Show, Think, and Tell: Thought-Augmented Fine-Tuning of Large Language Models for Video Captioning

Byoungjip Kim[1], Dasol Hwang[1], Sungjun Cho[1], Youngsoo Jang[1], Honglak Lee[1], Moontae Lee[1,2]
[1] LG AI Research, [2] University of Illinois at Chicago
{bjkim, dasol.hwang, sungjun.cho, youngsoo.jang, honglak, moontae.lee}@lgresearch.ai

Abstract

Large language models (LLMs) have achieved a great success in natural language processing, and have a significant potential for multi-modal applications. Despite the surprising zero-shot or few-shot ability, it is also required to effectively fine-tune pre-trained language models for specific downstream tasks. In this paper, we introduce CaptionT5, a video captioning model that fine-tunes T5 towards understanding videos and generating descriptive captions. To generate a more corespondent caption, CaptionT5 introduces thought-augmented fine-tuning for video captioning, in which a pre-trained language model is fine-tuned on thought-augmented video inputs. This resembles the process that human see a video, think of visual concepts such as objects and actions, and then tell a correct and natural sentence based on the thoughts. To automatically generate thoughts, we propose (1) CLIP-guided thought sampling that samples thoughts based on the similarity in an image-text multimodal embedding space by leveraging CLIP. We also propose (2) CLIP-guided caption ranking during decoding for further performance gains. Through experimentation on VATEX, MSRVTT, and YC2 datasets, we empirically demonstrate that CaptionT5 performs competitively against prior-art video captioning approaches without using encoders specialized for video data. Further experiments show that CaptionT5 is especially effective under small number of sampled video frames.

1. Introduction

Video captioning is a fundamental video-and-language task where the model is asked to generate a textual description that accurately describes the contents of an input video. Despite great potential in many applications including social media, robot vision, and surveillance [16], the task itself remains a big challenge as it requires the model to jointly detect visual cues in the given video and generate natural text from its spatio-temporal interpretation of the content,

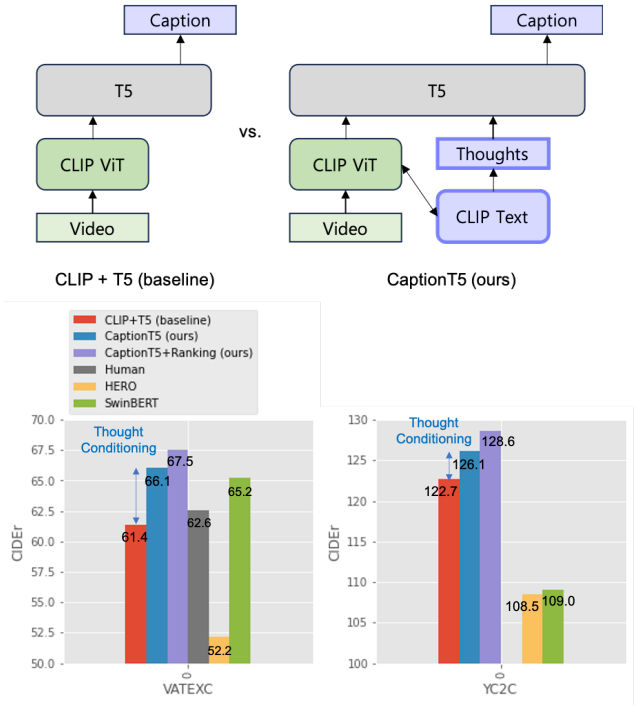


Figure 1. **Main idea of CaptionT5.** CaptionT5 learns to generate video captions by using a pre-trained large language models (T5) and a vision-language model (CLIP). By combining pre-traiend foundation models, CaptionT5 simply achievs human-level performance. In addition to a simple baseline, we introduce *thought-augmented fine-tuning* for video captioning, in which a pre-trained language model is fine-tuned on a dataset consisting of thought-augmented video inputs and corresponding captions. Using the standard benchmarks such as VATEX and YC2, we show that thought-augmented fine-tuning can effectively improve the performance of caption generation.

similar to the "Show and Tell" [33] or "Show, Attend, and Tell" [39] frameworks in image captioning.

While there has been various advances for video captioning, existing work suffer from different drawbacks that limit applicability. For instance, video-and-text multimodal

approaches such as HERO [15] and VL-Adapter [30] proposed using subtitles alongside video inputs [15, 30], but these require that human-annotated subtitles are available at hand, which is not the case with in-the-wild video data. An alternative would be to train and use a video-to-text retriever to extract related text from a large textual corpus [45], but any mismatch between the retrieved text and given video may lead to inaccurate captioning, limiting its performance and generalizability. More recently, SwinBERT [18] proposed employing a large-scale pre-trained Video Swin Transformer [20] to extract useful video frame features for video captioning, which led to state-of-the-art performance without the help of auxiliary textual input. However, this approach requires a video encoder pre-trained on large-scale general-purpose video datasets as well as dense video frame sampling for high quality inference, both of which incur lack of accessibility and high computational burden for practitioners.

