11]	TSF-B	GPT2	0	✗	5.79	19.77	13.45	12.16	24.49	12.48
Finetuned										
LaVila [64] + GPT2 [38]	TSF-B	GPT2	336M	✗	38.22	38.10	16.58	17.98	29.48	12.81
LaVila [64] + FLANT5 [14]	TSF-B	FT5-XL	586M	✗	39.13	38.77	16.88	20.12	30.06	13.17
LaViLa [64]	TSF-B	GPT2	258M	✗	24.63	33.31	15.30	6.54	23.97	10.95
HierVidCap	TSF-B	GPT2	339M	✗	41.74	39.04	18.21	28.06	32.27	14.26
HierVidCap	TSF-B	GPT2	339M	✓	**46.88**	**39.73**	**18.55**	29.34	32.64	**14.45**
HierVidCap-U	TSF-B	GPT2	113M	✓	45.60	39.33	18.17	**31.06**	**33.32**	14.16

(b) Results for medium-length segment description and long-range video summary generation.

Table 2. **Main Results on the Ego4D-HCap dataset.** All results are evaluated in standard CIDEr (C), ROUGE-L (R) and METEOR (M) metrics. We observe several interesting trends. First, finetuned methods perform significantly better than the zero-shot baselines. Second, the Video ReCap model achieves the best results in video captioning across all three hierarchies, surpassing strong prior baselines such as LaViLa [64]. Third, using LLM-generated pseudo annotations leads to a significant boost in performance. Lastly, the unified variant of the model produces competitive results while having a significantly smaller number of trainable parameters than our standard variant.

tions and finetunes a text-only GPT2 model for *segment description* and *video summary* generation while keeping the underlying LaViLa model frozen.

2. **LaViLa + FLAN-T5** [14, 64]. Similar to the above, a fully-finetuned text-based baseline that uses FLAN-T5 rather than GPT2 for *segment description* and *video summary* generation.
3. **LaViLa** [64]. A video-based baseline, finetuned end-to-end to generate *short-term captions, medium-length segment descriptions*, and *long-range video summaries* directly using video inputs. Note that this baseline uses the same video encoder, text decoder, and other experimental settings as our model.

5.2. Our Model Varients

1. **Video ReCap.** This variant of our model uses a shared video encoder but separate text decoders and video-language alignment modules to generate captions at different hierarchy levels (i.e., the weights across different hierarchies are not shared). Due to the increased model capacity of having specialized modules for each hierarchy, this variant typically produces the best performance.
2. **Video ReCap-U.** The unified variant using shared parameters across all hierarchies. Since it has a lot fewer trainable parameters than the previous variant, it is more efficient but performs slightly worse in certain settings.

6. Results and Analysis

6.1. Hierarchical Video Captioning Results

In Table 2, we present our main results for hierarchical video captioning. We use standard captioning metrics, including CIDEr [51], ROUGE-L [29], and METEOR [5] to evaluate our model on the hierarchical video captioning task. Based on these results, we observe several interesting trends. First, we note that zero-shot baselines (e.g., BLIP2 [27], BLIP2 + GPT3.5 [11], LaViLa + GPT3.5) perform considerably worse than the fully finetuned approaches (e.g., LaViLa [64], LaViLa + GPT2 [38], LaViLa + FLAN-T5 [14]), underscoring the significance of in-domain learning on the Ego4D-HCap dataset. Second, we observe that the best performing fully-finetuned text-based baseline LaViLa + FLAN-T5 [14] falls short of our model by 2.61% CIDEr on video segment description and 9.94% CIDEr on video summary generation, despite using significantly more trainable parameters (586M vs 339M). This indicates the benefits of using hierarchical video and text inputs rather than just text for video segment description and long-range video summary generation. Third, we notice that our best performing Video ReCap variant significantly improves upon the strong LaViLa baseline on clip captioning for Ego4D [19], outperforming it by 9.79% CIDEr while employing the same visual encoder, text decoder, and training data as our model. We note that while LaViLa uses a transformer resampler [1, 64], our model utilizes a Language Model-based alignment module (see Section 3.2), which we found very effective for this particular task.

