Exploring the Role of Audio in Video Captioning

Yuhan Shen[†] [*] Linjie Yang[‡] Longyin Wen[‡] Haichao Yu[‡] Ehsan Elhamifar[†] Heng Wang[‡]
[†] Northeastern University [‡] ByteDance

Abstract

Recent focus in video captioning has been on designing architectures that can consume both video and text modalities, and using large-scale video datasets with text transcripts for pre-training, such as HowTo100M. Though these approaches have achieved significant improvement, the audio modality is often ignored in video captioning. In this work, we present an audio-visual framework, which aims to fully exploit the potential of the audio modality for captioning. Instead of relying on text transcripts extracted via automatic speech recognition (ASR), we argue that learning with raw audio signals can be more beneficial, as audio has additional information including acoustic events, speaker identity, etc. *Our contributions are twofold. First, we observed that the model overspecializes to the audio modality when pre-training with both video and audio modality, since the ground truth (*i.e.*, text transcripts) can be solely predicted using audio. We proposed a Modality Balanced Pre-training (MBP) loss to mitigate this issue and significantly improve the performance on downstream tasks. Second, we slice and dice different design choices of the cross-modal module, which may become an information bottleneck and generate inferior results. We proposed new local-global fusion mechanisms to improve information exchange across audio and video. We demonstrate significant improvements by leveraging the audio modality on four datasets, and even outperform the state of the art on some metrics without relying on the text modality as the input.*

1. Introduction

Large-scale pre-training [3, 18, 29, 38, 47, 57, 65] plays a key role in boosting modern deep learning models. It is even more so for vision and language tasks, such as video captioning [1,10,32,36,41,45,56,63,64], where leveraging large video datasets with text supervision for pre-training is essential to achieve competitive results. However, manually annotating captions for video datasets is costly and not scalable. Thus existing video captioning datasets [46,59,61,68] are often limited in size.

[*]Work done during YS's internship at ByteDance.

To address this challenge, recent work collected datasets from instructional videos, where ASR transcripts can be used as text supervision, *e.g.*, How2 [42], CrossTask [71], HowTo100M [30], HD-VILA-100M [62], *etc*. This has established a new trend of pre-training on large-scale video datasets with text transcripts for video captioning [14, 28, 43]. We argue that text transcripts from ASR only includes partial information from audio, and hypothesize end-to-end learning using the audio modality can potentially lead to better performance, since audio can provide additional information (shown in Fig. 1) including acoustic events, speaker identity, *etc*.

More specifically, our paper seeks to better understand the following questions:

- *To what extent, can the audio modality improve video captioning?*
- *How can the potential of the audio modality be fully realized in an audio-visual framework for captioning?*

To this end, we start with a simple multi-modal pre-training framework for video captioning with ASR transcripts as supervision (shown in Fig. 2), and look into different components that may hinder the performance of the pre-trained audio-visual model on the downstream datasets.

First, we observed that simply jointly training of the audio and video modalities may result in degenerated models that overspecialize to audio modality and underfit on video modality. As text transcripts are used as video captions during pre-training, the model essentially learns to cheat and solve the ASR problem instead of extracting information from both visual and audio signals. To mitigate this issue, we proposed the Modality Balanced Pre-training (MBP) loss that takes into account both the unimodal losses and cross-modal loss. We introduce a weighting mechanism to balance different modalities during training. Fig. 4 shows that our MBP loss enforces the model to focus on the underfitted video modality and drives the final loss much smaller.

Second, we thoroughly investigated the design of the cross-modal fusion module, which is responsible for the information exchange between audio and video modality. An improperly designed cross-modal fusion module may become an information bottleneck and result in inferior performance for video captioning. We proposed new local-

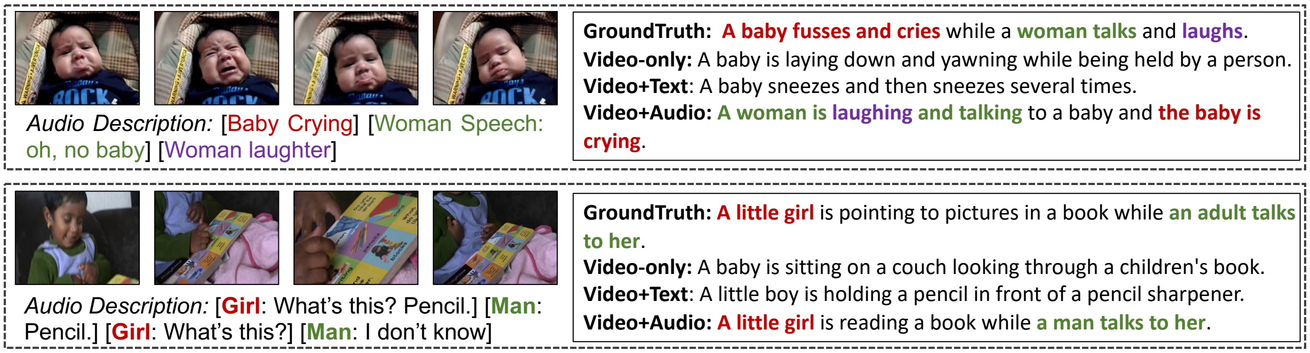


Figure 1. Audio provides critical complementary information in multi-modal video captioning. We show two examples of generated captions when we input 1) only video, 2) video and ASR text, and 3) video and audio. Audio can provide additional information that cannot be obtained from visual modality or ASR text, *e.g.*, sound of crying, laughter, and speaker gender.

global fusion modules to encourage the information flow across different modalities. We analyzed the relevance of the annotated captions to the audio modality on downstream datasets, and observed that the local fusion modules are more beneficial to the flow of fine-grained information like single words in speech, while the global fusion modules are more effective on holistic information like acoustic events or scenes. The local-global design is able to capture information at different granularities, and mingle audio and video information at different levels. Compared with existing designs, our local-global fusion has shown empirically better results.

