



# Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, Cordelia Schmid

## ► To cite this version:

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, et al.. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. CVPR 2023 - IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun 2023, Vancouver, Canada. hal-04039246

**HAL Id: hal-04039246**

**<https://inria.hal.science/hal-04039246>**

Submitted on 21 Mar 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

Antoine Yang<sup>†\*</sup> Arsha Nagrani<sup>§</sup> Paul Hongsuck Seo<sup>§</sup> Antoine Miech<sup>‡</sup>

Jordi Pont-Tuset<sup>§</sup> Ivan Laptev<sup>†</sup> Josef Sivic<sup>¶</sup> Cordelia Schmid<sup>§</sup>

<sup>§</sup>Google Research <sup>†</sup>Inria Paris and Département d’informatique de l’ENS, CNRS, PSL Research University

<sup>‡</sup>DeepMind <sup>¶</sup>Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University in Prague

<https://antoyang.github.io/vid2seq.html>

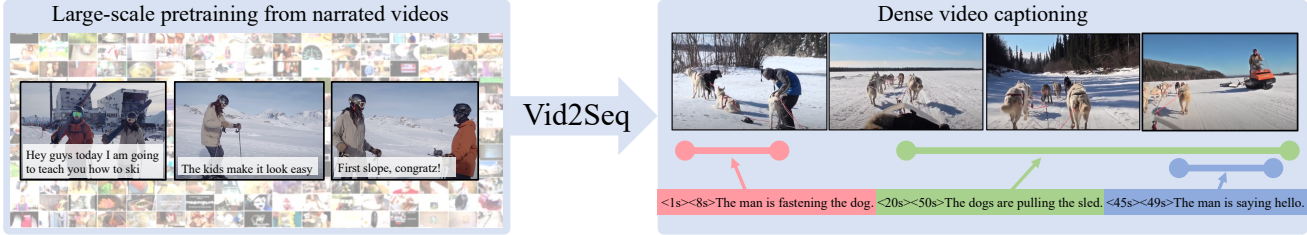


Figure 1. **Vid2Seq** is a visual language model that predicts dense event captions together with their temporal grounding in the video by generating a *single* sequence of tokens (right). This ability is enabled by large-scale pretraining on unlabeled narrated videos (left).

## Abstract

In this work, we introduce *Vid2Seq*, a multi-modal single-stage dense event captioning model pretrained on narrated videos which are readily-available at scale. The *Vid2Seq* architecture augments a language model with special time tokens, allowing it to seamlessly predict event boundaries and textual descriptions in the same output sequence. Such a unified model requires large-scale training data, which is not available in current annotated datasets. We show that it is possible to leverage unlabeled narrated videos for dense video captioning, by reformulating sentence boundaries of transcribed speech as pseudo event boundaries, and using the transcribed speech sentences as pseudo event captions. The resulting *Vid2Seq* model pretrained on the *YT-Temporal-1B* dataset improves the state of the art on a variety of dense video captioning benchmarks including *YouCook2*, *ViTT* and *ActivityNet Captions*. *Vid2Seq* also generalizes well to the tasks of video paragraph captioning and video clip captioning, and to few-shot settings. Our code is publicly available at [1].

## 1. Introduction

Dense video captioning requires the temporal localization and captioning of all events in an untrimmed video [46, 102, 131]. This differs from standard video captioning [63, 70, 80], where the goal is to produce a single caption for a given short video clip. Dense captioning is significantly more difficult, as it raises the additional complexity of localizing the events in minutes-long videos. However, it also benefits from long-range video information. This task is

potentially highly useful in applications such as large-scale video search and indexing, where the video content is not segmented into clips.

Existing methods mostly resort to two-stage approaches [37, 46, 100], where events are first localized and then captioned. To further enhance the inter-task interaction between event localization and captioning, some approaches have introduced models that jointly solve the two tasks [20, 102, 131]. However, often these approaches still require task-specific components such as event counters [102]. Furthermore, they exclusively train on manually annotated datasets of limited size [35, 46, 130], which makes it difficult to effectively solve the task. To address these issues, we take inspiration from recent sequence-to-sequence models pretrained on Web data which have been successful on a wide range of vision and language tasks [4, 11, 13, 105, 117].

First, we propose a video language model, called *Vid2Seq*. We start from a language model trained on Web text [78] and augment it with special *time tokens* that represent timestamps in the video. Given video frames and transcribed speech inputs, the resulting model jointly predicts all event captions and their corresponding temporal boundaries by generating a *single* sequence of discrete tokens, as illustrated in Figure 1 (right). Such a model therefore has the potential to learn multi-modal dependencies between the different events in the video via attention [94]. However this requires large-scale training data, which is not available in current dense video captioning datasets [35, 46, 130]. Moreover, collecting manual annotations of dense captions for videos is expensive and prohibitive at scale.

\*This work was done when the first author was an intern at Google.

Hence we propose to pretrain Vid2Seq by leveraging unlabeled narrated videos which are readily-available at scale. To do this, we reformulate sentence boundaries of transcribed speech as pseudo event boundaries, and use the transcribed speech sentences as pseudo event captions. We then pretrain Vid2Seq with a generative objective, that requires predicting the transcribed speech given visual inputs, and a denoising objective, which masks spans of transcribed speech. Note that transcribed speech may not describe the video content faithfully, and is often temporally misaligned with the visual stream [32, 43, 71]. For instance, from the example in Figure 1 (left), one can understand that the grey skier has descended a slope from the last speech sentence which is said *after* he actually descended the slope. Intuitively, Vid2Seq is particularly suited for learning from such noisy supervision as it jointly models *all* narrations and the corresponding timestamps in the video.

We demonstrate the effectiveness of our pretrained model through extensive experiments. We show the importance of pretraining on untrimmed narrated videos, the ability of Vid2Seq to use both the visual and speech modalities, the importance of the pretraining objectives, the benefit of joint caption generation and localization, as well as the importance of the language model size and the scale of the pretraining dataset. The pretrained Vid2Seq model achieves state-of-the-art performance on various dense video captioning benchmarks [35, 46, 130]. Our model also excels at generating paragraphs of text describing the video: without using ground-truth event proposals at inference time, our model outperforms all prior approaches including those that rely on such proposals [50, 76, 128]. Moreover, Vid2Seq generalizes well to the standard task of video clip captioning [9, 109]. Finally, we introduce a new few-shot dense video captioning setting in which we finetune our pretrained model on a small fraction of the downstream training dataset and show benefits of Vid2Seq in this setting.

In summary, we make the following contributions: (i) We introduce Vid2Seq for dense video captioning. Given multi-modal inputs (transcribed speech and video), Vid2Seq predicts a single sequence of discrete tokens that includes caption tokens interleaved with special *time tokens* that represent event timestamps. (ii) We show that transcribed speech and corresponding timestamps in unlabeled narrated videos can be effectively used as a source of weak supervision for dense video captioning. (iii) Finally, our pretrained Vid2Seq model improves the state of the art on three dense video captioning datasets (YouCook2, ViTT, ActivityNet Captions), two video paragraph captioning benchmarks (YouCook2, ActivityNet Captions) and two video clip captioning datasets (MSR-VTT, MSVD), and also generalizes well to few-shot settings.

Our code implemented in Jax and based on the Scenic library [19] is publicly released at [1].

## 2. Related Work

**Dense video captioning.** Dense video captioning lies at the intersection of event localization [25, 29, 33, 62, 65, 66, 84, 127] and event captioning [30, 63, 75, 97, 104]. The majority of existing methods for dense video captioning [37, 38, 46, 100, 103] consist of a temporal localization stage followed by an event captioning stage. To enrich inter-task interactions, recent works [8, 10, 20, 61, 73, 79, 82, 83, 100, 102, 131] jointly train the captioning and localization modules. In particular, Wang *et al.* [102] propose to view dense video captioning as a set prediction task, and jointly perform event localization and captioning for each event in parallel. In contrast, our model generates event boundaries and captions conditioned on the previously generated events. Deng *et al.* [20] propose to first generate a paragraph and then ground each sentence in the video. We also generate all captions as a single output sequence, however our output already includes event timestamps. Zhang *et al.* [125] propose to generate event boundaries sequentially, but separately perform event localization and single event captioning, and only use visual input. Most related to our work, Zhu *et al.* [133] also perform dense video captioning by generating a single output sequence. Their method, however, infers event locations directly from the timestamps of transcribed speech and, hence, can only detect events that closely follow the speech. In contrast, our model generates event timestamps as special tokens and can produce dense captions for videos with limited speech, as we demonstrate on the ActivityNet Captions dataset.

**Video and language pretraining.** Following the success of image-text pretraining [14, 21, 23, 24, 28, 34, 36, 39–41, 53–55, 58, 60, 68, 69, 85, 88, 92, 93, 99, 118–120, 124, 129], recent works have explored video-text pretraining [3–5, 26, 31, 32, 43, 49, 52, 56, 71, 72, 74, 80, 81, 89, 96, 98, 108, 111, 111–115, 121, 122]. These methods show strong improvements on various tasks such as text-video retrieval [5, 71], video question answering [112, 122] and video clip captioning [4, 80]. While these works mostly learn global video representations to tackle video-level prediction tasks, we here focus on learning detailed representations to address a dense prediction task requiring reasoning over multiple events in untrimmed videos. Several works have explored long-form video-text pretraining [90] and video-text pretraining for temporal localization tasks [7, 48, 64, 106, 110, 116]. However these works focus on video understanding tasks while our pretraining approach is tailored for a generative task that not only requires the model to reason over multiple events in the video, but also to describe them by natural language.

