HiTeA: Hierarchical Temporal-Aware Video-Language Pre-training

Qinghao Ye Guohai Xu Ming Yan* Haiyang Xu*
Qi Qian Ji Zhang Fei Huang

DAMO Academy, Alibaba Group
yeqinghao.yqh@alibaba-inc.com

Abstract

Video-language pre-training has advanced the performance of various downstream video-language tasks. However, most previous methods directly inherit or adapt typical image-language pre-training paradigms to video-language pre-training, thus not fully exploiting the unique characteristic of video, i.e., temporal. In this paper, we propose a ***Hi****erarchical* ***Te****mporal-****A****ware video-language pre-training framework,* ***HiTeA****, with two novel pre-training tasks for yielding temporal-aware multi-modal representation with cross-modal fine-grained temporal moment information and temporal contextual relations between video-text multi-modal pairs. First, we propose a cross-modal moment exploration task to explore moments in videos by mining the paired texts, which results in detailed video moment representation. Then, based on the learned detailed moment representations, the inherent temporal contextual relations are captured by aligning video-text pairs as a whole in different time resolutions with multi-modal temporal relation exploration task. Furthermore, we introduce the shuffling test to evaluate the temporal reliance of datasets and video-language pre-training models. We achieve state-of-the-art results on 15 well-established video-language understanding and generation tasks, especially on temporal-oriented datasets (*e.g.*, SSv2-Template and SSv2-Label) with 8.6% and 11.1% improvement respectively.* ***HiTeA*** *also demonstrates strong generalization ability when directly transferred to downstream tasks in a zero-shot manner.*

1. Introduction

Vision and language are two primary signals that constitute the real-world perception of humanity. With the success of image-language pre-training [8,22,25,47], video-language pre-training [23,26,27,34] has recently received increasing attention. Large-scale video-language pre-training helps the model to learn effective multi-modal representation, which has shown significant improvement on a variety of video-language downstream tasks, such as video-text retrieval, video question answering and video captioning [5, 33, 44, 48, 50, 53, 59].

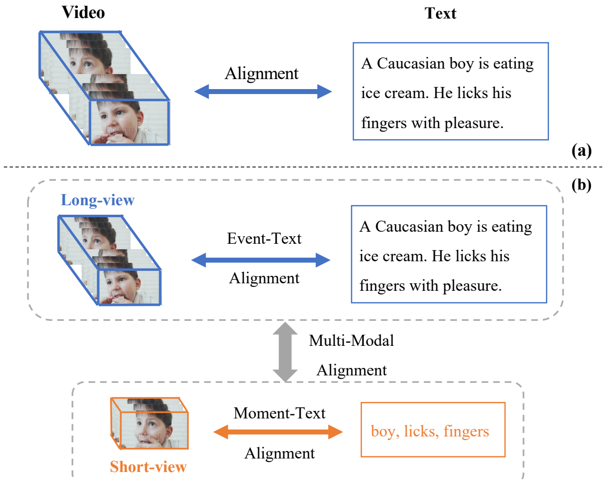
Video
Text
Alignment
A Caucasian boy is eating ice cream. He licks his fingers with pleasure.
(a)
(b)
Long-view
Event-Text
Alignment
A Caucasian boy is eating ice cream. He licks his fingers with pleasure.
Multi-Modal
Alignment
Moment-Text
Alignment
boy, licks, fingers
Short-view

Figure 1: Comparison between existing paradigms and ours for video-language pre-training. (a) Previous methods align video and text within global perspective as the pretext. (b) We introduce **HiTeA** by varying video in different temporal views and modeling cross-modal temporal information between moments and texts, as well as the temporal contextual relations between multi-modal pairs.

Inspired by the success of image-language pre-training paradigm, various methods [10,11,23,26,27] have been proposed to adapt it to video-language pre-training. ClipBERT [21] and Singularity [20] directly build on representations from image encoders and aggregate them via score aggregation function and temporal encoder. Furthermore, MIL-NCE [34] and Frozen [2] switch image encoder to video encoder for spatio-temporal video representation learning and align the video with corresponding text. In addition, some advanced pre-training tasks are designed through modeling entity [23], reconstructing masked patches [10] and predicting frame order [26, 61]. Despite their promising performance on downstream tasks, they treat video within global perspective illustrated in Figure 1(a), thus failing to consider fine-grained temporal information and temporal contextual relations which are essential to video-language pre-training.

Since untrimmed video contains various temporal details,

*Corresponding Author.

directly treating the video globally has two main limitations: (1) Less effective in modeling the fine-grained moment information including atomic actions and moments. As illustrated in Figure 1(b), we vary time resolutions and generate two views (long & short) for the input video. As a result, the short-view video clip tends to represent the moment information and the long-view video may express more event-level information. For example, the short-view video clip in Figure 1(b) only describes the moment of "lick fingers" rather than "eating ice cream". Such fine-grained moment information is hard to be captured by the long-view video under global event perspective; (2) Ignoring the temporal contextual relations implicitly existed in the video. Knowing the event expressed by the text, the moment "eating ice cream" can be inferred from the moment "lick fingers" shown by short-view video. However, such implicit temporal contextual relations between the moment and the event are rarely explored in previous works.