By combining the two contributions, we demonstrate that audio is crucial to video captioning and provides both speech and non-speech information. Fig. 1 shows a few examples on how our model effectively integrates the information from both the audio and video modality, and generates better captions than video-only and video-text variants.

We summarize our contributions as follows:

- Proposed to pre-train video captioning models based on video and audio modalities, and demonstrated the benefits of audio on four benchmarks.
- Proposed the MBP loss to balance different modalities automatically during training, and ease the issue of overspecialization to the audio modality.
- Did an extensive evaluation on the effects of different cross-modal fusion modules on audio-visual video captioning, and proposed a novel local-global fusion module to effectively integrate audio and video information.

2. Related Work

Video Captioning. Most works in video captioning [1, 10, 32, 36, 55, 56, 63, 67] focus on designing a better model to generate text descriptions given precomputed video features via an encoder-decoder framework. SwinBert [25] attempted to train the encoder-decoder framework directly from raw video pixels. In addition to visual modality, some works studied video captioning from visual data and ASR texts [13, 28, 43, 44, 46, 50]. A few prior works also studied audio-visual video captioning [7, 15, 40, 51], but they are often limited to small-scale video captioning datasets and precomputed input features. To the best of our knowledge, we propose the first end-to-end audio-visual video captioning framework.

Multi-Modal Pre-training. A growing number of works are investigating multi-modal pre-training in videos, *e.g.*, video-text pre-training [20, 28–30, 35, 47, 66] and video-text-audio pre-training [3, 4, 41, 65], which mostly adopt contrastive learning and/or masked language modeling to learn better representations for downstream tasks. As only encoders are trained for multiple modalities, a separate decoder needs to be trained on top of the encoders for generative tasks such as video captioning. MV-GPT [43] shows the benefits of pre-training an end-to-end encoder-decoder framework to video captioning. Unlike MV-GPT that relies on ASR text as input, our framework directly uses video and audio. A Textless Vision-Language Transformer (TVLT) [49] was recently proposed to take visual and audio inputs for multi-modal representation learning without ASR inputs. However, the pre-trained TVLT is a discriminative model that cannot be directly applied to generative tasks. While a multi-modal network receives more information and is expected to boost performance, recent works [16, 33, 37, 58] have identified a key challenge in training a multi-modal network that one modality may converge faster than other modalities and undermine the representation learning of other modalities. We propose a Modality Balanced Pre-training objective to mitigate this issue and facilitate a powerful audio-visual video captioning model.

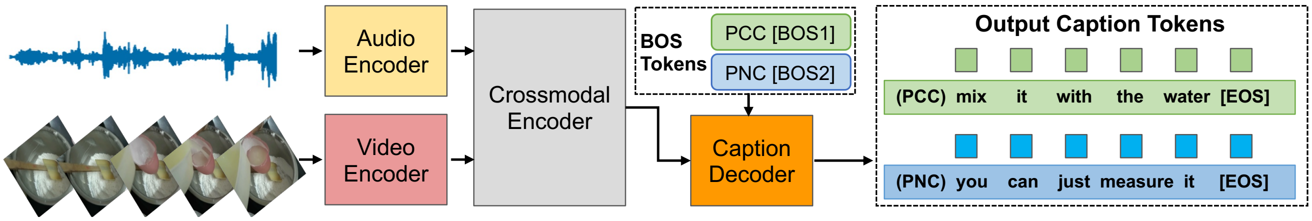


Figure 2. Overview of our audio-visual video captioning framework. We design two tasks for caption generation during pre-training: Predict Current Caption (PCC) and Predict Next Caption (PNC). For downstream fine-tuning, we only adopt PCC because the goal is to predict current caption given the input frames and audio.

Cross-Modal Fusion. Given the representations of multiple modalities, a cross-modal fusion module [17, 29, 30, 38] fuses these representations into a shared space to generate cross-modal representations. In order to fuse a sequence of representations generated by Transformers [53], there are two major types of cross-modal fusion modules: **merged fusion**, and **cross fusion**. In merged fusion, the two modalities are concatenated and fed into a Transformer block [21, 23, 28, 57]. In cross fusion, the two modalities are fed into different Transformer blocks with cross attention [27, 43, 44, 48, 52]. Besides, some recent works propose variants of cross-modal fusion modules that use bottleneck tokens [31] or prune single-modal units [60] to control the flow of cross-modal interaction.

3. Audio-Visual Video Captioning

In this section, we propose our methods for audio-visual video captioning. In Sec. 3.1, we present an overview of our framework. In Sec. 3.2, we describe our MBP loss to balance different modalities during pre-training. In Sec. 3.3, we investigate different cross-modal fusion modules and propose a local-global fusion module to improve information flow between audio and video at different granularities.

3.1. Framework Overview

As shown in Fig. 2, we follow the common practice in video captioning [28, 43] and use an encoder-decoder framework including four main modules: a video encoder, an audio encoder, a cross-modal encoder, and a caption decoder, all of which are Transformer architectures [53].

Given a video, the video encoder extracts a sequence of D-dim video embeddings $\varphi^v \in \mathbb{R}^{N_v \times D}$ from the frames, and the audio encoder extracts a sequence of D-dim audio embeddings $\varphi^a \in \mathbb{R}^{N_a \times D}$ from the audio spectrogram, where N_v and N_a are the numbers of video tokens and audio tokens. Then, we employ a cross-modal encoder $f_\Theta(\cdot)$ to generate multi-modal embeddings $\varphi^c \in \mathbb{R}^{(N_a+N_v) \times D}$ for cross-modal interaction. Finally, we use a decoder $g_{\Theta'}(\cdot)$ conditioned on φ^c to output the captions auto-regressively. By default, we use Video Swin Transformer [26] as the video encoder, and Audio Spectrogram Transformer [12] as the audio encoder.