We also note that the performance of LaViLa drops significantly for segment description and video summary generation, indicating its inability to handle long-range videos. In contrast, Video ReCap maintains strong performance on these longer video inputs, outperforming LaViLa by 17.11% CIDEr on segment description and 21.52% CIDEr on video summary generation. We also note that while Video ReCap uses more training parameters than LaViLa (258M vs. 339M), Video ReCap-U has significantly fewer training parameters (113M) than LaViLa but still outperforms LaViLa by substantial margins (+20.97% and +24.50% in CIDEr for segment description and video summary generation respectively). This indicates that the performance gain of our model comes from the recursive and hierarchical design and not from the larger capacity of the model. Our results also indicate that our model's performance can be further improved (5.14% CIDEr in segment description and 1.28% CIDEr in video summary) by incorporating LLM-based supervision (see Section 3.4). Lastly, the last two rows of Table 2 highlight the trade-off between the two variants of our model, i.e., Video ReCap achieves the highest performance across two out of three hierarchies, while the unified variant, Video ReCap-U, attains the second-best performance with significantly fewer trainable parameters.

6.2. Long-Range VideoQA on EgoSchema

Model	Input Feature	Ego4D Pretrain	QA Acc
Random	-	✗	20.0
GPT3.5 [11]	Question	✗	19.57
FrozenBiLM [59]	Video	✗	26.9
VIOLET [18]	Video	✗	19.9
mPLUG-Owl [62]	Video	✗	31.1
InternVideo [56]	Video	✗	32.1
EgoVLP [30]	Video	✓	34.86
EgoVLPv2 [37]	Video	✓	34.12
LaViLa [64] + GPT3.5 [11]	Captions	✓	44.27
Video ReCap + GPT3.5 [11]	Captions	✓	46.03
Video ReCap + GPT3.5 [11]	Hier. Captions	✓	**50.23**

Table 3. **Long-Range VideoQA on EgoSchema [33]** Our approach achieves state-of-the-art results, outperforming the previous best method, InternVideo, by a substantial margin of 18.13%. Furthermore, leveraging the hierarchical captions produced by our model leads to 4.2% and 5.96% boost in performance compared to short-clip captions generated by Video ReCap or LaViLa [64].

In Table 3, we validate the effectiveness of our hierarchical video model on the recently introduced long-range video question-answering (VideoQA) EgoSchma dataset [33]. EgoSchema contains over 5K human-curated multiple-choice question-answer pairs spanning 250 hours of real-world videos, requiring hierarchical reasoning over long videos. We use a simple two-stage approach to perform VideoQA on EgoSchema. First, given long EgoSchema video inputs, we generate hierarchical video captions like before. Afterward, we feed our generated hierarchical video captions as inputs to a text-only GPT3.5 [11] and prompt it to answer a question about a given video in a zero-shot manner. The simple framework performs very well on this benchmark despite the simplicity. We first observe that compared to the variant of our method that uses only short-term captions as inputs to GPT3.5, the variant that uses hierarchical video captions achieves a significant 4.2% boost in performance. We also compare our method with a similar baseline that uses LaViLa-generated short-term captions rather than our hierarchical video captions as inputs to GPT3.5 and show that our approach outperforms this baseline by 5.96%. This highlights the benefits of hierarchical video cues for long-range videoQA. Our results also indicate that our method outperforms the previous best model, InternVideo [56] by a large margin of 18.13%, setting a new state-of-the-art on this benchmark. We note, however, that since InternVideo was never pretrained on Ego4D, the comparison with our approach might be somewhat unfair. Thus, in our comparisons, we also include two recent methods, pretrained on Ego4D, EgoVLP [30] and EgoVLPv2 [37]. Note that for all evaluations, we removed all Ego4D videos used by the EgoSchema benchmark from our training set to avoid data leakage. Compared to EgoVLP and EgoVLP2, our approach still achieves the best results, outperforming these two baselines by a significant margin of 16%, indicating the superiority of our method.

6.3. Ablation Studies

Ablation of Input Modalities. Our model utilizes both video features and recursive text inputs (generated in the previous hierarchy) for the segment descriptions and video summaries. Note that we do not use any text inputs for clip captions as they define the base case of our recursive video model. Since we need to sparsely sample video features to fit long-range videos into GPU memory, we hypothesize that using text as an intermediate representation should complement the sparse video features. In Table 4, we compare our model with a non-recursive baseline (row 2), which only uses sparse video features and a recursive baseline (row 3), which only uses recursive text features. We observe that combining video and text inputs produces a +1.57% boost relative to video-only and a +1.64% boost compared to text-only baselines in CIDEr for segment description generation. Moreover, combining both inputs is more important for long-range video summary generation,

Input	Segment Description			Video Summary		
	C	R	M	C	R	M
Video	40.17	38.65	17.59	25.64	29.61	13.57
Text	40.10	38.02	17.41	23.23	29.17	13.31
Video + Text	**41.74**	**39.04**	**18.21**	**28.06**	**32.27**	**14.26**

Table 4. **Video-Language Input Ablation.** Using both sparse video features and recursive text inputs leads to better performance for both segment description and video summary generation.