To address these problems, we propose a **Hi**erarchical **Te**mporal-**A**ware video-language pre-training framework, **HiTeA**, for both multi-modal understanding and generation. Except for the standard pre-training tasks, **HiTeA** introduces two novel temporal-aware video-language pre-training tasks, named *cross-modal moment exploration* (CME) and *multi-modal temporal relation exploration* (MTRE), which not only model the fine-grained temporal moment information but also captures temporal contextual information hierarchically, yielding temporal-aware multi-modal representations for both understanding and generation. Specifically, we first generate the long-view and short-view videos with different time resolutions to build hierarchy of the input video. Then, based on the similarities of words and short-view video, we select the most relevant words as positive and leave the rest of the words as hard negatives. The CME pre-training task is applied to align the positive words and short-view video representations in the same embedding space. Moreover, to capture association between moments and the event for temporal contextual modeling, we match different views for the same video. However, directly matching two views visually would be noisy due to the background similarity [39]. To this end, we perform multi-modal pair alignment between video-text pairs via the MTRE pre-training task. More specifically, the short-view video guided by most relevant words and the long-view video guided by text will be aligned, which enables the model to extrapolate the contextual information from the short-view with language signal while enhancing temporal reasoning ability. Empowered by above two novel temporal-aware video-language pre-training tasks, **HiTeA** is capable of modeling temporal-aware multi-modal information revealed in video-text data including both fine-grained moment information and temporal contextual relations.

In spite of a good performance, recent studies [4, 20] reveal most video-language downstream datasets are biased towards still objects, scenes, *etc*., while the temporal dynamics are negligible. To evaluate the temporal performance of the video-language pre-training model and temporal reliance of downstream datasets, we introduce temporal shuffling test for these datasets. This enables a comprehensive evaluation of temporal modeling capability in the video-language pre-training field. Besides, our method achieves significant improvement on the datasets with heavy temporal reliance.

In summary, our key contributions are the followings:

- We propose a novel hierarchical temporal-aware video-language pre-training framework with both video-language understanding and generation capabilities.
- We introduce temporal-aware pre-training tasks to generate temporal-aware multi-modal representation through modeling fine-grained temporal moment information as well as capturing the temporal contextual relations between moment and event.
- Extensive experiments demonstrate the effectiveness and generalization ability of **HiTeA**, and it achieves state-of-the-art performance on 15 video-language downstream datasets including video-text retrieval, video question answering, and video captioning, especially on temporal-oriented datasets (*e.g*., SSv2-Template and SSv2-Label) with 8.6% and 11.1% improvement respectively.

2. Related Work

Video-Language Pre-training Benefiting from a large number of image/video-text pairs, video-language pre-training (VLP) exhibits superior capabilities on various video-text benchmarks. The method of VLP is constantly evolving. Traditional approaches [26, 30, 42, 62] leverage offline-extracted dense video features for pre-training to circumvent the expensive computation overhead. In contrast, ClipBERT [21] suggests that sparse sampling can enable affordable end-to-end learning and improve performance simultaneously. Recent emerging approaches [2, 11, 14, 21, 23, 27, 52, 57] adopt this strategy and propose new model architectures and pre-training tasks. Frozen [2] trains jointly on image and video datasets via video-text contrastive learning (VTC). ALPRO [23] proposes a new visually-grounded pre-training task combined with VTC, video-text matching (VTM) and masked language modeling (MLM) [9] to learn fine-grained region-entity alignment. LAVENDER [27] formulates all pre-training and downstream tasks as MLM so that a unified architecture can used for all video-text tasks. Apart from above representative works, frame order modeling (FOM) [26, 61] and masked video modeling (MVM) [10] are designed for VLP. However, the temporal characteristic of video still remains largely unexplored. To this end, we introduce a novel hierar-