Inspired by [43], we design two tasks during pre-training: Predict Current Caption (PCC) and Predict Next Caption (PNC). Given the audio-visual embeddings as the context, we feed two `BOS` (Beginning of Sentence) tokens to the caption decoder for caption generation, namely `BOS1` and `BOS2`, which initiate the prediction of the current and next caption respectively. The PNC task enforces the model to anticipate future events, which is more challenging and requires higher level of semantic understanding. Note that this multi-task training is only used for pre-training. For downstream fine-tuning, we only feed `BOS1` token to the decoder as the goal is to predict the current caption.

3.2. Modality Balanced Pre-training

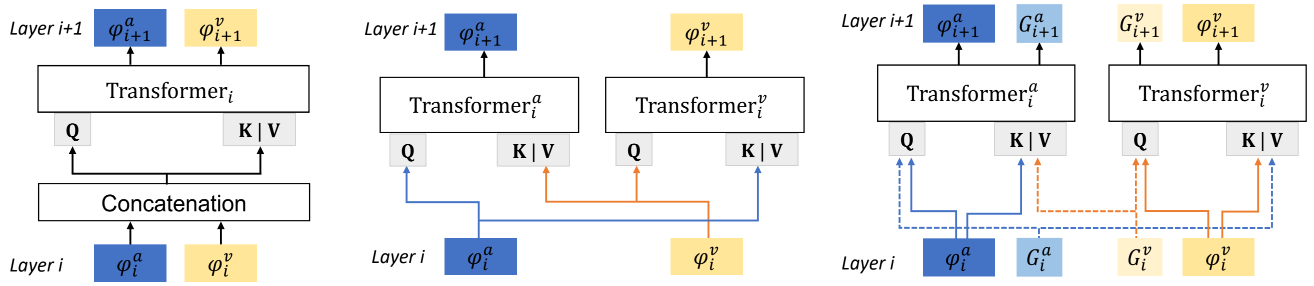
With captions as supervision, a commonly used objective is to minimize the negative log-likelihood:

$$L = \mathcal{L}(g_{\Theta'}(f_\Theta(\varphi^a, \varphi^v)), y), \quad (1)$$

where $\mathcal{L}$ is the cross entropy loss and y is the ground-truth caption. We refer to this loss as audio-video decoder loss, as both audio and video features are input into the decoder.

Prior works [3, 29, 43, 57, 65, 66] have proved that large-scale pre-training is essential to multi-modal learning. Thus, we pre-train our audio-visual video captioning model on a large-scale video dataset before fine-tuning it on downstream captioning datasets. Although there are some large-scale datasets for video-language pre-training, *e.g.*, HowTo100M [30] and HD-VILA-100M [62], they lack manually annotated captions and use ASR transcripts as text supervision. One significant issue based on our experiments (Fig. 4) is that the model tends to learn only from the audio and ignore the video modality, since the ASR caption can be solely derived from human speech in the audio. Even with the two tasks for current and future caption predictions, the model still favors the audio features when predicting the current caption.

To address the issue, we propose MBP to balance different modalities during training. We add losses for audio-only and video-only predictions to improve the learning of the two modalities. To measure how well the model exploits a



(a) Merged Fusion (b) Cross Fusion (c) Global Cross Fusion

Figure 3. Three different cross-modal fusion designs. Superscripts a, v denote audio and video modalities. $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$ represent query, key, and value in multi-head attention models. For clarity, we use dot lines to denote the flow of global tokens in (c).

certain modality, we first define the following two losses:

$$\begin{aligned} L_a &= \mathcal{L}(g_{\Theta'}(f_{\Theta}(\varphi^a, \mathbf{0})), y); \\ L_v &= \mathcal{L}(g_{\Theta'}(f_{\Theta}(\mathbf{0}, \varphi^v)), y), \end{aligned} \tag{2}$$

where we set video embeddings to all-zeros to get the audio-only decoder loss L_a, and set audio embeddings to all-zeros to get the video-only decoder loss L_v based on Eq. (1). We refer to L_a or L_v as mono-modal losses. If a mono-modal loss is small, it means that the corresponding modality is well utilized by the model. We then measure the gap between the multi-modal loss and the mono-modal losses by a Mono-to-Multi Discrepancy (MMD) index:

$$G_a = (L_a - L)^2; G_v = (L_v - L)^2, \tag{3}$$

where G_a and G_v measure the discrepancy between audio-only/video-only decoder loss and audio-video decoder loss, respectively. Inspired by G-Blend [58] that uses Overfitting-to-Generalization Ratio (OGR) to iteratively update training weights for different modalities, we guide the weights of mono-modal losses based on whether the modality is well utilized by the model:

$$L_{pretrain} = L + w_a L_a + w_v L_v, \tag{4}$$

where w_a and w_v are weights updated over iterations:

$$w_m^{(t)} = \beta w_m^{(t-1)} + (1-\beta)\tilde{w}_m^{(t)}, m \in \{a, v\}, \tag{5}$$

where $\beta \in (0, 1)$ is a smoothing hyperparameter, t is the iteration number, and $\tilde{w}_m^{(t)}$ is obtained using a softmax function over the MMD of two modalities at the current iteration:

$$\tilde{w}_m^{(t)} = \frac{\exp\left(\alpha G_m^{(t)}\right)}{\sum_{m'} \exp\left(\alpha G_{m'}^{(t)}\right)}, m \in \{a, v\}, \tag{6}$$

where $\alpha > 0$ is a temperature hyperparameter. If G_m is large for a certain modality, we will assign a higher weight w_m for modality m in Eq. (4). This strategy enforces the model to attend more to modality m and mitigate its over-specialization to the other modality. As the optimization progresses and the gaps G_a and G_v change over time, we dynamically adjust w_a and w_v according to Eq. (5) to enhance the underfitted modality. Our strategy can be easily extended to more than two modalities, which is out of the scope of this work.