A few works explore pretraining for dense video captioning. Zhang *et al.* [125] pretrain on ActivityNet Captions to improve the downstream performance on the same dataset. In contrast, we propose a pretraining method that does not rely on *any* manual annotation, and show its ben-

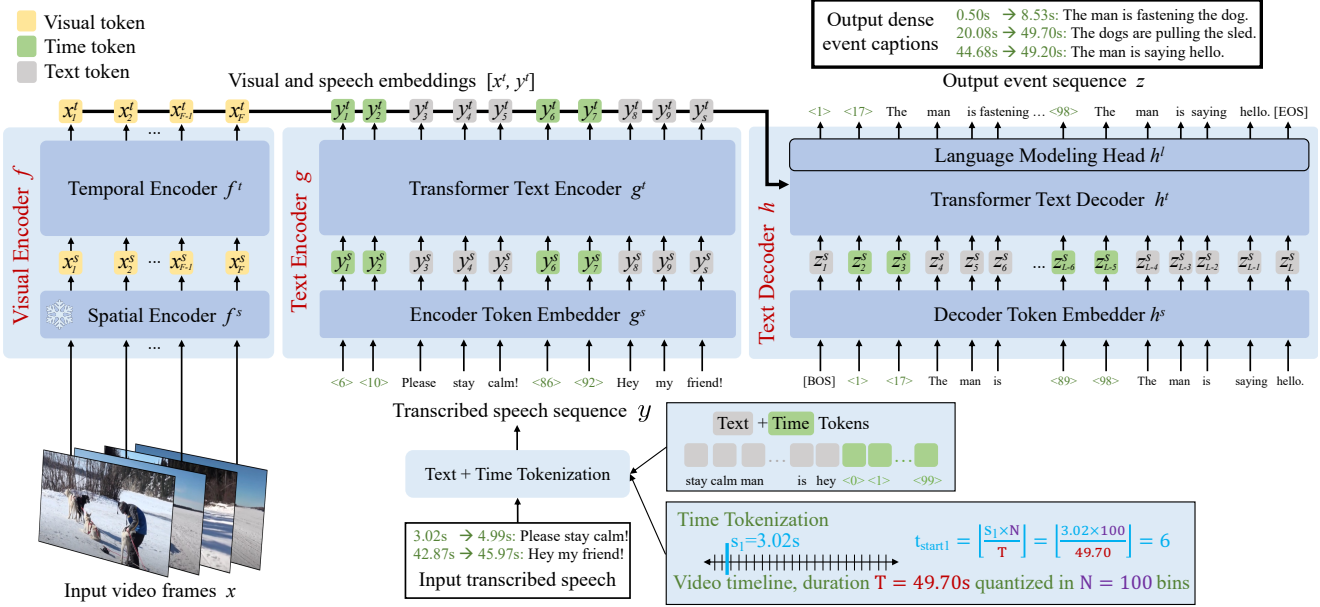


Figure 2. **Vid2Seq model overview.** We formulate dense event captioning as a sequence-to-sequence problem, using special *time tokens* to allow the model to seamlessly understand and generate sequences of tokens containing both textual semantic information and temporal localization information grounding each text sentence in the video. In detail, all input video frames  $x$  and the transcribed speech sequence  $y$  are first processed with a Visual Encoder  $f$  (a frozen Spatial Encoder  $f^s$  followed by a Temporal Encoder  $f^t$ ) and a Text Encoder  $g$  (a Token Embedder  $g^s$  followed by a Transformer Encoder  $g^t$ ), respectively. Then the Text Decoder  $h$  (composed of a Token Embedder  $h^s$ , a Transformer Encoder  $h^t$  and a Language Modeling Head  $h^l$ ) autoregressively generates the output event sequence  $z$  by cross-attending to the visual and speech embeddings  $x^t$  and  $y^t$ .

efits on multiple downstream datasets. Huang *et al.* [35] explore pretraining on narrated instructional videos, but only consider event captioning using ground truth proposals as their model does not handle localization. Finally, [35, 133] explore pretraining on a domain specific text-only dataset [45]. In contrast, we propose to pretrain on a generic video corpus [121] and show benefits on various domains.

**Unifying tasks as language modeling.** Recent works [11–13, 15, 17, 44, 59, 101, 117, 132] have shown that it is possible to cast various computer vision problems as a language modeling task, addressing object detection [11], grounded image captioning [117] or visual grounding [132]. In this work we also cast visual localization as a language modeling task. However, unlike prior work focused on image-level spatial localization, we address the different problem of event localization *in time*, in untrimmed videos.

### 3. Method

The goal of dense video captioning is to temporally localize and describe with natural language *all* events in an untrimmed input video. Therefore a key challenge is to effectively model the relationships between the different events in the video, as for example, it is easier to predict that the dogs are pulling the sled if we know that the man has just fastened a dog (see Figure 1 (right)). Furthermore, due to the dense nature of the task, there can be many events

in a long video and the requirement is to output a natural language caption for each event. Hence, another key challenge is that the manual collection of annotations for this task is particularly expensive. To tackle these challenges, we first develop a unified multi-modal model that jointly predicts event boundaries and captions as a single sequence of tokens, as explained in Section 3.1 and Figure 2. Second, we design a pretraining strategy that effectively leverages cross-modal supervision in the form of transcribed speech from unlabeled narrated videos by reformulating sentence boundaries as pseudo event boundaries, as presented in Section 3.2 and Figure 3.

#### 3.1. Model

We wish to design a model for dense video captioning that can capture relationships between events using visual and (transcribed) speech cues in order to effectively localize and describe these events in untrimmed minutes-long videos. To tackle this challenge, we cast dense video captioning as a sequence-to-sequence problem where the input and output sequences contain both the semantic information about the event in the form of natural language descriptions and the temporal localization of the events in the form of temporal timestamps. In addition, to best leverage both the visual and the language signal, we develop an appropriate multi-modal encoder-decoder architecture. As illustrated



in Figure 2, our architecture takes as input video frames  $x = \{x_i\}_{i=1}^F$  together with the transcribed speech sequence  $y = \{y_j\}_{j=1}^S$ . The output of our model is an event sequence  $z = \{z_k\}_{k=1}^L$ , where each event contains both its textual description and timestamps corresponding to the temporal event locations in the video. Below we explain the structure of the transcribed speech and event sequences constructed for our model as well as details of our model architecture.

**Sequence construction.** To model inter-event relationships in dense event captioning annotations (or the readily-available transcribed narration, see Section 3.2), we cast dense video captioning as predicting a single output sequence of tokens  $z$ . This output event sequence is constructed by leveraging a text tokenizer augmented with special *time tokens*. Furthermore, we enable our architecture to jointly reason about the semantic and temporal information provided in the transcript of the input narration by constructing the input transcript sequence  $y$  in a similar manner as the event sequence  $z$ . Details are given next.

**Time tokenization.** We start from a text tokenizer with a vocabulary size  $V$ , and augment it with  $N$  additional time tokens, resulting in a tokenizer with  $V + N$  tokens. The time tokens represent relative timestamps in a video, as we quantize a video of duration  $T$  into  $N$  equally-spaced timestamps. In detail, we use the SentencePiece tokenizer [47] with vocabulary size  $V = 32, 128$  and  $N = 100$ .

**Event sequence.** Our introduced tokenizer enables us to construct sequences that contain both video timestamps and text video descriptions. We next explain how we construct the output event sequence  $z$ . Note that videos have a variable number of events in standard dense video captioning datasets [35, 46, 130]. Each event  $k$  is characterized by a text segment, a start time and an end time. We first construct for each event  $k$  a sequence by concatenating its start time token  $t_{start_k}$ , its end time token  $t_{end_k}$  and its text tokens  $[z_{k_1}, \dots, z_{k_{l_k}}]$ . Then we order all these sequences in increasing order of their start times and concatenate them. In practice, each text segment ends with a dot symbol indicating the separation between different events. Finally, the event sequence is obtained by prepending and appending a BOS and an EOS tokens to indicate the start and the end of sequence, respectively, *i.e.*  $z = [BOS, t_{start_1}, t_{end_1}, z_{1_1}, \dots, z_{1_{l_1}}, t_{start_2}, \dots, EOS]$ .

**Transcribed speech sequence.** To enable the model to use both the transcribed speech and its corresponding timestamps, we convert the speech transcript into a speech sequence  $y$  similarly as the input training dense event captions  $z$ . This is done by segmenting the raw speech transcript into sentences with the Google Cloud API<sup>1</sup>, and using each transcribed speech sentence with its corresponding timestamps analogously as an event in the previously explained process.

<sup>1</sup><https://cloud.google.com/speech-to-text/docs/automatic-punctuation>.

**Architecture.** We wish to design an architecture that can effectively model relationships between different events in untrimmed minutes-long videos. To tackle this challenge, we propose a multi-modal encoder-decoder architecture, illustrated in Figure 2, that gradually refines and outputs the event sequence described above. In detail, given an untrimmed minutes-long video, the visual encoder  $f$  embeds its frames while the text encoder  $g$  embeds transcribed speech and the corresponding timestamps. Then a text decoder  $h$  predicts event boundaries and text captions using the visual and transcribed speech embeddings. The individual modules are described next.