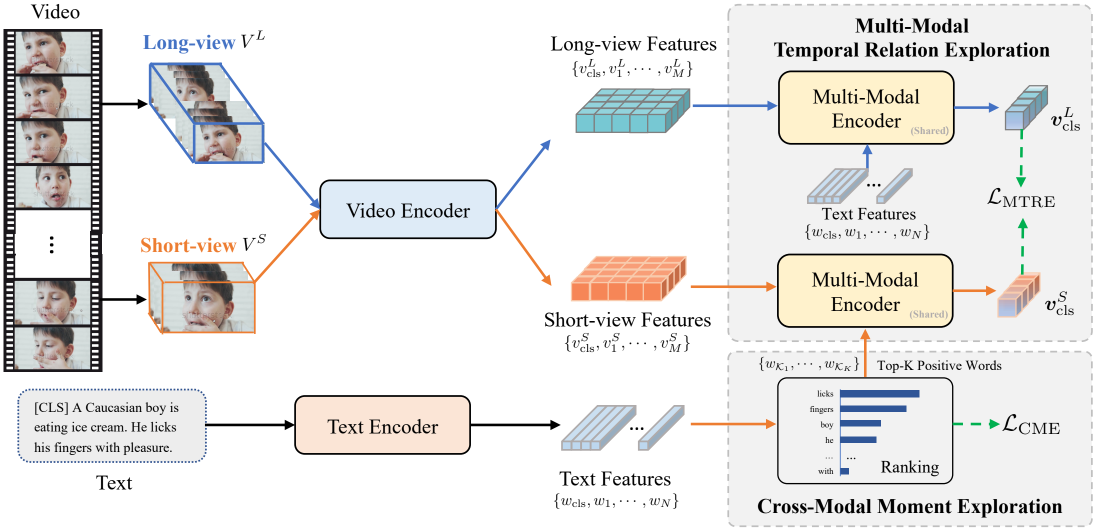
Video
Long-view V^L
Short-view V^S
Video Encoder
[CLS] A Caucasian boy is eating ice cream. He licks his fingers with pleasure.
Text
Text Encoder
Long-view Features
$\{v^L_{\text{cls}}, v^L_1, \cdots, v^L_M\}$
Short-view Features
$\{v^S_{\text{cls}}, v^S_1, \cdots, v^S_M\}$
Text Features
$\{w_{\text{cls}}, w_1, \cdots, w_N\}$
Multi-Modal Temporal Relation Exploration
Multi-Modal Encoder (Shared)
$\boldsymbol{v}^L_{\text{cls}}$
Text Features
$\{w_{\text{cls}}, w_1, \cdots, w_N\}$
$\mathcal{L}_{\text{MTRE}}$
Multi-Modal Encoder (Shared)
$\boldsymbol{v}^S_{\text{cls}}$
$\{w_{\mathcal{K}_1}, \cdots, w_{\mathcal{K}_K}\}$
Top-K Positive Words
licks
fingers
boy
he
...
with
Ranking
$\mathcal{L}_{\text{CME}}$
Cross-Modal Moment Exploration

Figure 2: Illustration of the proposed **HiTeA**. We first generate two different temporal views for the input video, where the long-view is the video itself and the short-view is randomly truncated from the input video. To explore the moment revealed in the short-view, *cross-modal moment exploration* (CME) selects the candidate words from the input text with $\mathcal{L}_{\text{CME}}$. Then, we perform *multi-modal temporal relation exploration* (MTRE) for modeling the temporal contextual relations between two video-text pairs with different views by $\mathcal{L}_{\text{MTRE}}$. Note that the multi-modal encoders and the text features are shared.

chical temporal-aware VLP framework which not only models the fine-grained moment information but also captures contextual relations with different temporal granularities.

Temporal Modeling The temporal characteristic acts as a vital role in VLP since it provides the model with the capabilities of reasoning and understanding causality. Previous efforts in this field can be roughly divided into three categories. First, several methods directly transfer image-text models to video-text tasks by simply concatenating video frame [22,24] or building an extra temporal encoder [31,32]. Second, some works [2, 10, 23, 27] switch the image encoder to video encoder for learning spatio-temporal contexts within videos. Third, HERO [26] and MERLOT [61] design FOM task to explicitly recover the correct temporal order of shuffled frames. Nonetheless, ATP [4] and Singularity [20] reveal the existence of a static appearance bias in popular video-language datasets, and they develop single-frame models to achieve strong performance, comparable or even better than above methods with explicit temporal modeling. Therefore, they recommend SSv2 [20] and NExT-QA [48] datasets to test the temporal ability of VLP models. Different from previous approaches, we vary the temporal resolutions and generate two views of video so as to construct the temporal hierarchy, which equips the model with the ability to learn both fine-grained temporal moment information and temporal contextual representation at the same time.

3. Method

3.1. Overview

Figure 2 sketches the overview of the **HiTeA**. In concrete, our model consists of two unimodal encoders for encoding video and text separately, a multi-modal encoder for video and text interaction, and a text decoder for generation which is omitted here for simplicity and detailed in Appendix.

For video representation, previous methods [21, 23, 27] encode the whole input video as a single-view feature, ignoring the rich temporal details contained in the video. Thus, we first treat the video into two views with different time resolutions to build hierarchy of the input video. Specially, the untrimmed video is regarded as a long-view video V^L for capturing event information, and a video segment is randomly truncated from the input video as the short-view for capturing moment information denoted as V^S. Then, we use the video encoder to encode an arbitrary view of video $V \in \mathbb{R}^{T\times H\times W}$ into a sequence of embeddings: $\mathcal{V} = \{v_{\text{cls}}, v_1, \cdots, v_M\} \in \mathbb{R}^{M\times D}$, where M is the number of flattened patches, D is hidden size, and v_{cls} is the embedding of the visual [CLS] token which provides global representation of the video. For text representation, we use the text encoder to transform the text T into a sequence of embeddings: $\mathcal{T} = \{w_{\text{cls}}, w_1, \cdots, w_N\} \in \mathbb{R}^{N\times D}$, where N is length of the text. After that, the multi-modal encoder takes video features $\mathcal{V}$ and text features $\mathcal{T}$ as inputs and yields the multi-modal representation $\boldsymbol{v}_{\text{cls}}$ for the video.