3.3. Cross-Modal Fusion: Beyond Cross Attention

In addition to the MBP objective, the cross-modal encoder is another component where different modalities interact with each other. In this section, we explore different design choices for the cross-modal fusion module to better leverage the audio modality for video captioning.

3.3.1 Background

Given N_q d-dim query vectors $\mathbf{Q} \in \mathbb{R}^{N_q \times d}$, and N_v key-value pairs, $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{N_v \times d}$, an attention function maps queries to output vectors with a scaled dot product: $\text{Att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}$. A Transformer (TFMR) layer consists of a Multi-Head Attention (MHA) module and a Feed-Forward Block (FFB), denoted by $\text{Transformer}(\mathbf{X}, \mathbf{Y}) = \text{FFB}(\text{MHA}(\mathbf{X}, \mathbf{Y}, \mathbf{Y}))$.

3.3.2 Cross-Modal Fusion

As mentioned in Sec. 3.1, we feed the audio tokens $\varphi^a \in \mathbb{R}^{N_a \times D}$ and the video tokens $\varphi^v \in \mathbb{R}^{N_v \times D}$ into a cross-modal encoder for cross-modal fusion. Below we introduce two popular fusion methods.

In **merged fusion** (Fig. 3a), the tokens of two modalities are concatenated and then passed into Transformer blocks:

$$\varphi^a_{i+1}, \varphi^v_{i+1} = \text{Transformer}_i([\varphi^a_i; \varphi^v_i], [\varphi^a_i; \varphi^v_i]). \tag{7}$$

Audio tokens φ^a_i and video tokens φ^v_i are the inputs of the i-th Transformer layer. Each audio token can attend to all audio and video tokens, and it is the same for video tokens.

In **cross fusion** (Fig. 3b), each modality has its own Transformer layers, and different modalities exchange information via cross attention, *i.e.*, one modality is used as the context (keys and values) of the other modality:

$$\begin{aligned} \varphi_{i+1}^a &= \text{Transformer}_i^a(\varphi_i^a, \varphi_i^v), \\ \varphi_{i+1}^v &= \text{Transformer}_i^v(\varphi_i^v, \varphi_i^a). \end{aligned} \tag{8}$$

Each layer has two modality-specific Transformers.

3.3.3 Global Cross Fusion

We use "local fusion" to denote the fusion modules in (7)(8) as the interaction is among local tokens, regardless of intra-modality or inter-modality. However, only using local interaction can be sub-optimal, as local tokens may be noisy and less informative for cross-modal interaction. To select more salient features, we introduce global tokens G^a and G^v for each modality (Fig. 3c), which serve as a global representation of the audio clip or the video clip. Instead of using all video tokens as the context for audio tokens, we use the global video token as the context:

$$\begin{aligned} \varphi_{i+1}^a, G_{i+1}^a &= \text{Transformer}_i^a([\varphi_i^a; G_i^a], [\varphi_i^a; G_i^v]), \\ \varphi_{i+1}^v, G_{i+1}^v &= \text{Transformer}_i^v([\varphi_i^v; G_i^v], [\varphi_i^v; G_i^a]). \end{aligned} \tag{9}$$

We restrict all cross-modal attention to be via the global tokens, and local tokens are only used for intra-modal attention. The global token of the first cross layer is learnable and is initialized with a Gaussian distribution.

3.3.4 Local-Global Fusion

Local and global tokens capture information in different granularities. Local tokens capture local features such as words in the speech or objects in a video frame, while global tokens capture high-level concepts like sounds of laughter or people gathering on a street. To leverage multigranular information, we propose to combine local fusion and global cross fusion. Let $\varphi_{i+1}^{a(G)}$ and $\varphi_{i+1}^{v(G)}$ denote the embeddings from global cross fusion in Eq. (9), and $\varphi_{i+1}^{a(L)}$ and $\varphi_{i+1}^{v(L)}$ denote the embeddings from local fusion in Eq. (7) or Eq. (8), we compute an average of these embeddings before feeding them into the next layer, *i.e.*,

$$\varphi_{i+1}^m = (\varphi_{i+1}^{m(G)} + \varphi_{i+1}^{m(L)})/2, \; m \in \{a, v\}. \tag{10}$$

This unified fusion module is named as "local-global fusion", and is able to progressively refine the tokens using both local and global guidance. We name this variant with merged fusion as "local-global merged fusion", and the variant with cross fusion as "local-global cross fusion".

4. Experiments

4.1. Pre-training and Downstream Datasets

We use the HowTo100M dataset [30] for pre-training, and four video captioning datasets, including YouCook2 [68], MSRVTT [61], VATEX [59] and ActivityNet-Captions [46] for evaluation. We use four evaluation metrics: BLEU-4 (B) [34], METEOR (M) [5], ROUGE-L (R) [24], and CIDEr (C) [54].