**Visual encoder.** The visual encoder operates on a sequence of  $F$  frames  $x \in \mathbb{R}^{F \times H \times W \times C}$  where  $H$ ,  $W$  and  $C$  are the height, width and the number of channels of each frame. A visual backbone  $f^s$  first encodes each frame separately and outputs frame embeddings  $x^s = f^s(x) \in \mathbb{R}^{F \times d}$ , where  $d$  is the embedding dimension. Then a transformer encoder [94]  $f^t$  models temporal interactions between the different frames, and outputs  $F$  contextualized visual embeddings  $x^t = f^t(x^s + x^p) \in \mathbb{R}^{F \times d}$ , where  $x^p \in \mathbb{R}^{F \times d}$  are learnt temporal positional embeddings, which communicate time information from visual inputs to the model. In detail, the visual backbone is CLIP ViT-L/14 [22, 77] at resolution  $224 \times 224$  pixels, pretrained to map images to text descriptions with a contrastive loss on Web-scraped image-text pairs. We keep the backbone frozen for efficiency.

**Text encoder.** The text encoder operates on a transcribed speech sequence of  $S$  tokens  $y \in \{1, \dots, V + N\}^S$ , where  $V$  is the text vocabulary size,  $N$  is the size of the vocabulary of time tokens and  $S$  is the number of tokens in the transcribed speech sequence. Note that the transcribed speech sequence includes time tokens to input the temporal information from the transcribed speech into the model. An embedding layer  $g^s \in \mathbb{R}^{(V+N) \times d}$  embeds each token independently and outputs semantic embeddings  $y^s = g^s(y) \in \mathbb{R}^{S \times d}$ . Then a transformer encoder  $g^t$  computes interactions in the transcribed speech sequence and outputs  $S$  contextualized speech embeddings  $y^t = g^t(y^s) \in \mathbb{R}^{S \times d}$ .

**Text decoder.** The text decoder generates the event sequence  $z$  by using the encoder embeddings, which are obtained by concatenating the visual and speech embeddings  $x^t$  and  $y^t$ . The text decoder is based on a causal transformer decoder  $h^t$  that cross-attends to the encoder outputs, and at each autoregressive step  $k$ , self-attends to the previously generated tokens  $\hat{z}_{<k}^t$  to output a contextualized representation  $z_k^t = h^t(h^s(\hat{z}_{<k}^t), x^t, y^t) \in \mathbb{R}^d$  where  $h^s \in \mathbb{R}^{(V+N) \times d}$  is the decoder token embedding layer. Then a language modeling head  $h^l \in \mathbb{R}^{d \times (V+N)}$  predicts a probability distribution over the joint vocabulary of text and time tokens in order to predict the next token in the event sequence, *i.e.*  $z_k^l = h^l(z_k^t) \in \mathbb{R}^{V+N}$ .

**Text initialization.** We initialize the text encoder and the

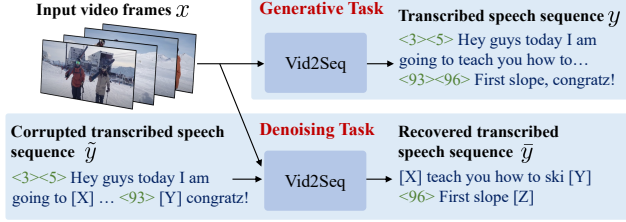


Figure 3. **Pretraining tasks.** To train Vid2Seq on unlabeled narrated videos, we design two pretraining objectives. **Top:** generative objective, given visual inputs  $x$  only, the task is to generate the transcribed speech sequence  $y$ . **Bottom:** denoising objective, given visual inputs  $x$  and the corrupted speech sequence  $\tilde{y}$ , the task is to generate the sequence of recovered speech segments  $\hat{\tilde{y}}$ .

text decoder with T5-Base [78] which has been pretrained on Web text corpora with a denoising loss. Therefore their implementation and parameters also closely follow T5-Base, e.g. they use relative positional embeddings and share their token embedding layer  $g^s = h^s \in \mathbb{R}^{(V+N) \times d}$ .

### 3.2. Training

In this Section, we describe how we leverage a large amount of unlabeled narrated videos to train the previously described dense event captioning model. We first present the pretraining method used to effectively train Vid2Seq using cross-modal supervision in readily-available narrated videos in Section 3.2.1 and Figure 3. Then we explain how we finetune our architecture for various downstream tasks including dense event captioning in Section 3.2.2.

#### 3.2.1 Pretraining on untrimmed narrated videos

We wish to leverage narrated videos for pretraining as they are easily available at scale [72, 121]. However these videos do not contain dense event captioning annotations. Therefore we use as supervisory signal the transcribed speech sentences and their corresponding timestamps. As speech transcripts are not always visually grounded and often temporally misaligned [32, 43, 71], we note that they only provide *weak* supervision. Furthermore, speech transcripts drastically differ from dense event captioning annotations. For instance, in the YT-Temporal-1B dataset [121], a video contains 120 speech sentences on average which is an order of magnitude more than the number of events in standard dense video captioning datasets [35, 46, 130]. Our Vid2Seq model is particularly suitable for using such weak supervision as it constructs the speech sequence similarly as a manually annotated event sequence, and jointly contextualizes the speech boundaries and semantic information on the level of potentially minutes-long videos (see Section 3.1) rather than at a shorter clip-level, enabling our model to learn long-term relationships between the different speech segments: in experiments we show that pretraining on entire minutes-long videos is highly beneficial.

We next describe the two proposed training objectives,

which are both based on a maximum likelihood objective. Formally, given visual inputs  $x$ , encoder text sequence  $y$  and a decoder target text sequence  $z$ , both objectives are based on minimizing the following loss:

$$\mathcal{L}_\theta(x, y, z) = -\frac{1}{\sum_{k=1}^{L-1} w_k} \sum_{k=1}^{L-1} w_k \log p_\theta(z_{k+1} | x, y, z_{1:k}), \quad (1)$$

where  $L$  is the length of the decoder target sequence,  $w_k$  is the weight for  $k$ -th token in the sequence, which we set to  $w_k = 1 \forall k$  in practice,  $\theta$  denotes the trainable parameters in the model and  $p_\theta$  is the output probability distribution over the vocabulary of text and time tokens.

**Generative objective.** This objective uses the transcribed speech as a (pseudo-)supervisory signal to teach the decoder to predict a sequence of events given visual inputs. Given video frames  $x$ , which are fed to the encoder, the decoder has to predict the transcribed speech sequence  $y$  (see Figure 3), which serves as a proxy dense event captioning annotation. Note that no text input is given to the encoder for this task as using transcribed speech both as input and target would lead the model to learn text-only shortcuts.

**Denoising objective.** As no text input is given to the encoder for the generative proxy task, the generative objective only trains the visual encoder and the text decoder, but not the text encoder. However when our model is used for dense video captioning, the text encoder has a significant importance as it encodes speech transcripts. Hence we introduce a denoising objective that aims at jointly aligning the visual encoder, the text encoder and the text decoder. Inspired by T5 [78] in the text domain, we randomly mask spans of (text and time) tokens in the transcribed speech sequence with a probability  $P$  and an average span length  $M$ . The encoder input is composed of the video frames  $x$  together with the corrupted speech sequence  $\tilde{y}$ , which contains sentinel tokens that uniquely identify the masked spans. The decoder then has to predict a sequence  $\hat{\tilde{y}}$  constructed with the corresponding masked spans for each sentinel token, based on visual inputs  $x$  and speech context  $\tilde{y}$  (see Figure 3).

#### 3.2.2 Downstream task adaptation

Our architecture and task formulation enables us to tackle dense video captioning with a generic language modeling training objective and inference procedure. Note that as a by-product of our generic architecture, our model can also be used to generate paragraphs about entire videos by simply removing the time tokens from the output sequence, and can also be easily adapted to video clip captioning with the same finetuning and inference recipe.

**Finetuning.** To finetune our model for dense video captioning, we use a maximum likelihood objective based on the event sequence (see Equation 1). Given video frames  $x$  and speech transcripts  $y$ , the decoder has to predict the event sequence  $z$ .

**Inference.** The text decoder autoregressively generates the event sequence by sampling from the model likelihood. In practice, we use beam search as we find that it improves the captioning quality compared with argmax sampling or nucleus sampling. Finally, the event sequence is converted into a set of event predictions by simply reversing the sequence construction process.

## 4. Experiments

This section demonstrates the effectiveness of our pre-trained Vid2Seq model and compares our method to the state of the art. We first outline our experimental setup in Section 4.1. We then present ablation studies in Section 4.2. The comparison to the state of the art in dense video captioning, video paragraph captioning and video clip captioning is presented in Section 4.3. Next, we present results in a new few-shot dense video captioning setting in Section 4.4. Finally, we show qualitative results in Section 4.5.

### 4.1. Experimental setup

**Datasets.** For pretraining, following prior work showing the benefits of pretraining on a diverse and large dataset [122], we use the **YT-Temporal-1B** dataset [121], which includes 18 million narrated videos collected from YouTube. We evaluate Vid2Seq on three downstream dense video captioning datasets: YouCook2 [130], ViTT [35] and ActivityNet Captions [46]. **YouCook2** has 2K untrimmed videos of cooking procedures. On average, each video lasts 320s and is annotated with 7.7 temporally-localized sentences. **ViTT** consists of 8K untrimmed instructional videos. On average, each video lasts 250s and is annotated with 7.1 temporally-localized short tags. **ActivityNet Captions** contains 20k untrimmed videos of various human activities. On average, each video lasts 120s and is annotated with 3.7 temporally-localized sentences. For video clip captioning, we use two standard benchmarks, **MSR-VTT** [109] and **MSVD** [9]. For all datasets, we follow the standard splits for training, validation and testing. Note that we only use videos available on YouTube at the time of the work, resulting in 10 to 20% less videos than in the original datasets.