In order to take full advantage of the different views of the video, we introduce *cross-modal moment exploration* (CEM) to explore the proper words or phrases from input text to align the short-view video with $\mathcal{L}_{\text{CME}}$ for capturing the fine-grained temporal moment information in Section 3.2. Furthermore, to model the temporal contextual relations between the short-view video containing moment information and the long-view video with event information, we propose *multi-modal temporal relation exploration* (MTRE) to match the multi-modal representation of short-view and long-view videos by $\mathcal{L}_{\text{MTRE}}$ in Section 3.3. Lastly, we introduce the overall pre-training objective for training the model in Section 3.4.

3.2. Cross-Modal Moment Exploration

The video with short temporal range (*i.e.*, short-view of video) with the paired text tends to be accompanied with fine-grained temporal moment information. However, the paired text partially describes the short-view of video bringing noise to the fine-grained moment representation thus degrading the performance. To this end, we propose a novel pre-training task named *cross-modal moment exploration* (CME), which enables the model to understand fine-grained moment information by leveraging the partially aligned text.

Formally, we first discover the possible positive words for the video in short-view by computing the cosine similarity of the word embedding sequence $\{w_1, \cdots, w_N\}$ from text encoder and the short-view video representation v^S_{cls} from video encoder as:

$$\mathcal{K} = \{\pi(1), \cdots, \pi(K)\}, \tag{1}$$

where $\pi : \{1, \cdots, N\} \rightarrow \{1, \cdots, N\}$ is a permutation function for ranking such that $s(w_{\pi(1)}, v^S_{\text{cls}}) \geq \cdots \geq s(w_{\pi(N)}, v^S_{\text{cls}})$, and $\mathcal{K}$ is the set of selected word indices, K is the number of possible selected words, and $s(x, y) = x^T y / \|x\|_2 \|y\|_2$ represents the cosine similarity between x and y. After obtaining the words for the video in short-view as the positive pair, the cross-modal moment exploration loss $\mathcal{L}_{\text{CME}}$ is computed with negative pairs from other words in the input text, which is defined as:

$$\mathcal{L}_{\text{CME}} = -\frac{1}{B}\sum_{i=1}^{B}\left(\frac{1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}\log\frac{\exp((v^S_{\text{cls}})_i^\top w_{i,k}/\tau)}{\sum_{n=1}^{N}\exp((v^S_{\text{cls}})_i^\top w_{i,n}/\tau)}\right), \tag{2}$$

where τ is the learnable temperature hyper-parameter that controls the sharpness of the output distribution, and it is initialized as 0.07. As a consequence, the model is able to understand moment information via the proposed cross-modal exploration scheme.

3.3. Multi-Modal Temporal Relation Exploration

While the video encoder has demonstrated its effectiveness in learning temporal representation implicitly [10, 23, 56], it remains a challenge to discover the inherent temporal contextual relations. As a result, the limited capabilities in temporal modeling deteriorate the downstream task in temporal reasoning. This is in particular a missing point for the existing video-language pre-training paradigm [11,21,23,27], which usually focuses on bridging video and text neglecting the function of text for guiding the video temporal contextual representation learning thus losing the temporal cues.

To this end, we introduce *multi-modal temporal relation exploration* (MTRE), a novel temporal-aware pre-training task that improves models' capacities in capturing temporal context in video with fine-grained text guidance by aligning multi-modal pairs. Specially, the short-view video V^S would represent moment information with respect to the whole video. On the contrary, the long-view video V^L expresses the event and topical information. To obtain the text-guided video features, we feed videos in different temporal views into the video encoder individually. Then, the text features are extracted and interact with the video features by the multi-modal encoder and yield text-guided video representations $\boldsymbol{v}^L_{\text{cls}} \in \mathbb{R}^D$ and $\boldsymbol{v}^S_{\text{cls}} \in \mathbb{R}^D$ as follows:

$$\boldsymbol{v}^L_{\text{cls}} = f(\{v^L_{\text{cls}}, v^L_1, \cdots, v^L_M\}, \{w_{\text{cls}}, w_1, \cdots, w_N\}), \tag{3}$$

$$\boldsymbol{v}^S_{\text{cls}} = f(\{v^S_{\text{cls}}, v^S_1, \cdots, v^S_M\}, \{w_{\text{cls}}, w_1, \cdots, w_N\}), \tag{4}$$

where $f(\mathcal{V}, \mathcal{T})$ represents the multi-modal encoder with video features $\mathcal{V}$ and text features $\mathcal{T}$. However, since the short-view of the video is partially aligned with the text, using the whole text is not reasonable for generating accurate text-guided video feature for short-view. Meanwhile, improper video-text pairs would yield noisy multi-modal representation thus degrading the performance of the model. Therefore, thanks to the positive words mined by *cross-modal moment exploration*, we can calibrate representation for short-view video by:

$$\boldsymbol{v}^S_{\text{cls}} = f(\{v^S_{\text{cls}}, v^S_1, \cdots, v^S_M\}, \{w_{\mathcal{K}_1}, \cdots, w_{\mathcal{K}_K}\}), \tag{5}$$

where $\mathcal{K}_i \in \mathcal{K}$ is the index of the set for selected positive words. Then, we aim to match the representation of produced text-guided video features in different granularities in order to enable the model to predict the past and the future from the short-view of video, which benefits for capturing the general structure of the video. Specifically, we adopt the SimSiam framework [6] for minimizing their negative cosine similarity:

$$\mathcal{D}(p^S, z^L) = -\frac{p^S}{\|p^S\|_2} \cdot \frac{z^L}{\|z^L\|_2}, \tag{6}$$

where $p^S = h(g(\boldsymbol{v}^S_{\text{cls}}))$ and $z^L = g(\boldsymbol{v}^L_{\text{cls}})$. The g and h are projection MLP head and prediction MLP head [7,13]. Minimizing $\mathcal{D}(p^S, z^L)$ is equivalent for minimizing the mean