Considering wide accessibility of large language models (LLMs) [3, 6, 25, 27] that can generate plausible text based on rich knowledge learned through training on large textual corpora, leveraging LLMs towards video captioning is relatively less explored. As video captioning requires accurate reasoning of interactions between objects and its surrounding in the video, it is evident that the power of LLMs in few-shot in-context learning [3] and prompt-based reasoning [35, 40, 42] would be of great use. The main challenge then resides in how to effectively tune pre-trained LLMs into understanding video data, as they are not previously trained on visual inputs.

In light of such challenge, we propose CaptionT5, a video captioning model that fine-tunes T5 [27] to understand video input and generate its corresponding caption (Figure 2). To generate a more corespondent caption, CaptionT5 introduces *thought-augmented fine-tuning for video captioning*, in which a pre-trained language model is fine-tuned on a dataset consisting of thought-augmented video inputs and corresponding captions. To automatically generate thoughts, we propose (1) *CLIP-guided thought sampling* that samples thoughts based on the similarity in an image-text multimodal embedding space by leveraging CLIP. CaptionT5 overall resembles how humans observe videos by a) detecting visual cues, b) developing conceptual thoughts regarding observed objects as well as their behaviors, and c) constructing a natural sentence that accurately captures those thoughts, namely a "Show, Think, and Tell" pipeline. We also propose a post-processing step with (2) *CLIP-guided caption ranking* again using CLIP as guidance for further improvement in inference quality.

Our experiments on standard benchmarks such as VATEX [34], MSRVTT [38], and YC2 [46] show that CaptionT5 achieves comparable or better accuracy than well-known video captioning models such as HERO [15] and SwinBERT [18]. More interestingly, CaptionT5 significantly outperforms SwinBERT given only a small number of video frames (2∼4 frames), demonstrating its high computational efficiency. We also provide comprehensive empirical analyses on how each component of CaptionT5 affects performance as well as how well CaptionT5 performs in a single-frame setting analogous to image captioning.

In summary, our main contributions are as follows:

- We introduce CaptionT5, a *thought-augmented fine-tuned language model* for video captioning. Thought-augmented fine-tuning aligns a pre-trained language model on a paired dataset consisting of thought-augmented video inputs and corresponding captions. (see Figure 2).
- To automatically generate thoughts, we propose (1) *CLIP-guided thought sampling* that samples thoughts based on the similarity in an image-text multimodal embedding space by leveraging CLIP. We also propose (2) *CLIP-guided caption ranking* during decoding for further performance gains.
- Using standard benchmarks including VATEX, MSRVTT, and YC2, we empirically show that CaptionT5 can achieve comparable or better accuracy than well-known video captioning models such as HERO and SwinBERT. (see Figure 1 and Table 1).

2. Related work

Large language models. Large language models (LLMs) are Transformer-based [31] models pre-trained on a large-scale text corpus. LLMs are typically pre-trained via self-supervised learning such as next token prediction and masked language modeling, capable of leveraging large-scale in-the-wild textual data without any human annotated labels. There are numerous LLMs available, examples of which include BERT [6], GPT [3, 25], T5 [27], PaLM [4], and Chinchilla [10]. Pre-trained LLMs exhibit interesting and unique abilities, such as GPT-3 [3] demonstrating few-shot in-context learning that can achieve competitive performance without fine-tuning to the given task. Recent work such as Chain-of-Thought [35], STaR [42], and ReAct [40] have also showed that thought-augmented prompting allows LLMs to solve complex downstream tasks such as arithmetic, commonsense, and symbolic reasoning. Inspired by such success, we introduce a framework that leverage prompt engineering on LLMs towards video captioning and explore how to effectively construct video-to-text multimodal prompts.

Vision-and-language models. Combining two different modalities, vision-and-language models (VLMs) [17, 22, 29] aim to jointly learn vision-language representations that can be used for downstream tasks such as text-based im-