Training Scheme	Segment Description			Video Summary		
	C	R	M	C	R	M
Init → Segment	36.81	38.70	17.17	-	-	-
Caption → Segment	**41.74**	**39.04**	**18.21**	-	-	-
Init → Video	-	-	-	8.62	26.33	11.24
Caption → Video	-	-	-	24.84	30.74	13.25
Caption → Segment → Video	-	-	-	**28.06**	**32.27**	**14.26**

Table 5. **Hierarchical Curriculum Learning.** Using the proposed curriculum learning scheme yields a performance boost of +4.93% in segment description and +19.44% in long-range video summary generation compared to training the model from GPT2 pretrained weights (Init).

where video+text inputs provide +2.42% and +4.83% gains compared to video-only and text-only variants. These experiments reveal that the recursive design of Video ReCap that utilizes both video and text input modalities is crucial for the hierarchical video captioning task.

Significance of Hierarchical Curriculum Learning. Next, we investigate the significance of our hierarchical curriculum learning scheme. Table 5 shows the importance of such a curriculum learning scheme. We observe that if we directly train our model on the segment description from GPT2 pretrained initialization, performance drops by a significant margin of 4.93% CIDEr. Moreover, the performance drop is even more catastrophic (-19.44%) for video summary generation without curriculum learning. Finally, we show that it is useful to progressively incorporate higher-level captions, starting from short-term captions, then transitioning to medium-length segment descriptions, and lastly, finishing with long-range video summaries. The variant that progresses from short-term caption to long-range video summary learning directly exhibits a 3.22% drop in CIDEr performance.

Importance of LLM-Based Supervision. Finally, we study the importance of LLM-based supervision for medium-length segment descriptions and long-range video summaries. In Table 6a, we show the performance of different LLM Teachers (e.g., GPT2 [38], and FLAN-T5 [14]) that we use to generate the pseudo ground truth data. We observe that FLAN-T5-Large achieves the best performance in all metrics. Hence, we use FLAN-T5-Large as our Teacher to generate pseudo-ground truth data for segment descriptions and long-range video summaries. Specifically, we produce 100K pseudo-annotations for segment descriptions

LLM	Segment Description			Video Summary		
	C	R	M	C	R	M
GPT2	96.47	46.96	23.13	40.06	33.06	14.76
GPT2-L	104.30	47.68	23.15	43.18	33.86	15.00
FLAN-T5-S	95.61	46.16	22.30	43.27	34.19	14.69
FLAN-T5-L	**125.67**	**50.61**	**26.06**	**52.08**	**36.99**	**19.93**

(a) Training an LLM Teacher.

Pseudo Ann.	Segment Description			Video Summary		
	C	R	M	C	R	M
✗	41.74	39.04	18.21	28.06	32.27	14.26
✓	**46.88**	**39.73**	**18.55**	**29.34**	**32.64**	**14.45**

(b) Supervision Using the best LLM Teacher (FLAN-T5-Large).

Table 6. **Importance of LLM Supervision. Top:** Given ground-truth short-term captions concatenated across varying temporal lengths, FLAN-T5-Large generates the highest quality pseudo-annotations for segment description and long-range video summary annotations. Using this LLM Oracle, we produce 100K pseudo-annotations for medium-length segment descriptions and 15K for long-range video summaries. **Bottom:** Combining LLM-generated annotations with manual annotations during training leads to a performance improvement of 5.14% CIDEr for segment description and 1.28% CIDEr for the video summary.

and 15K for video summaries. We combine these pseudo-annotations with the manually annotated data and train our model. Table 6b shows that utilizing supervision from LLMs provides a substantial performance boost in both segment description (+5.14% CIDEr gain) and video summary (+1.28% CIDEr improvement) generation performance.