**Implementation details.** We extract video frames at 1FPS, and subsample or pad the sequence of frames to  $F$  frames where we set  $F = 100$ . The text encoder and decoder sequence are truncated or padded to  $L = S = 1000$  tokens. Our model has 314M trainable parameters. We use the Adam optimizer [42]. We pretrain our model for 200,000 iterations with a batch size of 512 videos split on 64 TPU v4 chips, which lasts a day. We sum both pretraining objectives with equal weighting to get our final pretraining loss. More details are included in Appendix Section B.

**Evaluation metrics.** For captioning, we use CIDEr [95] (C) and METEOR [6] (M). For dense video captioning, we follow the commonly used evaluation tool [46] which calculates matched pairs between generated events and the

Pretraining input		YouCook2			ActivityNet		
Untrimmed	Time tokens	S	C	F1	S	C	F1
1.	<i>No pretraining</i>	4.0	18.0	18.1	5.4	18.8	49.2
2.	✗	5.5	27.8	20.5	5.5	26.5	52.1
3.	✓	6.7	35.0	23.3	5.6	27.4	52.2
4.	✓	<b>7.9</b>	<b>47.1</b>	<b>27.3</b>	<b>5.8</b>	<b>30.1</b>	<b>52.4</b>

Table 1. **Ablation showing the impact of using untrimmed videos and adding time tokens during pretraining.** When we use untrimmed video-speech inputs, time information from transcribed speech sentence boundaries is integrated via time tokens.

ground truth across IoU thresholds of  $\{0.3, 0.5, 0.7, 0.9\}$ , and compute captioning metrics over the matched pairs. However, these metrics do not take into account the story of the video. Therefore we also use SODA.c [27] (S) for an overall dense video captioning evaluation. To further isolate the evaluation of event localization, we report the average precision and average recall across IoU thresholds of  $\{0.3, 0.5, 0.7, 0.9\}$  and their harmonic mean, the F1 Score.

### 4.2. Ablation studies

The default Vid2Seq model predicts both text and time tokens, uses both visual frames and transcribed speech as input, builds on the T5-Base language model, and is pre-trained on untrimmed videos from YT-Temporal-1B with both the generative and denoising losses. Below we ablate the importance of each of these factors on the downstream dense video captioning performance by reporting results on YouCook2 and ActivityNet Captions validation sets.

**Pretraining on untrimmed narrated videos by exploiting transcribed speech sentence boundaries.** In Table 1, we evaluate the effectiveness of our pretraining task formulation that uses untrimmed videos and integrates sentence boundaries of transcribed speech via time tokens. In contrast, most video clip captioning pretraining methods [35, 70, 80] use short, trimmed, video-speech segments for pretraining. We adapt this strategy in our model and find that it indeed yields significant performance improvements over the baseline that uses no video-text pretraining (row 2 vs row 1). However, larger improvements are obtained by using untrimmed video-speech inputs (row 3 vs row 2). Moreover, using time tokens to integrate time information from transcribed speech drastically improves performance (row 4 vs row 3). This shows the benefits of exploiting sentence boundaries of transcribed speech via time tokens and of using untrimmed videos during pretraining. In Appendix Section C.2, we show additional ablations to quantify how the performance improves by pretraining on longer narrated videos that contain more speech sentences.

**Input modalities and pretraining objectives.** In Table 2, we analyze the importance of input modalities and pretraining tasks on the downstream dense video captioning performance. The model with visual inputs only (no transcribed speech as input) benefits significantly from pretraining with the generative objective (row 3 vs row 1). This shows the

	Finetuning Input		Pretraining loss	YouCook2			ActivityNet		
	Visual	Speech	Generative Denoising	S	C	F1	S	C	F1
1.	✓	✗	No pretraining	3.0	15.6	15.4	5.4	14.2	46.5
2.	✓	✓	No pretraining	4.0	18.0	18.1	5.4	18.8	49.2
3.	✓	✓	✓	5.7	25.3	23.5	<b>5.9</b>	<b>30.2</b>	51.8
4.	✓	✓	✓	2.5	10.3	15.9	4.8	17.0	48.8
5.	✓	✓	✓	<b>7.9</b>	<b>47.1</b>	<b>27.3</b>	5.8	30.1	<b>52.4</b>

Table 2. Effect of input modalities and pretraining losses.

	Captioning	Pretraining	YouCook2			ActivityNet		
			Recall	Precision	F1	Recall	Precision	F1
1.	✗	✗	17.8	19.4	17.7	47.3	57.9	52.0
2.	✓	✗	17.2	20.6	18.1	42.5	<b>64.1</b>	49.2
3.	✗	✓	25.7	21.4	22.8	52.5	53.0	51.1
4.	✓	✓	<b>27.9</b>	<b>27.8</b>	<b>27.3</b>	<b>52.7</b>	53.9	<b>52.4</b>

Table 3. Effect of joint captioning and localization on the localization performance. The variant that does not caption corresponds to a localization-only variant that only predicts time tokens.

	Language Model	Pretraining		YouCook2			ActivityNet		
		# Videos	Dataset	S	C	F1	S	C	F1
1.	T5-Small	15M	YTT	6.1	31.1	24.3	5.5	26.5	52.2
2.	T5-Base	∅	∅	4.0	18.0	18.1	5.4	18.8	49.2
3.	T5-Base	15K	YTT	6.3	35.0	24.4	5.1	24.4	49.9
4.	T5-Base	150K	YTT	7.3	40.1	26.7	5.4	27.2	51.3
5.	T5-Base	1M5	YTT	7.8	45.5	26.8	5.6	28.7	52.2
6.	T5-Base	1M	HTM	<b>8.3</b>	<b>48.3</b>	26.6	<b>5.8</b>	28.8	<b>53.1</b>
7.	T5-Base	15M	YTT	7.9	47.1	<b>27.3</b>	<b>5.8</b>	<b>30.1</b>	52.4

Table 4. Effect of language model size and pretraining data. HTM: HowTo100M [72], YTT: YT-Temporal-1B [121].

effectiveness of using the transcribed speech as a proxy annotation for dense video captioning pretraining. However, this model is pretrained with visual inputs only and its performance largely drops when it is finetuned with both visual and transcribed speech inputs (row 4 vs row 3). With both modalities, adding the denoising loss strongly benefits our model (row 5 vs rows 4 and 2). We conclude that the denoising objective benefits multi-modal reasoning.

**Effect of captioning on localization.** In Table 3, we compare the event localization performance of our model with a localization-only variant that only predicts event boundaries. We find that the model that jointly predicts event boundaries and captions localizes better and benefits more from pretraining than the localization-only baseline (row 4 vs row 3), which demonstrates the importance of contextualizing the noisy timestamps of the transcribed speech with the speech semantic content during pretraining.

**Model size and pretraining data.** In Table 4, we show that the language model size has a great importance on the performance, as the model with T5-Base outperforms its variant with T5-Small (row 7 vs row 1). We also evaluate the importance of the size of the pretraining dataset of narrated videos by constructing subsets such that larger subsets include the smaller ones. We find that scaling up the size of the pretraining dataset is beneficial, and that our pretraining method yields important benefits when only using 150K narrated videos for pretraining (row 4). We

Method	Backbone	YouCook2 (val)			ViTT (test)			ActivityNet (val)		
		S	C	M	S	C	M	S	C	M
MT [131]	TSN	—	6.1	3.2	—	—	—	—	9.3	5.0
ECHR [103]	C3D	—	—	3.8	—	—	—	3.2	14.7	7.2
PDVC [102]	TSN	4.4	22.7	4.7	—	—	—	5.4	29.0	8.0
PDVC [102] <sup>†</sup>	CLIP	4.9	28.9	5.7	—	—	—	<b>6.0</b>	29.3	7.6
UEDVC [125]	TSN	—	—	—	—	—	—	5.5	—	—
E2ESG [133]	C3D	—	25.0*	3.5	—	25.0	8.1	—	—	—
Vid2Seq (Ours)	CLIP	<b>7.9</b>	<b>47.1</b>	<b>9.3</b>	<b>13.5</b>	<b>43.5</b>	<b>8.5</b>	5.8	<b>30.1</b>	<b>8.5</b>

Table 5. Comparison to the state of the art for dense video captioning. \* Results provided by the authors. <sup>†</sup> Results of our experiments using the official codebase.

Method	Backbone	YouCook2 (val)		ViTT (test)		ActivityNet (val)	
		Recall	Precision	Recall	Precision	Recall	Precision
PDVC [102]	TSN	—	—	—	—	55.4	58.1
PDVC [102] <sup>†</sup>	CLIP	—	—	—	—	53.2	54.7
UEDVC [125]	TSN	—	—	—	—	<b>59.0</b>	<b>60.3</b>
E2ESG [133]	C3D	20.7*	20.6*	32.2*	32.1*	—	—
Vid2Seq (Ours)	CLIP	<b>27.9</b>	<b>27.8</b>	<b>42.6</b>	<b>46.2</b>	52.7	53.9

Table 6. Comparison to the state of the art for event localization. \* Results provided by the authors. <sup>†</sup> Results of our experiments using the official codebase.

further show that our pretraining method generalizes well to the HowTo100M dataset [72]. The model pretrained on HowTo100M (row 6) actually achieves best results on YouCook2, as these datasets are from a similar domain. Finally, we ablate the importance of the size and pretraining of the visual backbone in Appendix Section C.2.