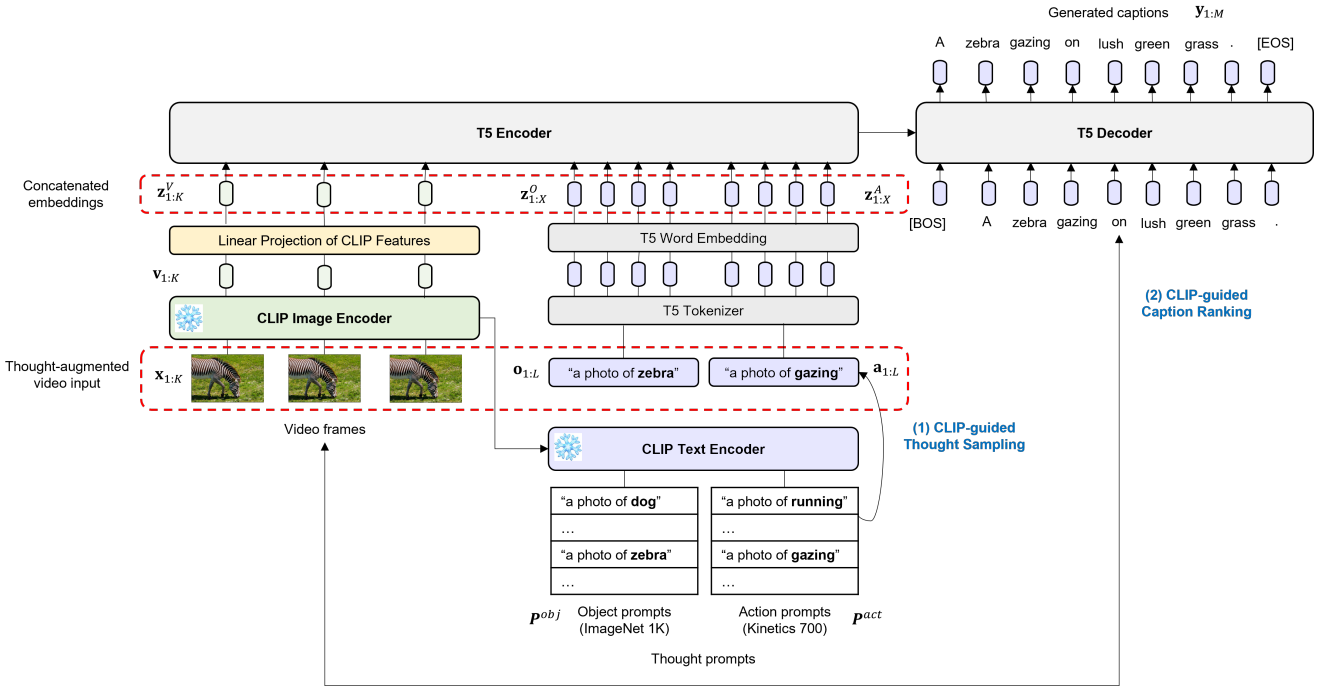


Figure 2. **Overview of CaptionT5.** CaptionT5 is a video captioning model that fine-tunes T5 [27] towards understanding video input and generating its corresponding caption. To generate a more corespondent caption, CaptionT5 introduces *thought-augmented fine-tuning for video captioning*, in which a caption is generated conditioned on the thought-augmented video input. To automatically generate thoughts, we propose *(1) CLIP-guided thought sampling* that samples thoughts based on the similarity in an image-text multimodal embedding space by leveraging CLIP [26]. To further improve caption generation, we provide *(2) CLIP-guided caption ranking* that selects the most correspondent caption among multiple generations by using the video-text similarity.

age retrieval, visual question answering (VQA), and image captioning. Based on the training mechanism, VLMs can be categorized into two groups: MLM-based models and contrastive learning-based models. MLM-based models include ViLBERT [22], VL-BERT [29], and Oscar [17]. Contrastive learning-based models include CLIP [26] and ALIGN [11]. Among them, CaptionT5 uses CLIP as the base VLM to encode video frames and sample contextual texts due to its strong zero-shot performance.

Video-and-language understanding. Extending VLMs to video data by viewing videos as a sequence of image frames, Video-and-language (VidL) models aim to infer textual information while interpreting spatio-temporal relations of objects in the video. Previous work include ActBERT [47], ClipBERT [13], VideoCLIP [37], Frozen in Time (FiT) [1], VIOLET [8], and MERLOT [43]. While these works mainly focus on discriminative downstream tasks such as video question answering and text-to-video retrieval, CaptionT5 tackles the generative task of video captioning.

Video captioning. Early methods proposed for video captioning such as Open-Book [45], HERO [14], and VL-Adapter [30] proposed multimodal frameworks that take video data as well as related text such as subtitles or sentences retrieved from a text corpus for textual guidance. These approaches lack applicability to in-the-wild video data where subtitles or sentences that exactly align with the contents of the video may not be available. More recently, SwinBERT [18] proposed leveraging a large pre-trained video encoder, achieving state-of-the-art performance on video captioning without any additional text annotations. However, this comes at a cost of requiring a large-scale general-purpose video datasets that are not as accessible as text or image data. Previous work have also shown that SwinBERT requires a large number of video frames for high-quality generation. Our CaptionT5, on the other hand, effectively reduces this computational burden by combining the image-and-text aligned feature space of CLIP and the textual reasoning capability of T5 via video-and-thought prompting.

3. Method

In this section, we describe the problem formulation of video captioning alongside notations used throughout the paper, then provide the details of CaptionT5.