Method	# PT Data	MSRVTT			DiDeMo			LSMDC			ActivityNet Caption		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ClipBERT [21]	0.2M	22.0	46.8	59.9	20.4	48.0	60.8	-	-	-	21.3	49.0	63.5
Frozen [2]	5M	31.0	59.5	70.5	31.0	59.8	72.4	15.0	30.8	39.8	-	-	-
ALPRO [23]	5M	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-	-	-	-
BridgeFormer [11]	5M	37.6	64.8	75.1	37.0	62.2	73.9	17.9	35.4	44.5	-	-	-
Singularity [20]	5M	36.8	65.9	75.5	47.4	75.2	84.0	-	-	-	43.0	70.6	81.3
LAVENDER [27]	5M	37.8	63.8	75.0	47.4	74.7	82.4	22.2	43.8	53.5	-	-	-
Models pre-trained on more data													
VIOLET [10]	183M	34.5	63.0	73.4	32.6	62.8	74.7	16.1	36.6	41.2	-	-	-
All-in-one [46]	138M	37.9	68.1	77.1	32.7	61.4	73.5	-	-	-	22.4	53.7	67.7
Clip4Clip [31]	400M	42.1	71.9	81.4	43.4	70.2	80.6	21.6	41.8	49.8	40.5	72.4	-
X-CLIP [32]	400M	46.1	73.0	83.1	45.2	74.0	-	23.3	43.0	-	44.3	74.1	-
HiTeA	5M	**44.4**	**69.3**	**78.9**	**51.8**	**79.1**	**85.3**	**27.1**	**46.2**	**54.5**	**45.1**	**73.5**	**84.2**

Table 1: **Performance comparison on text-to-video retrieval.** All results are reported on R@1/R@5/R@10. We gray out methods that use significantly more pre-training data for fair comparison. **# PT Data** is the number of video-text pairs for pre-training.

square error between p^S and z^L, which encourages the videos in different temporal magnitudes to be similar. Following [6, 13], we defined a symmetrized loss as:

$$\mathcal{L}_{\text{MTRE}} = \frac{1}{2}\left[\mathcal{D}(p^L, \texttt{sg}(z^S)) + \mathcal{D}(p^S, \texttt{sg}(z^L))\right], \quad (7)$$

where $\texttt{sg}(\cdot)$ is the stop-gradient operation that prevents the model from collapse during training [6].

3.4. Pre-training Objectives

Apart from the two proposed temporal-aware pre-training tasks, we follow proven video-text pre-training approaches [2, 23, 27] to adopt the standard pre-training tasks including video-text contrastive (VTC), video-text matching (VTM), masked language modeling (MLM), and prefix language modeling (PrefixLM) described in the related work. Precisely, VTC and VTM align the video and text from the global perspective, while MLM and PrefixLM contribute to multi-modal understanding and generation capabilities of the model. Details of these objectives are described in the Appendix. We simply combine these as the base training objective $\mathcal{L}_{base}$ for our model. Therefore, the full pre-training objective is computed as:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{CME}} + \mathcal{L}_{\text{MTRE}}. \quad (8)$$

4. Experiments

4.1. Experiment Setup

Pre-training Datasets Following the recent work [2, 11, 20, 23, 27], we pre-train our model on a webly-sourced video dataset WebVid-2M [2] with 2.5M video-text pairs and a image-text dataset Google Conceptual Captions (CC3M) [41] with 3M image-text pairs. Unlike previous methods, we do not pre-train our model on the large-scale video-text datasets like HowTo100M [34] with 136M video-text pairs and YT-Temporal-180M [61] due to the heavy computation.

Downstream Datasets We evaluate our pre-trained model on 18 video-language benchmarks including video-text retrieval, video question answering, and video captioning tasks. Specifically, video question answering (VideoQA) can be categorized as Multiple-Choice (MC) and Open-Ended (OE) settings. The evaluation datasets are briefly summarized in below. Details can be found in the Appendix.

- **Video-Text Retrieval**: MSRVTT [53], DiDeMo [1], LSMDC [38], ActivityNet Caption [18], SSv2-Label [20], and SSv2-Template [20];
- **VideoQA (MC)**: TGIF-Action, TGIF-Transition [15], MSRVTT-MC [58], LSMDC-MC [44], and NExT-QA [48];
- **VideoQA (OE)**: TGIF-Frame [15], MSRVTT-QA, MSVD-QA [50], LSMDC-FiB [33] and ActivityNet-QA [59].
- **Video Captioning**: MSRVTT [53] and MSVD [5].