### 4.3. Comparison to the state of the art

**Dense video captioning.** In Table 5, we compare our approach to state-of-the-art dense video captioning methods using cross-entropy training<sup>1</sup> on the YouCook2, ViTT and ActivityNet Captions datasets. Vid2Seq sets new state of the art on all three datasets. In particular, our method improves the CIDEr metric by 18.2 and 0.8 points on YouCook2 and ActivityNet Captions over PDVC. Our method also outperforms E2ESG [133] which uses in-domain text-only pretraining on Wikihow. These results demonstrate the strong dense event captioning ability of our pretrained Vid2Seq model.

**Event localization.** In Table 6, we evaluate the event localization performance of our dense video captioning model in comparison with prior work. On both YouCook2 and ViTT, Vid2Seq outperforms prior work [133] tackling dense video captioning as a single sequence generation task. However, our model underperforms compared to PDVC [102] and UEDVC [102] on ActivityNet Captions. We emphasize that our approach integrates less prior knowledge about temporal localization than both these approaches, which include task specific components such as event counters [102] or separately train a model for the localization subtask [125].

**Video paragraph captioning.** In Table 7, we compare our approach to state-of-the-art video paragraph captioning methods on the YouCook2 and ActivityNet Cap-

<sup>1</sup>We do not include methods directly optimizing the test metric [20, 73].



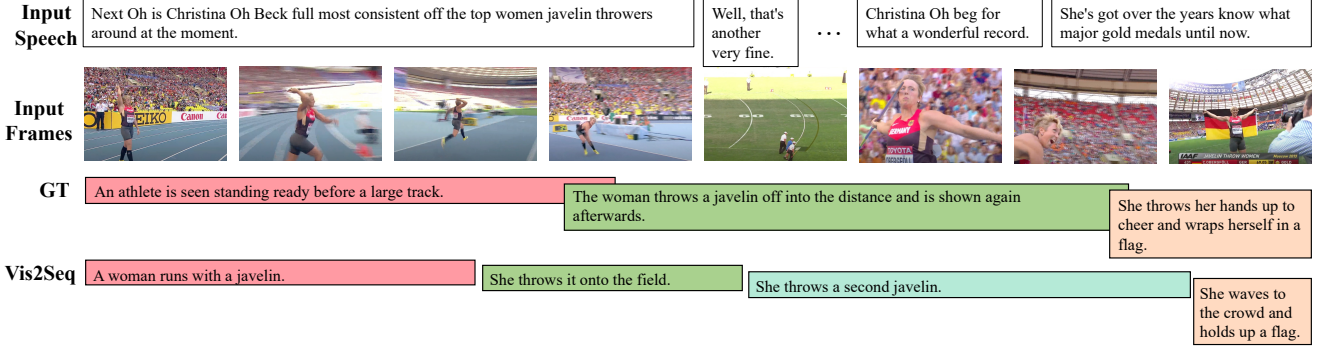


Figure 4. Example of dense event captioning predictions of Vid2Seq on ActivityNet Captions validation set, compared with ground-truth.

Method	Backbone	YouCook2 (val)		ActivityNet (val-ae)	
		C	M	C	M
<i>With GT Proposals</i>					
VTransformer [131]	V (ResNet-200) + F	32.3	15.7	22.2	15.6
Transformer-XL [18]	V (ResNet-200) + F	26.4	14.8	21.7	15.1
MART [50]	V (ResNet-200) + F	35.7	15.9	23.4	15.7
GVDSup [128]	V (ResNet-101) + F + O	—	—	22.9	16.4
AdvInf [76]	V (ResNet-101) + F + O	—	—	21.0	16.6
PDVC [102]	V + F (TSN)	—	—	27.3	15.9
<i>With Learnt Proposals</i>					
MFT [107]	V + F (TSN)	—	—	19.1	14.7
PDVC [102]	V + F (TSN)	—	—	20.5	15.8
PDVC [102] <sup>†</sup>	V (CLIP)	—	—	23.6	15.9
Vid2Seq (Ours)	V (CLIP)	<b>50.1</b>	<b>24.0</b>	<b>28.0</b>	<b>17.0</b>

Table 7. Comparison to the SoTA for video paragraph captioning.

<sup>†</sup> Results of our experiments using the official codebase. V/F/O refers to visual/flow/object features.

Method	Trained Parameters	Pretraining Data	MSR-VTT (test)		MSVD (test)	
			C	M	C	M
ORG-TL [126]	—	∅	50.9	28.8	95.2	36.4
SwinBERT [63]	229M	∅	53.8	29.9	120.6	41.3
Vid2Seq (Ours)	314M	∅	57.2	30.0	120.3	41.4
MV-GPT [80]	354M	HowTo100M	60.0	29.9*	—	—
Vid2Seq (Ours)	314M	HowTo100M	61.5	30.4	140.6	44.5
Vid2Seq (Ours)	314M	YT-Temporal-1B	<b>64.6</b>	<b>30.8</b>	<b>146.2</b>	<b>45.3</b>

Table 8. Comparison to the SoTA for video clip captioning. \* indicates results re-evaluated by the same evaluation toolkit.

Data		YouCook2			ViTT			ActivityNet		
		S	C	M	S	C	M	S	C	M
1.	1%	2.4	10.1	3.3	2.0	7.4	1.9	2.2	6.2	3.2
2.	10%	3.8	18.4	5.2	10.7	28.6	6.0	4.3	20.0	6.1
3.	50%	6.2	32.1	7.6	12.5	38.8	7.8	5.4	27.5	7.8
4.	100%	<b>7.9</b>	<b>47.1</b>	<b>9.3</b>	<b>13.5</b>	<b>43.5</b>	<b>8.5</b>	<b>5.8</b>	<b>30.1</b>	<b>8.5</b>

Table 9. **Few-shot dense event captioning**, by finetuning Vid2Seq using a small fraction of the downstream training dataset.

tions datasets. Vid2Seq outperforms all prior methods on both datasets, including the ones using ground-truth event boundary proposals at inference time [18, 50, 76, 102, 128, 131], showing strong video paragraph captioning ability.

**Video clip captioning.** In Table 8, we compare our approach to state-of-the-art video clip captioning methods on the MSR-VTT and MSVD datasets. Vid2Seq improves over prior methods in their respective pretraining data setting while using a comparable number of trained parameters. We conclude that our pretrained Vid2Seq model generalizes well to the standard video clip captioning setting.

## 4.4. Few-shot dense video captioning

To further evaluate the generalization capabilities of our pretrained Vid2Seq model, we propose a new few-shot dense video captioning setting where we finetune Vid2Seq using only a fraction of the downstream training dataset. From Table 9, we observe important improvements when using 10% compared to 1% of training data (row 2 vs 1). In Appendix Section C.1 we further show that pretraining is essential in this few-shot setting.

## 4.5. Qualitative examples

In Figure 4, we show an example of dense event captioning predictions from Vid2Seq. This example shows that our model can predict meaningful event boundaries and captions, and that the predicted captions and boundaries differ considerably from the transcribed speech input (showing the importance of the visual tokens in the input). More examples are provided in Appendix Section A.

## 5. Conclusion

We introduced Vid2Seq, a visual language model that performs dense video captioning by generating a single sequence of tokens including both text and time tokens given multi-modal inputs. We showed that Vid2Seq benefits from large-scale pretraining on unlabeled untrimmed narrated videos by leveraging transcribed speech sentences and corresponding temporal boundaries. Vid2Seq achieves state-of-the-art results on various dense event captioning datasets, as well as multiple video paragraph captioning and standard video clip captioning benchmarks. Finally, we believe the sequence-to-sequence design of Vid2Seq has the potential to be extended to a wide range of *other* video tasks such as temporally-grounded video question answering [51, 56, 57] or temporal action localization [16, 67, 123].

**Acknowledgements.** The work was partially funded by a Google gift, the French government under management of Agence Nationale de la Recherche as part of the "Investissements d’avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), the Louis Vuitton ENS Chair on Artificial Intelligence, the European Regional Development Fund under project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15 003/0000468). We thank Anurag Arnab, Minsu Cho, Anja Hauth, Ashish Thapliyal, Bo Pang, Bryan Seybold and the entire Ganesha team for helpful discussions.