Problem formulation. Video captioning can naturally be formulated as a sequence-to-sequence generation problem. We consider each video as a sequence of K images, denoted as $\mathbf{x}^i_{1:K} = [x^i_1, ..., x^i_K]$, with each image having height H, width W, and C channels (*i.e.,* $x \in \mathbb{R}^{H\times W\times C}$). Similarly, a caption is a sequence of words denoted as $\mathbf{y}^i_{1:M} = [y^i_1, ..., y^i_M]$ with each word belonging to a set of possible words $\mathcal{Y}$ (*i.e.* $y \in \mathcal{Y}$). Given a dataset consisted of video-caption pairs $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}^N_{i=1}$, we can train the model to generate the caption $\mathbf{y}^i$ given video $\mathbf{x}^i$ by using the next-token prediction objective below:

$$\mathcal{L}_{LM} = -\sum_{i=1}^{N}\sum_{j=1}^{M} \log P_\theta(y^i_j|\mathbf{x}^i_{1:K}, \mathbf{y}^i_{1:j-1}). \quad (1)$$

3.1. CaptionT5

Consider how humans would perform the task of video captioning. While watching the video, we typically think of answers to conceptual questions in our own language such as "What objects are shown?" and "What actions do they take?". Given those thoughts, we cohesively merge them together alongside the video, and generate a natural sentence that describes the overall content. Similar to this "Show, Think, and Tell" pipeline, our CaptionT5 performs *video-and-thought prompting* that extends video frame representations with contextual text, or thoughts, to be fed into T5 for effective caption generation. The following subsections describe each of the three steps in detail, and the overall illustration of CaptionT5 can be found in Figure 2.

3.1.1 Show — Video frame encoding

Given an input video $\mathbf{x}_{1:K}$, CaptionT5 first generates a representation for each frame $\mathbf{v}_{1:K}$ by using a frozen CLIP image encoder f with feature dimension D_{CLIP} as follows:

$$\mathbf{v}_{1:K} = [v_1, ..., v_K] = [f(x_1), .., f(x_K)], \quad (2)$$

with $f : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{D_{CLIP}}$.

Since image features from CLIP are not aligned with word embeddings from T5, CaptionT5 trains a linear projection $\mathbf{E}^V$ to align the two features. We also add trainable positional embeddings $\mathbf{E}^V_{pos}$ as features from CLIP do not capture the temporal aspect in videos. The overall video frame encoding procedure can be written as

$$\mathbf{z}^V_{1:K} = [z_1, ..., z_K] = [v_1\mathbf{E}^V, ..., v_K\mathbf{E}^V] + \mathbf{E}^V_{pos}, \quad (3)$$

with $\mathbf{E}^V \in \mathbb{R}^{D_{CLIP}\times D_{LM}}$. $\mathbf{E}^V_{pos}$ and $\mathbf{z}^V_{1:K} \in \mathbb{R}^{K\times D_{LM}}$ where D_{LM} denotes the feature dimension of T5.

3.1.2 Think — CLIP-guided thought sampling

The next step of CaptionT5 samples contextual text prompts that represent conceptual thoughts on the video while leveraging the multimodal feature space of CLIP.

Specifically, we consider two different types of thought prompts: object prompts and action prompts. The object prompt set $\mathcal{O} = \{o_k\}^{N_{obj}}_{k=1}$ contains object prompts of the form "`a photo of {object}`" that describe what objects are present in the video. The action prompt set $\mathcal{A} = \{a_k\}^{N_{act}}_{k=1}$ contains action prompts of the form "`a photo of {action}`" that describe what actions the object is performing. As one of possible implementations, we construct the object prompt set based on 1,000 object categories in ImageNet-1K [28] (i.e., $N_{obj} = 1000$), and the action prompt set based on the 700 action categories in Kinetics [12] (i.e., $N_{act} = 700$). Note that CaptionT5 is not limited to these thought prompt sets, but can be extended to more effective thought prompt sets.

CaptionT5 encodes object prompts o_k and action prompts a_k through the frozen CLIP text encoder g to obtain their respective features $h^{obj}_k = g(o_k)$ and $h^{act}_k = g(a_k)$, where $h^{obj}_k, h^{act}_k \in \mathbb{R}^{D_{CLIP}}$. It then computes two probability distributions, one defined on $\mathcal{O}$ and another on $\mathcal{A}$ based on the similarity between the thought features and the given video features. To represent the entire set of video frames $\mathbf{v}_{1:K}$, we use the mean-pooled image features $\overline{v} = \frac{1}{K}\sum_{i=1}^{K} v_i$. More formally, the video-thought similarity distribution can be defined as follows:

$$P_{obj}(o_k|\mathbf{v}_{1:K}) \approx P_{obj}(o_k|\overline{v}) = \frac{\exp((\overline{v}\cdot h^{obj}_k))/\tau)}{\sum_{l=1}^{N_{obj}}\exp((\overline{v}\cdot h^{obj}_l))/\tau)} \quad (4)$$

$$P_{act}(a_k|\mathbf{v}_{1:K}) \approx P_{act}(a_k|\overline{v}) = \frac{\exp((\overline{v}\cdot h^{act}_k)/\tau)}{\sum_{l=1}^{N_{act}}\exp((\overline{v}\cdot h^{act}_l))/\tau)}. \quad (5)$$

Here, τ is a temperature parameter that is set as 10^{-3} in our experiments.