Implementation Details Our implementation of **HiTeA** is based on PyTorch [35]. In detail, we instantiate the video encoder with MViT-Base model [28] pretrained on ImageNet-21K [37]. The text encoder is initialized from first six layers of pre-trained BERT-Base [9], and the multi-modal encoder is initialized with last six layers of pre-trained BERT-Base. We pre-train **HiTeA** for 10 epochs, using a batch size of 16 on 8 NVIDIA A100 GPUs. We use AdamW [17] optimizer with a weight decay of 0.02 and betas (0.9, 0.98). The learning rate is first warmed up to 5e-5 in the first 1000 iterations, and decays following a cosine schedule. During pre-training stage, following with [10, 14, 23], we sparsely sample 4 frames for short and long view while preserving their order in-between and resize them to 224 × 224. The duration of short view is restricted as the 1/8 of the whole video duration. K is empirically set to 5. The MLM mask ratio is set to 15%. Details of fine-tuning stage are described in Appendix.

		TGIF			MSRVTT		LSMDC		MSVD	ActivityNet
Method	#PT Data	Action	Transition	Frame	MC	QA	MC	FiB	QA	QA
ClipBERT [21]	0.2M	82.8	87.8	60.3	88.2	37.4	-	-	-	-
ALPRO [23]	5M	-	-	-	-	42.1	-	-	46.3	-
Singularity [20]	5M	-	-	-	92.0	42.7	-	-	-	41.8
LAVENDER [27]	5M	96.6	**99.1**	72.2	96.6	44.2	**86.0**	**56.9**	55.4	-
Clover [14]	5M	94.9	98.0	71.4	95.0	43.9	83.2	54.1	51.9	-
Models pre-trained on more data										
VIOLET [10]	183M	92.5	95.7	68.9	91.9	43.9	82.8	53.7	47.9	38.9
JustAsk [54]	69M	-	-	-	-	41.5	-	-	46.3	38.9
MERLOT [61]	180M	94.0	96.2	69.5	90.9	43.1	81.7	52.9	-	41.4
All-in-one [46]	283M	95.5	94.7	66.3	92.3	46.8	84.4	-	48.3	-
HiTeA	5M	**96.8**	98.8	**72.5**	**97.2**	**45.4**	85.8	54.6	**55.6**	**45.1**

Table 2: **Performance comparison on video question answering.** Accuracy is reported for evaluation. We gray out methods that use significantly more pre-training data for fair comparison.

Method	# PT Data	MSRVTT	MSVD
UniVL [30]	180M	49.9	-
SwinBERT [29]	-	53.8	120.6
MV-GPT [40]	53M	60.0	-
CLIP4Caption [43]	400M	57.7	-
LAVENDER [27]	5M	58.0	142.9
HiTeA	5M	**62.5**	**145.1**

Table 3: **Performance comparison on video captioning.** CIDEr [45] is reported for evaluation.

Method	MSRVTT-Ret.	LSMDC-Ret.	MSRVTT-QA	MSVD-QA
Frozen [2]	31.0/59.5/70.5	15.0/30.8/39.8	-	-
ALPRO [23]	33.9/60.7/73.2	-/-/-	42.1	46.3
BridgeFormer [11]	37.6/64.8/75.1	17.9/35.4/44.5	-	-
Singularity [20]	36.8/65.9/75.5	-/-/-	42.7	-
TimeSformer ($\mathcal{L}_{\text{base}}$)	37.6/62.0/72.3	19.2/38.7/48.0	41.5	49.0
HiTeA with TimeSformer	**39.7/65.0/75.1**	**21.8/40.7/49.9**	**43.7**	**52.4**

Table 4: Performance comparison of different SoTA methods with TimeSformer [3]. For text-to-video retrieval, Recall@1/5/10 are reported. For video question answering task, we report the Top-1 accuracy.

4.2. Comparison to Prior Arts

In this section, we compare **HiTeA** with numerous state-of-the-art video-language pre-training methods on several downstream datasets under fine-tuning setting.

4.2.1 Text-to-Video Retrieval

Table 1 summarizes the results on MSRVTT [53], DiDeMo [1], LSMDC [38], and ActivityNet Caption [18] under fine-tuning settings. Our method outperforms all of the existing video-language pre-training model by a large margin under the same data scale. In particular, our method yields 6.6% lift in terms of R@1 on MSRVTT dataset while only exploiting 5M video-text pairs. Note that we also include the comparison with the recent works that utilize the powerful encoder from CLIP [36], our method still can be comparable with them even surpass them, which shows the validness of the proposed method. Besides, we can notice that our method achieves the best result among all of listed methods on LSMDC dataset, which proves that our model can leverage the various moments presented in fruitful movie clips with cross-modal moment exploration.