7. Conclusions and Future Work

We introduce Video ReCap a recursive video captioning model adept at producing hierarchical captions for videos spanning diverse temporal granularities—from brief clip captions to extensive hour-long summaries. The incorporation of a curriculum learning scheme inspired by human psychology and an LLM-based supervision strategy enhances the model's efficacy in tackling the hierarchical video captioning problem. Beyond its primary focus, our model's hierarchical captions also proves advantageous for long-range video question answering. Additionally, the curated Ego4D-HCap dataset will be released, intended to catalyze ongoing progress in video understanding research. Some promising future directions include real-time caption generation, interactive video understanding, and video-based dialoguing.

Acknowledgements. We thank Feng Cheng, Yan-Bo Lin, Ce Zhang, Yue Yang, and Soumitri Chattopadhyay for their helpful discussions. Authors from UNC Chapel Hill were supported by the NIH Award R01HD11107402, Sony Faculty Innovation award, Laboratory for Analytic Sciences via NC State University, and ONR Award N00014-23-1-2356.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 4, 7

[2] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023. 3

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 1

[4] Albert Bandura. Social cognitive theory: An agentic perspective. *Asian journal of social psychology*, 2(1):21–41, 1999. 1, 4

[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, 2005. 6

[6] Siddhant Bansal, Chetan Arora, and CV Jawahar. My view is the best view: Procedure learning from egocentric videos. In *European Conference on Computer Vision*, pages 657–675. Springer, 2022. 3

[7] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Hierarchical boundary-aware neural encoder for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1657–1666, 2017. 2

[8] R.G. Barker and H.F. Wright. *Midwest and Its Children: The Psychological Ecology of an American Town*. Row, Peterson, 1954. 1, 4

[9] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 3, 5

[10] Matthew Botvinick and David C Plaut. Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological review*, 111(2):395, 2004. 1, 4

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5, 6, 7

[12] David Chen and William B 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, 2011. 3

[13] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 1

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 6, 8

[15] Richard P Cooper and Tim Shallice. Hierarchical schemas and goals in the control of sequential behavior. 2006. 1, 4

[16] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[18] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22898–22909, 2022. 7

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2, 5, 6

[20] Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202, 2017. 1, 2

[21] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. *arXiv preprint arXiv:2011.11760*, 2020. 3

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 5

[23] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. Natural language description of human activities from video images based on concept hierarchy of actions. *International Journal of Computer Vision*, 50:171–184, 2002. 2

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[25] Weiyu Lan, Xirong Li, and Jianfeng Dong. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1549–1557, 2017. 2

[26] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. *arXiv preprint arXiv:2005.05402*, 2020. 1, 2

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with

frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 5, 6

[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 *Conference on Empirical Methods in Natural Language Processing*, 2020. 3

[29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[30] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 7

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 5

[32] 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 arXiv:2002.06353*, 2020. 1

[33] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *arXiv preprint arXiv:2308.09126*, 2023. 2, 7

[34] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1029–1038, 2016. 2

[35] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6504–6512, 2017. 1, 2

[36] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8347–8356, 2019. 1, 2

[37] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 7

[38] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3, 5, 6, 8

[39] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123:94–120, 2017. 3

[40] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE international conference on computer vision*, pages 433–440, 2013. 1, 2

[41] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arxiv 2019. *arXiv preprint arXiv:1910.01108*, 2019. 3

[42] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21096–21106, 2022. 3

[43] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 1, 2

[44] Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. From deterministic to generative: Multimodal stochastic rnns for video captioning. *IEEE transactions on neural networks and learning systems*, 30 (10):3047–3058, 2018. 2

[45] Yale Song, Gene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4d goal-step: Toward hierarchical understanding of procedural activities. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 3

[46] Chen Sun and Ram Nevatia. Semantic aware video transcription using random forest classifiers. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 772–786. Springer, 2014. 2

[47] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 1

[48] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 1, 2

[49] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 3

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[51] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[52] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko.

Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 2

[53] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018. 2

[54] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022. 1

[55] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 3

[56] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 7

[57] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3

[58] Ran Xu, Caiming Xiong, Wei Chen, and Jason Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the AAAI conference on artificial intelligence*, 2015. 2

[59] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *ArXiv*, abs/2206.08155, 2022. 7

[60] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 1, 2

[61] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015. 2

[62] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178, 2023. 7

[63] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *European Conference on Computer Vision*, 2018. 3

[64] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 2, 4, 5, 6, 7

[65] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. 3

[66] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 3