Given the two probability distributions, CaptionT5 samples L number of prompts from each distribution, resulting in a set of L object prompts and L action prompts, which we use $\mathbf{o}_{1:L} \subset \mathcal{O}$ and $\mathbf{a}_{1:L} \subset \mathcal{A}$ to denote, respectively.

After sampling object and action thought prompts, we concatenate all $2L$ prompts into a single text sequence $[\mathbf{o}_{1:L}, \mathbf{a}_{1:L}]$, then tokenize and featurize the sequence through the learnable T5 embedding layer. As a result, we obtain word embeddings $[\mathbf{z}^O, \mathbf{z}^A] \in \mathbb{R}^{K_T\times D_{LM}}$ as our final thought representation where K_T denotes the total number of text tokens from all the thoughts.

3.1.3 Tell — Caption generation

CaptionT5 concatenates the video frame embeddings $\mathbf{z}^V_{1:K}$ and thought embeddings $[\mathbf{z}^O, \mathbf{z}^A]$ together into a single sequence $\mathbf{z} = [\mathbf{z}^V_{1:K}, \mathbf{z}^O, \mathbf{z}^A]$, and feeds it to the T5 encoder. Given a video-caption pair dataset $\mathcal{D} = \{(\mathbf{x}^i, \mathbf{y}^i)\}^N_1$, the T5

model is fine-tuned by using the following loss function:

$$\mathcal{L}_{VaT} = -\sum_{i=1}^{N}\sum_{j=1}^{M} \log P_\theta(y_j^i | \mathbf{x}_{1:K}^i, \mathbf{o}_{1:L}^i, \mathbf{a}_{1:L}^i, \mathbf{y}_{1:j-1}^i). \tag{6}$$

Note that object prompts $\mathbf{o}_{1:L}^i$ and action prompts $\mathbf{a}_{1:L}^i$ are not part of the dataset $\mathcal{D}$, but are automatically generated by using the CLIP-guided thought sampling.

3.1.4 Reflect — CLIP-guided caption ranking

During inference after training, we can leverage the feature space of CLIP once again for similarity-based caption ranking. Similar to language modeling, the quality of video captioning is largely affected by which decoding algorithm. For example, decoding via beam search leads to generation quality that depends on hyperparameters such as beam length [5]. To address this issue, we generate multiple captions from fine-tuned CaptionT5, and rank the generated captions by using the similarity between the mean-pooled input video frame features $\overline{v}$ from the CLIP image encoder and the features of each generated caption $c_i \in \mathbb{R}^{D_{CLIP}}$ from the CLIP text encoder. Then, CaptionT5 selects the caption with the largest similarity by

$$i^* = \text{argmax}_i(\overline{v} \cdot c_i). \tag{7}$$

As later explored in our experiments, we find that this post-processing step in decoding leads to notable improvements.

4. Experiments

In this section, we provide experimental results on video captioning of CaptionT5. Firstly, we present our experiment setup, and then provide main quantitative results. In addition, we provide a further analysis of the proposed methods. Finally, we provide some example video captions generated by CaptionT5.

4.1. Experimental setup

Datasets. We conduct experiments on two standard open-domain video captioning datasets.

- **VATEX** [34] is a large open-domain video dataset that contains about $41.2K$ video clips. It consists of train, validation, public test, and private test set (for leaderboard), and each set contains $26.0K$, $3.0K$, $6.0K$, and $6.2K$ videos, respectively. Each video clip has 10 ground-truth captions. We used the official train set ($26.0K$ videos / $260.0K$ captions) for training and the public test set ($6.0K$ videos) for evaluation.
- **MSR-VTT** [38] is a open-domain video dataset that contains about $10K$ video clips. Each video clip has 10 ground-truth captions. We used the split knownd as MSR-VTT-1kA [41] that contains $9K$ videos for training and $1K$ videos for evaluation.
- **YC2** [46] is a cooking domain video dataset that contains about 15.4K video clips. Each video clip has one ground-truth caption. We used the official train set ($10.3K$ videos / $3.5K$ captions) for training and the validation set ($3.5K$ videos) for evaluation.

Evaluation metric. Diverse metrics such as BLEU [23], METEOR [2], and CIDEr [32] are used for evaluating the quality of generated video captions. In our experiments, we consider CIDEr [32] as our main evaluation metric, as it is known that semantic similarity-based metrics are more correlated with human evaluation than word matching-based metrics such as BLEU [23]. Furthermore, we use CIDEr [32] for fair comparison as many works such as HERO [15] and SwinBERT [18] have mainly used the metric for video captioning.

Baselines. We compare CaptionT5 against a total of three baselines, two of which are multimodal approaches that assume each video input is accompanied by an additional textual input such as subtitles. The remaining baseline is a video-only approach which does not assume any additional annotations besides the video itself.