4.2.2 Video Question Answering

Table 2 lists the results of **HiTeA** and current state-of-the-art approaches on nine VideoQA datasets. It can be noticed that our method achieves the best performance in most of VideoQA datasets even with less pre-training data. Specifically, it achieves absolute improvement 1.1% on TGIF-FrameQA, 2.2% on MSRVTT-MC, 1.5% on MSRVTT-QA, 0.2% on MSVD-QA, and 3.3% on ActivityNet-QA. We believe the moments learned by the cross-modal exploration are useful for finding the clue of answers in VideoQA.

4.2.3 Video Captioning

Table 3 compares **HiTeA** with existing methods on video captioning datasets MSRVTT and MSVD. As shown in the table, although we use less pre-training data than compared approaches, **HiTeA** still obtains significant improvement compared to those large-scale pre-trained models. On MSRVTT Caption, our method surpasses SoTA method MV-GPT [40] by 2.5% CIDEr. Note that MV-GPT is pre-trained for multi-modal video captioning and it leverages the ASR transcripts from audio as the additional input. By contrast, our method only utilizes video as the input during generation.

4.3. Discussion

In this section, we discuss the temporal characteristics of our model and the datasets.

Generalization on Plain Backbone. Table 4 delivers the performance of SoTA methods with the plain visual backbone. We instantiate the video encoder with TimeSformer [3] pretrained on ImageNet-21K [37] following [2, 11, 23]. We can observe that **HiTeA** with TimeSformer significantly

Method	MSRVTT-Retrieval				SSv2 Template-Retrieval				NExT-QA (Hard)		MSVD-QA
	R@1	R@5	R@10	AveR	R@1	R@5	R@10	AveR	Acc@C	Acc@T	Acc.
$\mathcal{L}_{\text{base}}$	40.0	68.0	77.1	61.7	80.5	100.0	100.0	93.5	44.0	46.4	52.7
$\mathcal{L}_{\text{base}} + \mathcal{L}_{\text{MTRE}}$	41.6	69.1	78.2	63.0	83.3	98.9	100.0	94.1	46.3	46.4	54.8
$\mathcal{L}_{\text{base}} + \mathcal{L}_{\text{CME}}$	42.0	69.3	**79.7**	63.7	83.9	99.4	100.0	94.4	46.3	48.3	54.3
$\mathcal{L}_{\text{base}} + \mathcal{L}_{\text{CME}} + \mathcal{L}_{\text{MTRE}}$	**44.4**	**69.3**	78.9	**64.2**	**85.6**	**100.0**	**100.0**	**95.2**	**47.8**	**48.6**	**55.6**

Table 5: Evaluation of the proposed methods on four downstream video-language tasks. For text-to-video retrieval, R@1, R@5, R@10, and the average are reported. For video question answering, we report the accuracy.

Method	# PT Data	SSv2-Label			SSv2-Template		
		R@1	R@5	R@10	R@1	R@5	R@10
Frozen [2]	5M	-	-	-	52.9	94.8	99.4
Clip4Clip [31]	400M	43.1	71.4	80.7	77.0	96.6	98.3
Singularity [20]	5M	44.1	73.5	82.2	77.0	98.9	99.4
HiTeA	5M	**55.2**	**81.4**	**89.1**	**85.6**	**100.0**	**100.0**

Table 6: Comparison of existing methods on Something-to-Something (SSv2) text-to-video retrieval.

Method	# PT Data	Acc@C	Acc@T	Acc@D	Acc.
Full Set					
Human	-	87.6	88.6	90.4	88.4
HCRN [19]	-	45.9	49.3	53.7	48.2
HGA [16]	-	46.3	50.7	59.3	49.7
VGT [49]	0.18M	53.4	56.4	69.5	56.9
HGA* [16]	400M	46.8	52.1	59.3	50.4
ATP [4]	400M	51.3	50.2	66.8	54.3
HiTeA	5M	**62.4**	**58.3**	**75.6**	**63.1**
Hard Split					
ATP [4]	400M	38.4	36.5	/	/
HGA [16]	-	43.3	45.3	/	/
HiTeA	5M	**47.8**	**48.6**	**/**	**/**

Table 7: Comparison of existing methods on NExT-QA [48]. We report accuracy on the Causal (C), Temporal (T), Descriptive (D) splits and overall accuracy on validation set. * stands for using CLIP as the initialization of visual encoder.

Dataset	Original ↑	Shuffled ↓	Gap ↑
MSRVTT [53]	64.2	63.3	0.9
DiDeMo [1]	72.1	70.2	1.9
LSMDC [38]	42.6	41.7	0.9
ActivityNet Caption [18]	67.6	66.8	0.8
SSv2 Template [20]	95.2	72.4	22.8
SSv2 Label [20]	76.7	73.5	3.2

Table 8: Dependency on temporal information for text-to-video retrieval datasets with temporal shuffling test. The average recall of Recall@1, Recall@5, and Recall@10 are reported. We evaluate the performance drop when shuffling the input during inference. “Original” and “Shuffled” denote the original and shuffled input videos, respectively, and “Gap” is the difference between the Original and Shuffled metric. The larger "Gap" indicates the dataset relies on temporal information, and the model utilizes more temporal information to solve the task.