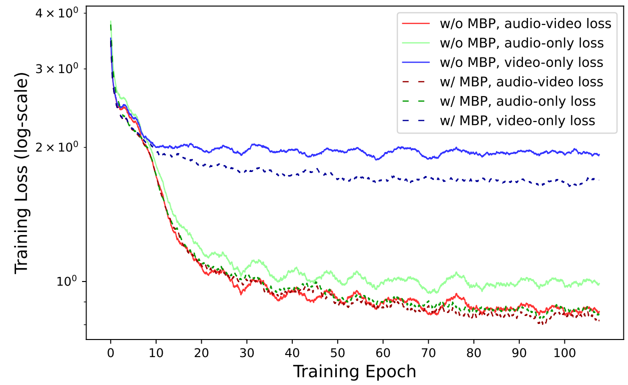


Figure 4. The pre-training losses without and with MBP. (Solid lines: without MBP. Dotted lines: with MBP.)

HowTo100M consists of $1.2M$ YouTube instructional videos. We download videos with ASR transcripts and audio from YouTube, and remove unavailable videos, resulting in $1.08M$ videos in total. Following [28], we start with a single ASR sentence and iteratively expand the video clip forward/backward by adding nearby sentences until the clip is longer than 5 seconds. **YouCook2** contains $2,000$ cooking videos with $15.4k$ video clips. Each video clip is annotated with a single sentence. **MSRVTT** contains $10k$ open-domain video clips for video captioning. Each video clip is annotated with 20 captions. **VATEX** consists of $41,250$ videos. Each video is annotated with 10 English captions and 10 Chinese captions, and we use the English captions. **ActivityNet-Captions** is a video paragraph captioning dataset consisting of $100k$ captions for $20k$ long videos. To be consistent with the other datasets, we train our model on sentence-level captions. Following [69], we compose paragraph-level captions by simply concatenating sentence-level captions and evaluate the performance at paragraph-level. Please see the supplementary material for more details.

The performance of video captioning is closely tied to the language styles of the annotations, so we analyzed how relevant the annotated captions are to the audio modality on each dataset in the supplementary material. As an observation, a large portion of YouCook2 captions are mentioned in speech, and the captions on VATEX are most relevant to acoustic events, scenes, or sound patterns among the four datasets. We will show how the audio relevance affects the performance in Sec. 4.3.

MBP	PCC	PNC	YouCook2				MSRVTT				VATEX				ActivityNet-Captions			
			B	M	R	C	B	M	R	C	B	M	R	C	B	M	R	C
G-Blend [58]	✓	✓	19.4	23.4	48.6	208.5	44.0	29.6	62.4	55.1	38.0	24.9	52.1	68.7	11.6	16.1	30.6	25.3
	✓		15.9	20.2	43.0	166.8	41.4	28.5	60.7	47.6	31.2	21.9	47.7	50.7	10.0	14.6	28.4	20.1
✓	✓		18.5	22.7	47.0	192.4	44.7	29.9	62.7	53.5	36.9	24.7	51.7	67.5	11.3	15.5	30.2	24.7
	✓	✓	17.2	21.7	45.8	184.2	42.0	28.4	60.8	48.4	31.2	22.0	48.0	51.4	10.4	14.9	28.9	20.2
✓	✓	✓	**20.6**	**24.2**	**49.6**	**217.0**	**46.0**	**30.6**	**64.0**	**57.0**	**38.8**	**25.9**	**52.9**	**73.5**	**11.7**	**16.1**	**30.7**	**26.1**

Table 1. Ablation studies on multi-modal pre-training with our audio-visual captioning framework from Fig. 2. MBP: Modality Balanced Pre-training; PCC: Predict Current Caption; PNC: Predict Next Caption.

Fusion	YouCook2				VATEX			
	B	M	R	C	B	M	R	C
cross	19.5	23.4	48.9	211.2	37.8	24.7	52.0	68.9
merged	19.9	23.6	49.1	210.7	38.4	25.1	52.3	71.1
global	18.6	22.9	48.0	202.3	**39.2**	25.5	52.8	72.6
local-global cross	19.9	23.9	49.2	213.9	38.6	25.8	**53.0**	73.4
local-global merged	**20.6**	**24.2**	**49.6**	**217.0**	38.8	**25.9**	52.9	**73.5**

Table 2. Ablation studies on cross-modal fusion modules.

4.2. Experimental Setup

Video Encoder: We sample 16 frames from each video clip. The frames are fed into the Video Swin Transformer [26] initialized with the weights pre-trained on Kinetics 600 [6] and tokenized into $N_v = 8 \times 7 \times 7 = 392$ video tokens. Then we add a linear layer to project the dimension of each video token to $D = 768$, to be consistent with the other modules.

Audio Encoder: We first extract log mel spectrogram of the audio. Following [65], each audio is resampled to 22,050Hz and divided into frames of 1536 samples with hop length of 588. Then we apply 64 mel-scale filters. We use a 12-layer Transformer on audio spectrogram to output $N_a = 64$ audio tokens with feature dimension of 768.

Cross Encoder: We use a 3-layer Transformer as the cross-modal encoder. The feature embeddings of different modalities will be added with the position embedding and token type embedding to distinguish the position and modality of the tokens. We comprehensively compare the results of the cross-modal fusion methods introduced in Sec. 3.3. Overall, local-global merged fusion performs best, so we use local-global merged fusion unless stated otherwise.

Caption Decoder: We use a 3-layer Transformer as the caption decoder. In training, we use causal masking to ensure that only history inputs are used. In testing, we use beam-search with beam width of 5 for caption generation.

Training Details: We pre-train the model on HowTo100M for 100 epochs using Adam optimizer [19]. The base learning rate for pre-training is 10^{-4} and we use a linear decay learning rate schedule with a warm-up of 10% training epochs as in [28]. For fine-tuning, we set the initial learning rate as 10^{-5}. It is noted that we employ modality balancing during pre-training but not during downstream fine-tuning, to allow the model to adapt to the more informative modality for caption generation.