- **Open-Book** [45] proposes to use similar sentences retrieved from training corpus. Since it retrieves similar sentences, they are not exactly matched with a given video. Therefore, it can limit the performance and lack the generalization ability.
- **HERO** [14] is another multi-modal baseline that uses subtitles of a video in addition to video frames. Using a combination of ResNet-152 [9] and SlowFast [7] for video encoding, and RoBERTa [19] for text decoding, HERO shows strong performance on various video-and-language tasks such as video retrieval and question answering [15].
- **SwinBERT** [18] is a Transformer-based video captioning model. It employs Video Swin Transformer [20] for encoding a sequence of dense video frames, and a Transformer encoder for generating captions conditioned on the encoded video representations.

Implementation details. We implement CaptionT5 using PyTorch [24], HuggingFace Transformers [36], and OpenAI CLIP [26]. Following VALUE [15], we use the AdamW [21] optimizer with a linear learning rate scheduling after $5K$ warm-up steps. We use a base learning rate of $1.5e-3$ and batch size of 256. Further details on the chosen hyperparameters can be found in the supplementary section.

4.2. Quantitative results

Comparison with other models. Table 1 shows the video captioning performance measurements in CIDEr from Cap-

Method	Vision			Language		Benchmark Score (CIDEr ↑)		
	Video Encoder	Object Detector	# Frames	Contextual Text	Text Decoder	VATEX	MSR-VTT	YC2
Human [15]	-	-	-	-	-	62.7	-	-
ORG-TRL [44]	InceptResnetV2, C3D	Faster R-CNN	≥ 8	✗	LSTM	49.7	50.9	-
Open-Book [45]	InceptResnetV2, C3D	✗	≥ 8	Retrieved sentences	LSTM	57.5	52.9	-
HERO [15]	ResNet-152, SlowFast	✗	≥ 8	Subtitles	ROBERTa	58.1	-	108.5
SwinBERT [18]	Video Swin	✗	8	✗	Transformer Encoder	65.2	47.6	-
SwinBERT [18]	Video Swin	✗	64	✗	Transformer Encoder	72.7	55.3	109.0
CaptionT5 (ours)	CLIP-ViT	✗	8	Objects & Actions (sampled by CLIP)	T5-Large	**67.5**	**66.1**	**128.6**

Table 1. **Comparison of video captioning models.** CaptionT5 achieves comparable or better performance, when compared with the state-of-the-art video captioning models. SwinBERT achieves human-level video captioning performance (CIDEr 62.7) by employing Video Swin Transformer that can encode better video representations. In contrast, CaptionT5 provides better performance than SwinBERT by harnessing large language models (T5) and pre-trained vision-and-language models (CLIP).

	Components			Score (CIDEr ↑)	
	(1) VP	(2) TP	(3) CR	VATEX	YC2
Human [15]	-	-	-	62.6	-
SwinBERT [18]	-	-	-	65.2	108.0
CaptionT5	✓	✗	✗	61.4	122.7
	✓	✓	✗	66.1	126.1
	✓	✓	✓	**67.5**	**128.6**

Table 2. **Effect of thought prompting and caption raking.** Each acronym means the followings: (1) VP: Video Prompting, (2) TC: Thought Prompting, (3) CR: Caption Ranking

tionT5 and other baselines on VATEX and MSR-VTT. Comparing multi-modal models that use subtitles or retrived sentences against the uni-modal ORG-TRL, we find that the additional text is indeed helpful in generating higher quality video captions, but is not sufficient to surpass human-level performance. On the other hand, CaptionT5 shows performance that exceeds human-level performance.

Effect of thought prompting and caption ranking. Table 2 shows the CIDEr results on the VATEX and YC2 test set after incrementally ablating each component in CaptionT5. When removing caption ranking, the performance decreases by 1.4, and the additional removal of the thought prompting procedure decreases it further by 4.7, resulting in sub-human-level performance. This shows that while both thought prompting and caption ranking are helpful for video captioning, the "thinking" step of generating contextual text prompts is more crucial towards high-quality caption generation.

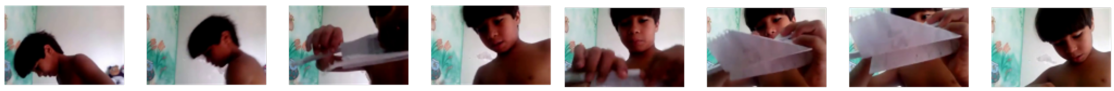
Method	# Video Frames			
	2	4	8	16
SwinBERT [18]	47.4	48.2	65.2	**68.4**
CaptionT5 (ours)	**60.0**	**64.2**	**67.5**	67.5
δ	+12.6	+16.0	+2.3	-0.9

Table 3. **Comparison with SwinBERT on varying number of video frames.** While performing comparably with large number of video frames, CaptionT5 outperforms SwinBERT significantly when the number of frames is small.