## References

- [1] Vid2Seq project webpage. <https://antoyang.github.io/vid2seq.html>. 1, 2
- [2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 14
- [3] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*, 2021. 2
- [4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1, 2
- [5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 2
- [6] Satantjeev 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*, 2005. 6
- [7] Meng Cao, Tianyu Yang, Junwu Weng, Can Zhang, Jue Wang, and Yuexian Zou. Locvtp: Video-text pre-training for temporal localization. In *ECCV*, 2022. 2
- [8] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iPerceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. In *WACV*, 2021. 2
- [9] David L Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 2, 6, 14
- [10] Shaoxiang Chen and Yu-Gang Jiang. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In *CVPR*, 2021. 2
- [11] Ting Chen, Saurabh Saxena, Lala Li, David J Fleet, and Geoffrey Hinton. Pix2seq: A language modeling framework for object detection. In *ICLR*, 2022. 1, 3
- [12] Ting Chen, Saurabh Saxena, Lala Li, Tsung-Yi Lin, David J Fleet, and Geoffrey Hinton. A unified sequence interface for vision tasks. In *NeurIPS*, 2022. 3
- [13] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 1, 3
- [14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, 2020. 2
- [15] Zhiyang Chen, Yousong Zhu, Zhaowen Li, Fan Yang, Wei Li, Haixin Wang, Chaoyang Zhao, Liwei Wu, Rui Zhao, Jinqiao Wang, et al. Obj2seq: Formatting objects as sequences with class prompt for visual tasks. In *NeurIPS*, 2022. 3
- [16] Feng Cheng and Gedas Bertasius. TALLformer: Temporal action localization with long-memory transformer. In *ECCV*, 2022. 8
- [17] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 3
- [18] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *ACL*, 2019. 8
- [19] Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, and Yi Tay. Scenic: A jax library for computer vision research and beyond. In *CVPR*, 2022. 2
- [20] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video captioning. In *CVPR*, 2021. 1, 2, 7
- [21] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021. 2
- [22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 4
- [23] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. In *NeurIPS*, 2022. 2
- [24] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, 2022. 2
- [25] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *ECCV*, 2016. 2
- [26] 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. 2
- [27] Soichiro Fujita, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. SODA: Story oriented dense video captioning evaluation framework. In *ECCV*, 2020. 6
- [28] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020. 2
- [29] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *ICCV*, 2017. 2

- [30] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 2017. 2
- [31] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *CVPR*, 2022. 2
- [32] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *CVPR*, 2022. 2, 5
- [33] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *CVPR*, 2016. 2
- [34] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *CVPR*, 2022. 2
- [35] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. In *ACL-IJCNLP*, 2020. 1, 2, 3, 4, 5, 6, 14, 17
- [36] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, 2021. 2
- [37] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *BMVC*, 2020. 1, 2
- [38] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *CVPR Workshops*, 2020. 2
- [39] 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 *ICML*, 2021. 2
- [40] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. In *ICCV*, 2021. 2
- [41] Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021. 2
- [42] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6, 14
- [43] Dohwan Ko, Joonmyung Choi, Juyeon Ko, Shinyeong Noh, Kyoung-Woon On, Eun-Sol Kim, and Hyunwoo J Kim. Video-text representation learning via differentiable weak temporal alignment. In *CVPR*, 2022. 2, 5
- [44] Alexander Kolesnikov, André Susano Pinto, Lucas Beyer, Xiaohua Zhai, Jeremiah Harmsen, and Neil Houlsby. Uvim: A unified modeling approach for vision with learned guiding codes. In *NeurIPS*, 2022. 3
- [45] Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*, 2018. 3
- [46] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1, 2, 4, 5, 6, 14
- [47] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *ACL*, 2018. 4
- [48] Jie Lei, Tamara L Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *NeurIPS*, 2021. 2
- [49] 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 *CVPR*, 2021. 2
- [50] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara L Berg, and Mohit Bansal. MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *ACL*, 2020. 2, 8, 14
- [51] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. TVQA: Localized, compositional video question answering. In *EMNLP*, 2018. 8
- [52] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *CVPR*, 2022. 2
- [53] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 2
- [54] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2
- [55] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2
- [56] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*, 2020. 2, 8
- [57] 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 *NeurIPS Track on Datasets and Benchmarks*, 2021. 8
- [58] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 2
- [59] Wanhua Li, Zhexuan Cao, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Label2Label: A language modeling framework for multi-attribute learning. In *ECCV*, 2022. 3
- [60] 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 *ECCV*, 2020. 2
- [61] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, 2018. 2

- [62] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *AAAI*, 2020. 2
- [63] 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 *CVPR*, 2022. 1, 2, 8
- [64] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *NeurIPS*, 2022. 2
- [65] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. BMN: Boundary-matching network for temporal action proposal generation. In *ICCV*, 2019. 2
- [66] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *ECCV*, 2018. 2
- [67] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. In *IEEE Transactions on Image Processing*, 2022. 8
- [68] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2
- [69] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 2
- [70] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. UniViLM: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 1, 6, 17
- [71] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 2, 5
- [72] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 2, 5, 7, 14
- [73] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, 2019. 2, 7
- [74] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022. 2
- [75] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017. 2
- [76] Jae Sung Park, Marcus Rohrbach, Trevor Darrell, and Anna Rohrbach. Adversarial inference for multi-sentence video description. In *CVPR*, 2019. 2, 8
- [77] 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 *ICML*, 2021. 4, 17
- [78] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 1, 5
- [79] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *ICCV*, 2019. 2
- [80] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *CVPR*, 2022. 1, 2, 6, 8, 17
- [81] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *CVPR*, 2021. 2
- [82] Zhiqiang Shen, Jianguo Li, Zhou Su, Minjun Li, Yurong Chen, Yu-Gang Jiang, and Xiangyang Xue. Weakly supervised dense video captioning. In *CVPR*, 2017. 2
- [83] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *ACL*, 2019. 2
- [84] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016. 2
- [85] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *CVPR*, 2022. 2
- [86] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014. 15
- [87] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. In *TMLR*, 2022. 17
- [88] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*, 2019. 2
- [89] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *ICCV*, 2019. 2
- [90] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. In *NeurIPS*, 2022. 2
- [91] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 15
- [92] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 2
- [93] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *NeurIPS*, 2021. 2



- [94] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1, 4
- [95] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDER: Consensus-based image description evaluation. In *CVPR*, 2015. 6
- [96] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 2
- [97] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *CVPR*, 2018. 2
- [98] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *CVPR*, 2022. 2
- [99] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. UFO: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021. 2
- [100] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018. 1, 2
- [101] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 3
- [102] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *ICCV*, 2021. 1, 2, 7, 8, 14
- [103] Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. Event-centric hierarchical representation for dense video captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020. 2, 7
- [104] Xin Wang, Wenhui Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *CVPR*, 2018. 2
- [105] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *ICLR*, 2022. 1
- [106] Zixu Wang, Yujie Zhong, Yishu Miao, Lin Ma, and Lucia Specia. Contrastive video-language learning with fine-grained frame sampling. In *AACL-IJCNLP*, 2022. 2
- [107] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *ECCV*, 2018. 8
- [108] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, 2021. 2
- [109] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2, 6, 14
- [110] Mengmeng Xu, Erhan Gundogdu, Maksim Lapin, Bernard Ghanem, Michael Donoser, and Loris Bazzani. Contrastive language-action pre-training for temporal localization. *arXiv preprint arXiv:2204.12293*, 2022. 2
- [111] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022. 2
- [112] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 2
- [113] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Learning to answer visual questions from web videos. *IEEE TPAMI*, 2022. 2
- [114] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-temporal video grounding with transformers. In *CVPR*, 2022. 2
- [115] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022. 2
- [116] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *ICCV*, 2021. 2
- [117] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. Crossing the format boundary of text and boxes: Towards unified vision-language modeling. In *ECCV*, 2021. 1, 3
- [118] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graph. In *AAAI*, 2020. 2
- [119] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 2
- [120] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2
- [121] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanyang Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. MERLOT Reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 2022. 2, 3, 5, 6, 7, 14
- [122] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. In *NeurIPS*, 2021. 2, 6
- [123] Chenlin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *ECCV*, 2022. 8
- [124] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying localization and vision-language understanding. In *NeurIPS*, 2022. 2

- [125] Qi Zhang, Yuqing Song, and Qin Jin. Unifying event detection and captioning as sequence generation via pre-training. In *ECCV*, 2022. 2, 7
- [126] 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 *CVPR*, 2020. 8
- [127] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 2
- [128] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, 2019. 2, 8, 14
- [129] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI*, 2020. 2
- [130] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1, 2, 4, 5, 6, 14
- [131] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. 1, 2, 7, 8
- [132] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. SeqTR: A simple yet universal network for visual grounding. In *ECCV*, 2022. 3
- [133] Wanrong Zhu, Bo Pang, Ashish Thapliyal, William Yang Wang, and Radu Soricut. End-to-end dense video captioning as sequence generation. In *COLING*, 2022. 2, 3, 7

## Appendix

In this Appendix, we present the following additional material:

- (i) Additional qualitative examples of dense video captioning predictions (Section A).
- (ii) Additional information about our experimental setup (Section B);
- (iii) Additional experimental results (Section C), including an ablation on the importance of pretraining for few-shot dense video captioning (Section C.1) and additional ablation studies in the standard fully-supervised dense video captioning setting (Section C.2).

### A. Qualitative examples of dense video captioning predictions

In Figure 4, we show qualitative results of dense event captioning by our Vid2Seq model. Here in Figures 5 and 6, we show additional results on examples from the YouCook2 and ActivityNet Captions datasets. These results show that Vid2Seq can predict meaningful dense captions and event boundaries in diverse scenarios, with or without transcribed speech input, *e.g.* series of instructions in cooking recipes (Figure 5) or actions in human sports or leisure activities (first three examples in Figure 6). The last example in Figure 6 illustrates a failure case where the model hallucinates events that are not visually grounded such as ‘one man hats off to the camera’.

### B. Experimental setup

In this section, we complement the information provided in Section 4.1 about the datasets we use (Section B.1). We also give additional implementation details (Section B.2).

#### B.1. Datasets

**YT-Temporal-1B** [121] consists of 18.821M unlabeled narrated videos covering about 150 years of video content for pretraining. Compared with HowTo100M [72], this dataset was created to cover a wider range of domains and not only instructional videos.

**HowTo100M** [72] consists of 1.221M unlabeled narrated instructional videos covering about 15 years of video content for pretraining.