4.3. Results and Discussions

Multi-modal pre-training objectives. We evaluate the effects of our pre-training objectives in Tab. 1. Our proposed MBP improves the performance by a large margin on four datasets. The addition of Predicting Next Caption leads to a remarkable boost as well. We also compare MBP with G-Blend [58]: G-Blend aims to reduce overfitting, which happens when the model performs well on the training set but fails to generalize, whereas MBP aims to reduce overspecialization, which happens when the model performs well on a single modality but fails on the other modalities. Hence, G-Blend updates mono-modal weights by computing Overfitting-to-Generation Ratio while MBP uses Mono-to-Multi Discrepancy. As shown in Tab. 1, G-Blend also improves the performance, but MBP consistently outperforms G-Blend, showing its suitability in avoiding overspecialization in audio-visual captioning pre-training. In Fig. 4, we show the training curve with and without MBP, including multi-modal loss (audio-video loss) and mono-modal losses (audio-only loss and video-only loss). Note that for the experiments without MBP, audio-only loss and video-only loss are only computed for analysis and not back-propagated. Although the two audio-video losses have similar scales, both audio-only loss and video-only loss are reduced with MBP, especially for video-only loss. We conjecture that the model reduces overspecialization to the audio modality and learns to better utilize the video modality with MBP. Though there is only a small decrease in the audio-video training loss, the performance on downstream tasks are significantly improved (see Tab. 1), which demonstrates the effectiveness of our pre-training objective.

Cross-modal fusion modules. We conduct ablation studies on cross-modal fusion in Tab. 2. Comparing the two local fusion modules, merged fusion performs favorably against cross fusion. Compared with local fusion modules, global cross fusion performs better on VATEX but performs worse on YouCook2. We note that annotated captions are very related to audio events and scenes on VATEX while

Modalities	YouCook2				MSRVTT				VATEX				ActivityNet			
	B	M	R	C	B	M	R	C	B	M	R	C	B	M	R	C
A	13.4	17.6	38.5	138.9	32.3	23.9	54.3	24.8	15.4	15.6	37.6	14.2	6.3	11.2	24.5	6.6
V	11.7	18.0	41.5	139.2	42.6	28.5	61.0	52.0	36.4	24.6	51.2	67.3	9.7	14.4	27.8	21.3
V+A	20.6	24.2	49.6	217.0 (77.8↑)	46.0	**30.6**	64.0	57.0 (5.0↑)	38.8	25.9	52.9	73.5 (6.2↑)	11.7	**16.1**	30.7	26.1 (4.8↑)
V+T	19.1	23.3	48.7	205.6 (66.4↑)	43.3	29.1	61.7	53.4 (1.4↑)	37.7	25.2	52.0	69.0 (1.7↑)	11.2	15.9	29.4	24.9 (3.6↑)
V+A+T	**20.9**	**24.4**	**49.9**	**221.6** (82.4↑)	**46.4**	30.2	**64.1**	**57.3** (5.3↑)	**39.1**	**26.3**	**53.4**	**73.7** (6.4↑)	**11.8**	**16.1**	**30.9**	**26.4** (5.1↑)

Table 3. The performance when we input different modalities (V: video, A: audio, T: text). The top two results are in **bold** and underline. We also show the improvement of multi-modal video captioning over the video-only method in terms of CIDEr.

Method	Pre-training dataset/model	Inputs	B	M	R	C
MV-GPT †	HT100M + GPT-2	V+T	21.9	27.1	49.4	221.0
MART	-	V	8.0	15.9	-	36.0
SwinBert	-	V	9.0	15.6	37.3	109.0
ActBert	HT100M	V	5.4	13.3	30.6	65.0
M-MASS	YT8M-cook + Recipe1M	V+T	12.0	18.3	39.0	123.0
Value	HT100M + TV	V+T	12.4	18.8	40.4	130.4
UniVL	HT100M	V+T	17.4	22.4	46.5	181.0
Ours	HT100M	V+A	20.6	24.2	49.6	217.0
Ours	HT100M	V+T	19.1	23.3	48.7	205.6
Ours	HT100M	V+A+T	**20.9**	**24.4**	**49.9**	**221.6**

(a) YouCook2

Method	Pre-training dataset/model	Inputs	B	M	R	C
MV-GPT †	HT100M + GPT-2	V+T	48.9	38.7	64.0	60.0
CLIP4VLA †	HT100M + AudioSet + CLIP	V+A+T	46.7	31.1	64.4	58.0
SwinBert	-	V	45.4	**30.6**	64.1	55.9
MMCNN	-	V+A	42.7	28.5	61.5	47.2
MGSA	-	V+A	45.4	28.6	-	50.1
Decembert	HT100M	V+T	45.2	29.7	**64.7**	52.0
UniVL	HT100M	V+T	41.8	28.9	60.8	50.0
Ours	HT100M	V+A	46.0	**30.6**	64.0	57.0
Ours	HT100M	V+T	43.3	29.1	61.7	53.4
Ours	HT100M	V+A+T	**46.4**	30.2	64.1	**57.3**

(b) MSRVTT

Method	Pre-training dataset/model	Inputs	B	M	R	C
CLIP4VLA †	HT100M + AudioSet + CLIP	V+A+T	36.4	25.0	54.7	59.7
SwinBert	-	V	38.7	26.2	53.2	73.0
MGRMP	-	V	34.2	23.5	50.3	57.6
Value	HT100M + TV	V+T	32.9	24.1	50.1	58.1
Ours	HT100M	V+A	38.8	25.9	52.9	73.5
Ours	HT100M	V+T	37.7	25.2	52.0	69.0
Ours	HT100M	V+A+T	**39.1**	**26.3**	**53.4**	**73.7**