Effect of the number of video frames. In Table 3, we further investigate the performance of CaptionT5 under varying number of video frames, with comparisons against SwinBERT. As suggested in the original paper, we find that the performance of SwinBERT consistently decreases as we reduce the number of sampled video frames [18]. Meanwhile, CaptionT5 retains its performance much better, showing a smaller decrease of only 7.5 CIDEr when reducing the number of frames from 16 to 2. This implies that leveraging well-aligned image-and-text feature space leads to more data-efficient video captioning compared to Transformer-based video encoders, that performs well even under sparse frame sampling. This is especially useful in practice as the time and memory cost of Transformer-based models increase quadratically to the total number of tokens.

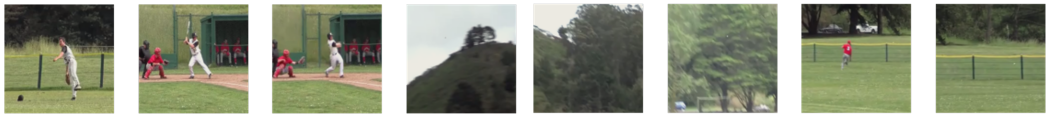
4.3. Qualitative results

We provide qualitative results in Figure 3 and 4. The results are generated by CaptionT5 trained in the setting where 8 video frames are given, and 7 object prompts and 7 action prompts are retrieved. Note that thought prompts are retrieved by similarity sampling for training, but they are retrieved by top-k search for inference to generate determin-



VideoID		G0mjFqytJt4_000152_000162
CaptionT5	Object prompts	A photo of envelope. A photo of carton. A photo of sombrero. A photo of quill, quill pen. A photo of sarong. A photo of wing. A photo of pole.
	Action prompts	A photo of making paper aeroplanes. A photo of ripping paper. A photo of applying cream. A photo of poking bellybutton. A photo of tapping pen. A photo of pinching. A photo of beatboxing.
	Generated caption	**A young boy is demonstrating how to fold a paper airplane.**
SwinBERT		A young boy is showing how to make a paper airplane
GT1		A boy is talking and fiddling with a few pieces of paper in his hands.
GT2		A young boy in his bathroom as he explains how to make a paper airplane.

Figure 3. **Example captions generated by CaptionT5 on VATEX dataset.**



VideoID		video7021
CaptionT5	Object prompts	A photo of baseball. A photo of ballplayer, baseball player. A photo of pole. A photo of swing. A photo of pitcher, ewer. A photo of projectile, missile. A photo of cliff, drop, drop-off.
	Action prompts	A photo of catching or throwing baseball. A photo of hitting baseball. A photo of catching or throwing softball. A photo of throwing ball (not baseball or American football). A photo of throwing axe. A photo of swinging baseball bat. A photo of catching or throwing frisbee.
	Generated caption	**A baseball player is hitting the ball with his bat.**
GT1		Baseball player hits ball.

Figure 4. **Example captions generated by CaptionT5 on MSR-VTT dataset.**

istic results.

In Figure 3, we provide the result on VATEX [34] dataset. Note that VATEX contains the ground-truth captions that are descriptive and long. Despite this complexity, CaptionT5 generates a semantically reasonable caption for a given video. For example, CaptionT5 generates *"A young boy is demonstraing how to fold a paper airplan"*, while the ground-truth is *"A young boy in his bathroom as he explains how to make a paper airplane"*.

To show how effective the sampled thought prompts are, we also provide object and action prompts that are sampled conditioned on a given video. For example, CaptionT5 samples semantically correct object prompts such as ``A photo of wing''. Also, CaptionT5 samples semantically meaningful action prompts such as ``A photo of making paper aeroplanes''.

Figure 4 shows the results on MSR-VTT [38] dataset. Also, we provide sampled thought prompts for each given video, to show the semantic relevance of the sampled thought prompts. CaptionT5 generates a semantically meaningful caption for each given video. For example, CaptionT5 generates *"A baseball player is hitting the ball with his bat"*, while the ground-truth is *"Baseball player hits ball"*. Also, for the same video, CaptionT5 captures semantically meaningful object prompts such as ``A photo of baseball'', and action prompts such as ``A photo of catching or throwing baseball'', respectively

4.4. Ablation study

For further analysis, we perform comprehensive ablation studies on the VATEX dataset to assess how each major component of CaptionT5 contributes to its performance and how different thought retrieval settings and caption sampling methods affect inference quality.