**YouCook2** [130] has 1,790 untrimmed videos of cooking procedures. On average, each video lasts 320s and is annotated with 7.7 temporally-localized imperative sentences. The dataset is split into 1,333 videos for training and 457 videos for validation.

**ViTT** [35] consists of 7,672 untrimmed instructional videos from the YouTube-8M dataset [2]. Compared to YouCook2,

ViTT was created to better reflect the distribution of instructional videos in the wild. On average, each video lasts 250s and is annotated with 7.1 temporally-localized short tags. The dataset is split into 5,476, 1,102 and 1,094 videos for training, validation and testing, respectively. Videos in the validation and test sets are provided with multiple sets of dense event captioning annotations. Following [35], we treat each set of annotations as a single example during evaluation and discard videos with more than 3 sets of annotations.

**ActivityNet-Captions** [46] contains 14,934 untrimmed videos of various human activities. Different from YouCook2 and ViTT where most videos contain transcribed speech content, we find that 68% of videos in ActivityNet Captions do not have transcribed narration. On average, each video lasts 120s and is annotated with 3.7 temporally-localized sentences. The dataset is split into 10,009 and 4,925 videos for training and validation, respectively. Videos in the validation set are provided with two sets of dense video captioning annotations. Following prior work [102], we use both sets of annotations for evaluation, by computing the average of the scores over each set for SODA.c and by using the standard evaluation tool [46] for all other dense event captioning metrics. For video paragraph captioning, we follow [102] and report results on the ‘val-ae’ split that includes 2,460 videos [50, 128].

**MSR-VTT** [109] consists of 10,000 open domain video clips. The duration of each video clip is between 10 and 30 seconds. 20 natural language descriptions are manually annotated for each clip. The dataset is split into 6,513, 497 and 2,990 videos for training, validation and testing, respectively.

**MSVD** [9] consists of 1,970 open domain video clips. The duration of each video clip is between 10 and 30 seconds. Each video clip has roughly 40 manually annotated captions. The dataset is split into 1,200, 100 and 670 videos for training, validation and testing, respectively.

#### B.2. Implementation details

**Architecture.** The visual temporal transformer encoder  $f^t$ , the text encoder  $g^t$  and the text decoder  $h^t$  all have 12 layers, 12 heads, embedding dimension 768, and MLP hidden dimension of 2048. The text encoder and decoder sequences are truncated or padded to  $L = S = 1000$  tokens during pretraining, and  $S = 1000$  and  $L = 256$  tokens during finetuning. At inference, we use beam search decoding where we track the top 4 sequences and apply a length normalization of 0.6.

**Training.** We use the Adam optimizer [42] with  $\beta = (0.9, 0.999)$  and no weight decay. During pretraining, we use a learning rate of  $1e^{-4}$ , warming it up linearly (from

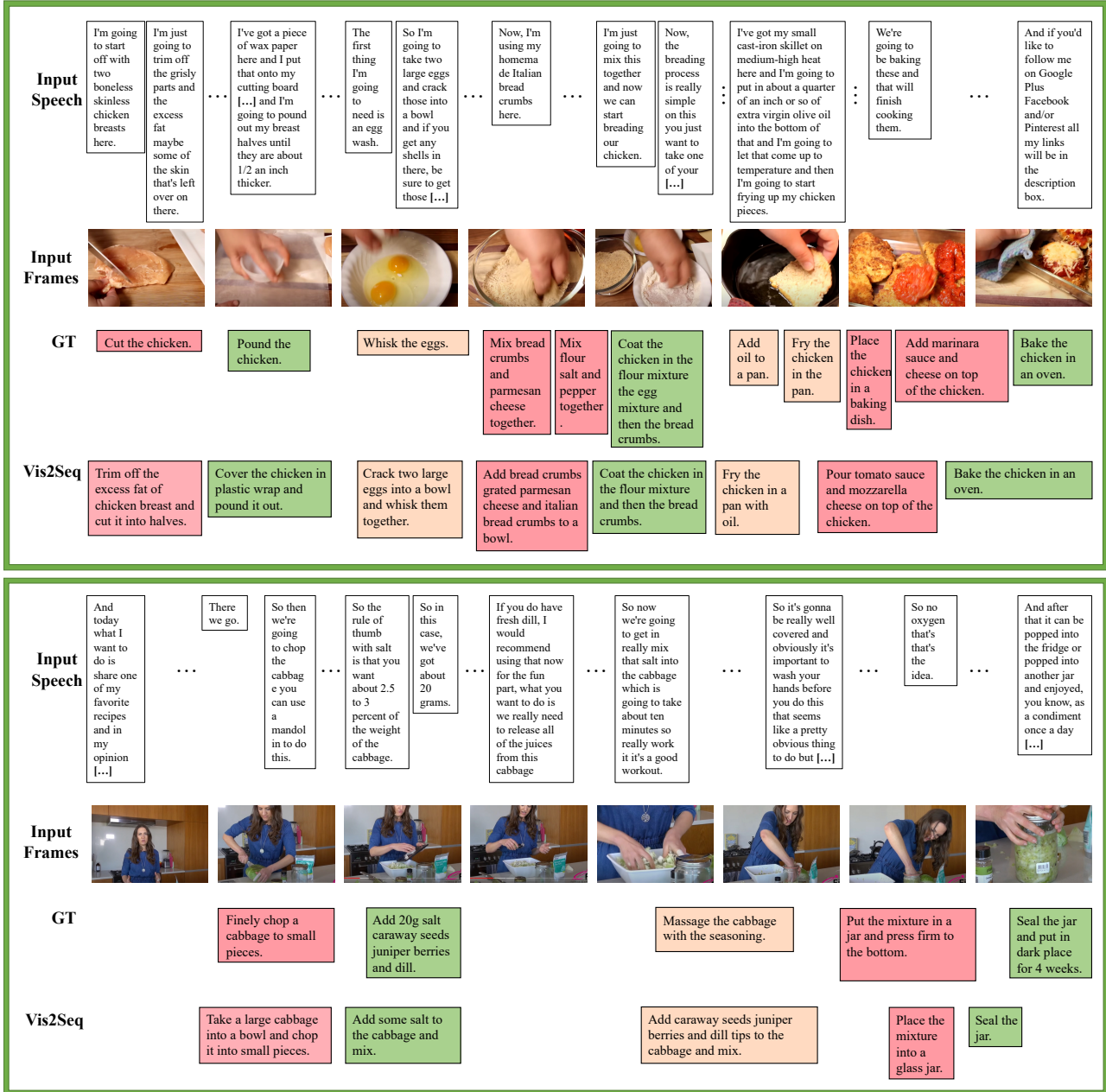


Figure 5. Examples of dense event captioning predictions of Vid2Seq on the validation set of YouCook2, compared with ground-truth.









0) for the first 1000 iterations, and keeping it constant for the remaining iterations. During finetuning, we use a learning rate of  $3e-4$ , warming it up linearly (from 0) for the first 10% of iterations, followed by a cosine decay (down to 0) for the remaining 90%. During finetuning, we use a batch size of 32 videos split on 16 TPU v4 chips. We finetune for 40 epochs on YouCook2, 20 epochs on ActivityNet Captions and ViTT, 5 epochs on MSR-VTT and 10 epochs on MSVD. We clip the maximum norm of the gradient to 0.1 during pretraining, and 1 during finetuning. For data

augmentation, we use random temporal cropping. For regularization, we use label smoothing [91] with value 0.1 and dropout [86] with probability 0.1.









## C. Experiments

In this section, we provide additional experiments that complement the results presented in Section 4. We first show the importance of pretraining in our proposed few-shot setting in Section C.1. Then we provide additional







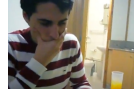



<b>Input Speech</b>	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
<b>Input Frames</b>								
<b>GT</b>	A weightlifter is standing on a stage.	He lifts the barbell before dropping it.			He jumps up and down in excitement.			
<b>Vis2Seq</b>	A very strong man is shown in a competition	He lifts a very heavy weight over his head.			He then drops the weight to the ground before shaking his hands.			

<b>Input Speech</b>	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
<b>Input Frames</b>								
<b>GT</b>	A man walks up to parallel bars while spectators, competitors, and officials are in the background.	The man performs a routine on the parallel bars.			The man finishes his routine and dismounts			
<b>Vis2Seq</b>	A man walks up to a set of uneven bars.	He mounts the bars, then spins himself around.			He does a handstand, then dismounts.			

<b>Input Speech</b>	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
<b>Input Frames</b>								
<b>GT</b>	A man is seen looking at the camera and leads into him playing a poker game with others.		One man deals cards and chips while speaking to one another.			They continue playing and speaking to one another.		
<b>Vis2Seq</b>	A man is sitting behind a table playing poker.		He deals cards to the people, then he puts them on the table.			The man puts the cards on the table, and puts the chips in the middle.		









<b>Input Speech</b>	Ø	Ø	Ø	Ø	Ø	Ø	Ø	Ø
<b>Input Frames</b>								
<b>GT</b>	A group of children are seen swimming in a pool.		The kids hit a ball back and forth in the water.			They fight over the ball, trying to get it into the goal.		
<b>Vis2Seq</b>	A picture of a sky is shown and leads into a group of boys playing a game of water polo.		The camera pans around a small group of kids playing and then a man chases a ball around.			The boys continue playing and one man hats off to the camera.		