(c) VATEX

Method	Pre-training dataset	Inputs	B	M	R	C
VTransformer	ActivityNet	V	9.31	15.54	-	21.33
Transformer-XL	ActivityNet	V	10.25	14.91	-	21.71
MART ‡	ActivityNet	V	9.78	15.57	30.63	22.16
COOT ‡	HT100M	V	10.85	15.99	**31.45**	**28.19**
Ours	HT100M	V+A	11.73	**16.14**	30.68	26.11
Ours	HT100M	V+T	11.22	15.94	29.40	24.92
Ours	HT100M	V+A+T	**11.83**	**16.14**	30.93	26.38

(d) ActivityNet-Captions

Table 4. Comparison to SOTA. The top two results are in **bold** and underline. † use pretrained GPT-2 [39] or CLIP [38], so the results are not comparable. ‡ use relationships between sentences to generate paragraphs.

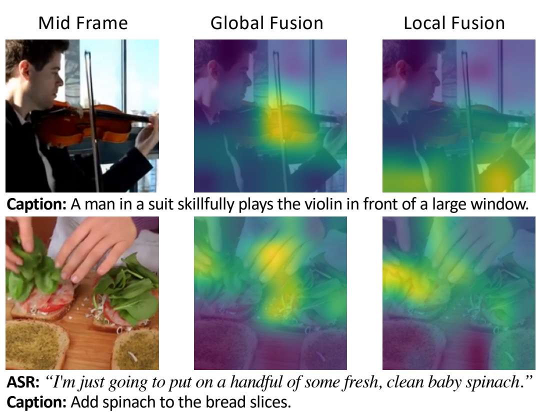


Figure 5. Attention maps from the audio modality to RGB space for global or local fusion. Top: VATEX. Bottom: YouCook2.

the captions are largely covered by speech on YouCook2. The results show that global cross fusion helps the flow of holistic information, *e.g.*, acoustic events, while local fusion helps fine-grained information, *e.g.*, single words in speech. Moreover, adding global cross fusion improves on both local fusion modules. Overall, local-global merged fusion performs best, showing the advantages of using both local and global tokens for cross-modal fusion. Since there are only a few global tokens, they do not bring much extra computational cost. Compared with merged fusion, local-global merged fusion increases FLOPs by only 6.96% (from 316G to 338G), but produces substantially better results.

Attention maps. Fig. 5 visualizes the attention maps for global and local fusion modules using Attention Rollout [2]. We visualize the attention maps summed over all the frames in the video clip as in [31]. We observe that global fusion focuses on salient regions related to acoustic events (piano), while local fusion attends to the key words in speech ("*put*" and "*spinach*").

Results with different modalities. Tab. 3 shows the results when we input different modalities. For audio-only and video-only methods, we zero-mask the other modality at the cross-modal encoder. For V+T setting, we use BERT [18] to replace the audio encoder. For V+A+T setting, we use global cross fusion between video and audio, and use merged fusion between video and text, to leverage the salient information in audio and text. On all four datasets, integrating audio (V+A) significantly improves compared to the video-only method. Besides, comparing V+A with V+T, audio consistently outperforms text. Particularly, on VATEX, compared with the video-only method, V+A improves CIDEr by 6.2%, while V+T only improves by 1.7%, showing that audio conveys more information than ASR texts. The comparisons between V+A+T and V+T also demonstrate the benefits of incorporating audio for video captioning. In addition, V+A+T further improves the CIDEr score by 4.6% on YouCook2 on the basis of the V+A method, probably because the ASR text contains more accurate semantic words than audio. On the other datasets where the speech is not dominant in the annotated captions, V+A+T performs very closely to V+A, showing that audio can cover the major information in speech while bypassing the time-consuming process of ASR [49]. In Fig. 1, we show two qualitative results on VATEX. In the first video, audio provides more information about laughter and crying, which the ASR text does not contain. In the second video, V+T mistakenly recognizes the little girl as a boy and misses the information that a man talks to her, while V+A corrects the error and adds the missing description by leveraging the information in audio.

Comparison with SOTA. In Tab. 4, we compare our methods with state-of-the-art (SOTA) methods on four datasets. On YouCook2, the V+T methods, *i.e*., M-MASS [14], Value [22], and UniVL [28], significantly outperform the video-only methods MART [36], ActBert [70], and SwinBert [25], demonstrating the importance of speech in enhancing results on this dataset. Our V+A method further improves the performance by directing learning from the audio. On MSRVTT, compared with two prior audio-visual video captioning works, MMCNN [51] and MGSA [7], our V+A method increases the CIDEr score by 6.9%, highlighting the benefits of pre-training. Notably, our V+A method also outperforms the SOTA video-only method SwinBert, which uses densely sampled video frames, and V+T methods, Decembert [50] and UniVL [28], both pre-trained on HowTo100M. On VATEX, our V+A method substantially outperforms the video-only method MGRMP [8] and the V+T method Value [22], and performs similarly to the SOTA SwinBert, excelling in two metrics. A line of recent works in video captioning like MV-GPT [43] and CLIP4VLA [41] leverage large-scale pre-trained text generative models, *e.g*. GPT-2 [39], or vision-text models, *e.g*. CLIP [38], which is not the focus of our work, so a fair comparison is infeasible. We mark their results in gray for reference, but we note that our model is still comparable or even better in some metrics. On ActivityNet-Captions, our method clearly outperforms VTransformer [69] and Transformer-XL [9]. Notably, without using any relationships between sentences to generate paragraphs, our method can achieve better performance than MART [36] and COOT [11], which exploit relationships across sentences [11, 36] during both training and inference. Please refer to the supplementary material for more qualitative results.