Effect of thought prompt retrieval methods. Next, we test how the similarity sampling procedure for thought retrieval plays a role in caption quality. Note that for the following ablation experiments, we mainly compare against

Vision	Language		VATEX
# Frames	# Thoughts	**Thought Retrieval**	(CIDEr ↑)
8	-	None	61.42
8	7 × 2	Top-k Search	62.94
8	7 × 2	Similarity Sampling	**66.10**

Table 4. **Effect of thought prompt retrieval methods.**

Vision	Language		VATEX
# Frames	# Thoughts	**Thought Type**	(CIDEr ↑)
8	-	None	61.42
8	7 × 1	Object	65.03
8	7 × 1	Action	65.93
8	7 × 2	Object & Action	**66.10**

Table 5. **Effect of thought prompt types.**

the CaptionT5 baseline without caption ranking which returns 66.1 CIDEr, in order to compare while disregarding the effect from additional post-processing. Table 4 shows that replacing similarity sampling with top-k search leads to a CIDEr, still outperforms the model without any thought retrieval, but performs significantly worse (-3.16 CIDEr) compared to CaptionT5 with similarity sampling. This indicates that the stochasticity from the sampling process acts as an effective regularizer that renders the model more robust and generalizable towards previously unseen videos.

Effect of thought prompt types. We compare the effect of ablating each thought prompt type in Table 5. We find that not considering object thoughts lead to a slightly worse performance than not considering action thoughts. We conjecture that this may be due to video captions focusing more on the action being done over time, rather than the object itself which can easily be identified by a single frame. One other possibility is that predicting the correct object given its action is an easier task for LLMs compared to predicting the action given the object. Nonetheless, considering at least one type of thought prompts lead to a significant gain in performance (+3.61), while using both types return highest-quality video captions.

Effect of the number of thought prompts. Table 6 shows results from varying the number of thought prompts to sample (i.e. L). Interestingly, we find that sampling one object and one action thought prompt still returns a performance of 65.02, which is comparable to SwinBERT (65.2). This indicates that the "thinking" procedure of visual concepts through CLIP is highly effective, and the trained thought prompt distributions are well-trained towards choosing the correct object and action.

Vision		Language	VATEX
# Frames	# Tokens	**# Thoughts**	(CIDEr ↑)
8	8	-	61.42
8	8	1 × 2	65.02
8	8	3 × 2	65.27
8	8	5 × 2	66.07
8	8	7 × 2	**66.10**

Table 6. **Effect of the number of thought prompts.**

Vision	Language	**Caption Sampling Method**	
# Frames	# Thoughts	Beam Search	Similarity Ranking
2	7 × 2	57.28	60.00
4	7 × 2	63.83	64.21
8	7 × 2	66.10	**67.52**

Table 7. **Effect of caption ranking.**

Effect of caption ranking. Finally, we test the effect of similarity-based caption ranking we proposed for better caption decoding. Table 7 shows results from decoding with beam search vs. with caption ranking across various number of video frames. For beam search, we use a length of 6 tokens. Interestingly, we find that caption ranking consistently outperforms beam search, with a larger performance gap appearing when less frames are used. This shows that while a autoregressive decoding can suffer from semantic errors, leveraging the CLIP feature space and choosing the caption that best aligns with corresponding video frames is an effective strategy in narrowing down semantically compatible samples from the decoder.

5. Conclusion

We propose CaptionT5, a video captioning model that fine-tunes a pre-trained language model T5 to understand video input and generate its corresponding caption. As T5 is not previously trained on visual inputs, CaptionT5 uses the multimodal feature space of CLIP to provide video-and-thought prompts to align T5 with visual cues in video frames. We find that using features in CLIP to rank generated captions during decoding leads to additional boost in caption quality. Experiments on VATEX, MSRVTT, and YC2 datasets show that CaptionT5 outperforms models that rely on additional annotations such as subtitles as well as models pre-trained on large-scale video data.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 3

[2] 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. 5

[3] 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. 2

[4] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2

[5] Eldan Cohen and Christopher Beck. Empirical analysis of beam search performance degradation in neural sequence models. In *International Conference on Machine Learning*, pages 1290–1299. PMLR, 2019. 5

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 5

[8] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 3

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[10] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 2

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 3

[12] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 4

[13] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 3

[14] 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, 2020. 3, 5

[15] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. In *Advances in neural information processing systems*, 2021. 2, 5, 6, 11

[16] Sheng Li, Zhiqiang Tao, Kang Li, and Yun Fu. Visual to text: Survey of image and video captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(4):297–312, 2019. 1

[17] 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, 3

[18] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 2, 3, 5, 6

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. In *International Conference on Learning Representations*, 2020. 5

[20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 2, 5

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*. 5, 11

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2, 3

[23] 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, 2002. 5

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5, 11

[25] 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. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 5, 11

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21 (140):1–67, 2020. 2, 3

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 4

[29] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2, 3

[30] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 2, 3

[31] 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

[32] 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. 5

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 1

[34] 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. 2, 5, 7, 11

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, 2022. 2

[36] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 5, 11

[37] 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. *arXiv preprint arXiv:2109.14084*, 2021. 3

[38] 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. 2, 5, 7, 11

[39] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1

[40] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022. 2

[41] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 5

[42] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*. 2

[43] 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*, 34:23634–23651, 2021. 3

[44] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020. 6

[45] Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Ying Deng, and Weiming Hu. Open-book video captioning with retrieve-copy-generate network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9837–9846, 2021. 2, 3, 5, 6

[46] 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. 2, 5, 11

[47] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 3