Figure 6. Examples of dense event captioning predictions by Vid2Seq on the validation set of ActivityNet Captions, compared with ground-truth. The first three examples show successful predictions, while the last example illustrates a failure case where the model hallucinates events that are not visually grounded (‘one man hats off to the camera’). Note that in all of these videos, there is no transcribed speech.

ablation studies in the standard fully-supervised setting in Section C.2, where we ablate various factors including pre-training on long narrated videos, the pretraining dataset and the size of the visual backbone, the time tokenization process and the number of time tokens, the sequence construction process, the temporal positional embeddings and the

initialization of the language model.

### C.1. Importance of pretraining in few-shot settings

In Section 4.2, we show the benefits of our pretraining method in the fully-supervised setting, *i.e.* when using 100% of the downstream training dataset. In Table 10, we

Data	Pretrain	YouCook2			ViTT			ActivityNet		
		S	C	M	S	C	M	S	C	M
1. 1%	✗	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1
2. 1%	✓	2.4	10.1	3.3	2.0	7.4	1.9	2.2	6.2	3.2
3. 10%	✗	0.1	0.0	0.2	3.3	0.4	3.3	3.4	11.9	4.6
4. 10%	✓	3.8	18.4	5.2	10.7	28.6	6.0	4.3	20.0	6.1
5. 50%	✗	1.8	8.5	2.4	6.5	18.7	3.9	4.6	13.1	6.3
6. 50%	✓	6.2	32.1	7.6	12.5	38.8	7.8	5.4	27.5	7.8
7. 100%	✗	4.0	18.0	4.6	7.9	21.2	6.2	5.4	18.8	7.1
8. 100%	✓	<b>7.9</b>	<b>47.1</b>	<b>9.3</b>	<b>13.5</b>	<b>43.5</b>	<b>8.5</b>	<b>5.8</b>	<b>30.1</b>	<b>8.5</b>

Table 10. **Impact of our pretraining on few-shot dense event captioning**, by finetuning Vid2Seq using a small fraction of the downstream training dataset.

	Max number of narrations	YouCook2			ActivityNet		
		S	C	F1	S	C	F1
1.	No pretraining	4.0	18.0	18.1	5.4	18.8	49.2
2.	1	6.0	32.1	22.1	5.1	22.9	48.1
3.	10	6.5	34.6	23.6	5.4	27.1	50.3
4.	$\infty$	<b>7.9</b>	<b>47.1</b>	<b>27.3</b>	<b>5.8</b>	<b>30.1</b>	<b>52.4</b>

Table 11. **Ablation showing the importance of pretraining on long narrated videos**, by varying the maximum number of narration sentences that a randomly cropped video can cover.  $\infty$  means the cropping is unrestricted and can sample arbitrarily long videos.

	Pretraining Data	Model	YouCook2			ActivityNet		
			S	C	F1	S	C	F1
1.	ImageNet	ViT-B/16	6.6	40.2	24.3	4.5	17.2	49.3
2.	CLIP	ViT-B/16	7.7	46.3	26.5	5.6	28.4	51.7
3.	CLIP	ViT-L/14	<b>7.9</b>	<b>47.1</b>	<b>27.3</b>	<b>5.8</b>	<b>30.1</b>	<b>52.4</b>

Table 12. **Ablation on the pretraining data and model size of the visual backbone  $f^s$** .

further show that our pretraining method has a considerable importance in the few-shot setting defined in Section 4.4, *i.e.* when using a smaller fraction of the downstream training dataset. In particular, our pretraining method enables our Vid2Seq model to have a non zero performance when using only 1% of the downstream training dataset (rows 1 and 2).

## C.2. Additional ablation studies

We here complement ablation studies reported in Section 4.2, using the same default settings, evaluation metrics and downstream datasets.

**Pretraining on long narrated videos.** In Table 1, we show the benefits of pretraining on untrimmed videos in comparison with the standard practice of pretraining on short, trimmed, video-speech segments [35, 70, 80]. In Table 11, we further evaluate the importance of sampling long narrated videos during pretraining. By default, at each training iteration, we randomly temporally crop each narrated video without constraints, resulting in a video that can span

	Tokenization	$N$	YouCook2			ActivityNet		
			S	C	F1	S	C	F1
1.	Absolute	20	0.3	0.2	0.9	3.2	23.0	23.1
2.	Absolute	100	3.5	25.7	12.0	4.8	25.5	41.5
3.	Absolute	500	<b>7.9</b>	39.8	24.3	5.4	28.1	48.6
4.	Relative	20	7.2	39.6	23.7	5.6	29.0	49.4
5.	Relative	100	<b>7.9</b>	<b>47.1</b>	<b>27.3</b>	<b>5.8</b>	<b>30.1</b>	52.4
6.	Relative	500	7.2	40.0	25.0	5.7	28.6	<b>52.5</b>

Table 13. **Ablation on time tokenization (relative or absolute) and the number of time tokens  $N$** .

	Dot symbol between segments	Time tokens Position	YouCook2			ActivityNet		
			S	C	F1	S	C	F1
1.	✗	After text	7.9	48.3	26.7	5.6	29.8	51.1
2.	✓	After text	<b>8.3</b>	<b>50.9</b>	26.2	5.7	<b>30.4</b>	51.8
3.	✗	Before text	8.0	50.0	<b>27.3</b>	5.6	28.2	50.7
4.	✓	Before text	7.9	47.1	<b>27.3</b>	<b>5.8</b>	30.1	<b>52.4</b>

Table 14. **Ablation on the sequence construction process.**

over hundreds of transcribed speech sentences. We here evaluate a baseline that constrains this cropping process such that the cropped video only spans over a given maximum number of narration sentences. Even with a maximum of 10 narration sentences, this baseline significantly underperforms our model trained in default settings where we sample longer untrimmed narrated videos (rows 1, 2 and 3). This demonstrates that our model benefits from pretraining on long narrated videos.

**Visual features.** In Table 4, we show the benefits of scaling up the size of the pretraining dataset of narrated videos and the size of the language model. In Table 12, we further analyze the importance of the pretraining dataset and size of the visual backbone  $f^s$ . We find that CLIP pretraining [77] considerably improves over ImageNet pretraining [87] with the same ViT-B/16 visual backbone model (row 2 vs 1). Furthermore, scaling up the visual backbone size from ViT-B/16 to ViT-L/14 brings additional improvements (row 3 vs 2).

**Time tokenization and number of time tokens.** In Table 13, we further ablate the time tokenization process presented in Section 3.1. Our default time tokens represent relative timestamps in a video, as we quantize a video of duration  $T$  into  $N$  equally-spaced timestamps. Another possibility is to use time tokens that represent absolute timestamps in the video, *i.e.* the  $k$ -th token represents the  $k$ -th second in the video. For both these variants, we vary the number of time tokens  $N$ . For the relative time tokens, increasing  $N$  makes the quantization more fine-grained but also spreads the data into more time tokens. On the other hand, for the absolute time tokens, increasing  $N$  increases the video du-

	Temporal embeddings	YouCook2			ActivityNet		
		S	C	F1	S	C	F1
1.	$\times$	6.8	42.0	24.9	5.3	27.0	50.6
2.	$\checkmark$	<b>7.9</b>	<b>47.1</b>	<b>27.3</b>	<b>5.8</b>	<b>30.1</b>	<b>52.4</b>

Table 15. Ablation on the temporal positional embeddings.

	Language Model Initialization	Video-text Pretraining	YouCook2			ActivityNet		
			S	C	F1	S	C	F1
1.	$\times$	$\times$	0.9	4.2	7.6	4.3	23.7	41.2
2.	$\checkmark$	$\times$	4.0	18.0	18.1	5.4	18.8	49.2
3.	$\times$	$\checkmark$	<b>8.8</b>	<b>51.3</b>	<b>28.4</b>	5.7	28.7	51.2
4.	$\checkmark$	$\checkmark$	7.9	47.1	27.3	<b>5.8</b>	<b>30.1</b>	<b>52.4</b>

Table 16. Ablation on language model initialization and pre-training.

ration that the time tokens can cover. We find that the best dense video captioning results are obtained with the relative time tokens and  $N = 100$  time tokens (row 5).

**Sequence construction.** In Table 14, we further ablate the sequence construction process presented in Section 3.1. Our default sequence inserts the start and end time tokens of each segment before its corresponding text sentence. Another possibility is to insert time tokens after each corresponding text sentence. We find that both variants achieve similar results (rows 2 and 4), with the default sequence (row 4) resulting in slightly higher event localization performance (F1 Score) but slightly lower dense captioning results overall. Furthermore, we observe that the dot symbols indicating the separation between different events have low importance (rows 1 and 2, rows 3 and 4).

**Temporal positional embeddings.** In Table 1, we show that time tokens in the speech sequence provide temporal information about the speech transcript to our model. In Table 15, we also evaluate the importance of the temporal positional embeddings which communicate temporal information from the visual stream to our model. We find that these temporal embeddings are beneficial (row 2 vs 1).

**Language model initialization and pretraining.** In Table 4, we show the benefits of using T5-Base instead of T5-Small. In Table 16, we further investigate the importance of initializing the language model from weights pre-trained on Web text. Without pretraining on narrated videos, we find that text-only initialization is helpful (rows 1 and 2). Interestingly, after pretraining on narrated videos, we find that text-only initialization has little importance (rows 3 and 4), as it slightly improves the performance on ActivityNet Captions while resulting in a slight drop of performance on YouCook2. We believe that this may be because

of the domain gap between Web text and the imperative-style dense captions in YouCook2, which are more similar to transcribed speech in YT-Temporal-1B.