5. Discussions on Societal Impact

Video captioning makes videos more accessible to all users, including users with accessibility issues, *e.g*. lowvision and blind users. Furthermore, audio-visual video captioning also benefits users with hearing disability by including text descriptions of the audio modality. However, our framework is a data-driven system, so the quality of generated captions may be biased to the distribution of training data. As our pre-training data are obtained from online YouTube videos, our system may produce harmful captions that contain toxic contents or social biases present in the training data. To avoid undesirable effects, careful examination is required before adopting the outputs of our framework for real-world applications. Another ethical concern on using YouTube videos is how to protect user privacy. In our experiments, we download the videos currently available on YouTube. Therefore, if a user deletes a video from their YouTube channel, we will not be able to use the videos.

6. Discussions on Limitations

Our approach is not always successful in the fusion of visual and audio modalities. A potential direction is to design a more intelligent cross-modal fusion module to dynamically update the weights of different modalities for caption generation. Besides, the dominance of the modality is dataset-dependent, and we pre-trained our audio-visual video captioning model on the HowTo100M dataset, where audio dominates due to the use of ASR transcripts as text supervision. We did not verify the effectiveness of our proposed modality balancing pre-training strategy on vision-dominant datasets due to the lack of such large-scale pre-training datasets with video-text pairs. We believe that our proposed approach has the potential to be advantageous in those settings, and it could be explored in future research.

7. Conclusion

We present an end-to-end pre-training framework for audio-visual video captioning. A novel modality balanced pre-training loss is proposed to balance the learning of different modalities during pre-training, which demonstrates the effectiveness for audio-visual video captioning. We also comprehensively investigate different cross-modal fusion modules for audio-visual fusion and propose a new local-global fusion module. Our model can capture different types of information in human speech and background sounds, achieving comparable or even better results against the models using ASR text or complex language models.

References

[1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12487–12496, 2019. 1, 2

[2] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics. 7

[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. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 1, 2, 3

[4] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33:25–37, 2020. 2

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 5

[6] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 6

[7] Shaoxiang Chen and Yu-Gang Jiang. Motion guided spatial attention for video captioning. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8191–8198, 2019. 2, 8

[8] Shaoxiang Chen and Yu-Gang Jiang. Motion guided region message passing for video captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1543–1552, 2021. 8

[9] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019. 8

[10] 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*, 19(9):2045–2055, 2017. 1, 2

[11] Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *Advances in neural information processing systems*, 33:22605–22618, 2020. 8

[12] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. 3

[13] Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. A case study on combining asr and visual features for generating instructional video captions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 419–429, 2019. 2

[14] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 470–490, 2020. 1, 8

[15] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *British Machine Vision Conference (BMVC)*, 2020. 2

[16] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[17] 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

[18] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1, 7

[19] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2015. 6

[20] 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. 2

[21] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11336–11344, 2020. 3

[22] 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 *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021. 8

[23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[25] 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, 8

[26] 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. 3, 6

[27] 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. 3

[28] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 1, 2, 3, 5, 6, 8

[29] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 1, 2, 3

[30] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1, 2, 3, 5

[31] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 3, 7

[32] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Niebles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020. 1, 2

[33] Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. Adamml: Adaptive multi-modal learning for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7576–7585, 2021. 2

[34] 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

[35] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference on Learning Representations*, 2020. 2

[36] Wenjie Pei, Jiyuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8347–8356, 2019. 1, 2, 8

[37] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8238–8247, 2022. 2

[38] 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. 1, 3, 7, 8

[39] 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. 7, 8

[40] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8908–8917, 2019. 2

[41] Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, Sipeng Zheng, and Qin Jin. Accommodating audio modality in clip for multimodal processing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. 1, 2, 8

[42] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018. 1

[43] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 1, 2, 3, 8

[44] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16877–16887, 2021. 2, 3

[45] Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, Yu Qiao, and Yiran Zhong. Fine-grained audible video description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10585–10596, June 2023. 1

[46] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6382–6391, 2019. 1, 2, 5

[47] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019. 1, 2

[48] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5100–5111, 2019. 3

[49] Zineng Tang, Jaemin Cho, Yixin Nie, and Mohit Bansal. Tvlt: Textless vision-language transformer. *Advances in neural information processing systems*, 2022. 2, 8

[50] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426, 2021. 2, 8

[51] Yapeng Tian, Chenxiao Guan, Justin Goodman, Marc Moore, and Chenliang Xu. Audio-visual interpretable and controllable video captioning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition workshops*, 2019. 2, 8

[52] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019. 3

[53] 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. 3

[54] 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

[55] Deepali Verma, Arya Haldar, and Tanima Dutta. Leveraging weighted cross-graph attention for visual and semantic enhanced video captioning network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2465–2473, 2023. 2

[56] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018. 1, 2

[57] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 1, 3

[58] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 2, 4, 6

[59] 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. 1, 5

[60] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12186–12195, 2022. 3

[61] 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. 1, 5

[62] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 1, 3

[63] Chenggang Yan, Yunbin Tu, Xingzheng Wang, Yongbing Zhang, Xinhong Hao, Yongdong Zhang, and Qionghai Dai. Stat: Spatial-temporal attention mechanism for video captioning. *IEEE transactions on multimedia*, 22(1):229–241, 2019. 1, 2

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

[65] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 1, 2, 3, 6

[66] 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. 2, 3

[67] 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

[68] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1, 5

[69] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748, 2018. 5, 8

[70] 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. 8

[71] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 1