Dataset	Original ↑	Shuffled ↓	Gap ↑
MSRVTT-QA [50]	45.4	45.2	0.2
MSVD-QA [50]	55.6	55.5	0.1
TGIF-FrameQA [15]	72.5	72.1	0.4
ActivityNet-QA [59]	45.1	45.0	0.1
NExT-QA (Hard) [48]	47.1	45.6	0.5

Table 9: Dependency on temporal information for video question answering datasets by temporal shuffling test. We report the accuarcy for each dataset. For NExT-QA dataset, we evaluate with the hard split of the validation set [4].

outperforms other SoTA methods (*e.g.*, ALPRO, BridgeFormer, and Singularity) with the same pre-training data and the number of frames. In addition, it shows that our proposed temporal-aware pre-training tasks can boost the performance of TimeSformer ($\mathcal{L}_{\text{base}}$) with 2.6% average improvement on MSRVTT-Retrieval and 3.4% on MSVD-QA. More related discussions are demonstrated in the Appendix.

Impact of Loss Terms. We investigate the contribution of individual loss terms and the results are shown in Table 5. It can be observed that the combining both $\mathcal{L}_{\text{CME}}$ and $\mathcal{L}_{\text{MTRE}}$ improves the performance of text-to-video retrieval and video question answering by at least 1.7% and 2.9% in Average Recall and Average accuracy respectively. In addition, we also find that the performance of $\mathcal{L}_{\text{CME}}$ surpasses that of $\mathcal{L}_{\text{MTRE}}$ on MSRVTT retrieval dataset that largely dominated by the appearance information. This can be explained that the cross-modal moment exploration loss not only select the positive verbs for the video from the text but also choose the acting object for alignment, which can boost the retrieval performance.

Evaluation on Temporal-aware Tasks. Lei *et al.* [20] reveal that the previous four retrieval datasets are prone to being biased for appearance while rarely relying on temporal information, thus introducing Something-to-Something v2 (SSv2) Template and SSv2 Label retrieval datasets to test models' true temporal modeling capability. In particular, SSv2 Template retrieval task requires a deeper understanding of the moment and temporal relation since no objects information are presented. The performance on these datasets are summarized in Table 6. It can be observed that **HiTeA** achieves significant improvement with +8.5% gains in terms of R@1 on these two temporal-oriented text-to-video retrieval datasets, which demonstrates the effectiveness of our proposed method through exploring fine-grained moment information and modeling temporal relation. In addition, we evaluate our model on NExT-QA [48] dataset that explicitly designed for temporal and causal understanding.

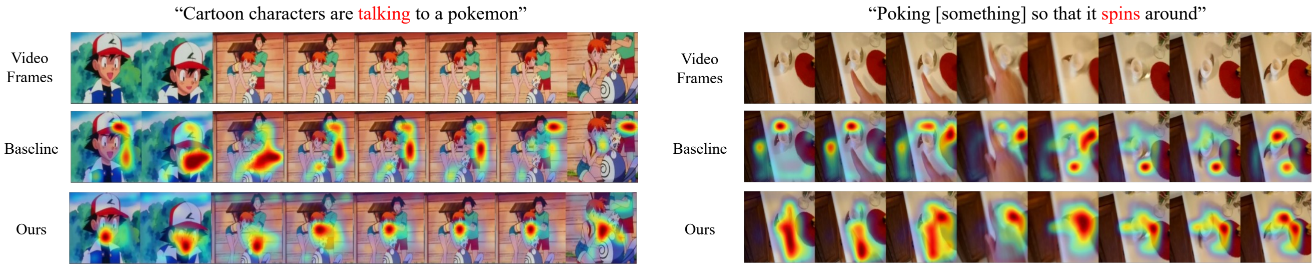
“Cartoon characters are talking to a pokemon”
Video Frames
Baseline
Ours
“Poking [something] so that it spins around”
Video Frames
Baseline
Ours

Figure 3: Visualizations of learned cross-attention maps from multi-modal encoder. We present samples from MSRVTT [53] and SSv2 Template [20] retrieval dataset. **HiTeA** attends to the patches related to objects motion by tracking trajectory.

As presented in Table 7, our method significantly surpasses its competitive counterparts, even those methods equipped with powerful image-text pre-trained encoders. Quantitatively, **HiTeA** obtains an absolute improvement +9% on the causality split with the help of intrinsic temporal relation. Recently, Buch *et al*. [4] filter out the trivial question for the dataset, and build the hard split for causality and temporal related questions for evaluate the causality and temporal of the model. As we can see in the table, even for the questions that heavily rely on causality, our model can still achieves a relative gain of 4.1% on the model with specific design for VideoQA, which indicates that our model do not solely depend on static appearance.

Temporal Reliance of Datasets. Previous methods [11,21, 23] only evaluate the performance of models on the existing datasets to demonstrate the superiority of the methods. However, Buch *et al*. [4] and Lei *et al*. [20] reveal that the most of the evaluation are biased towards the static concepts. Here, we investigate the temporal reliance for the evaluated datasets by introducing the temporal shuffling test. Specifically, we compute the performance changes between running inference on the ordered video versus its shuffled version. The large performance drop indicates the dataset has less spatial bias and needs for temporal information. Table 8 and Table 9 conclude the performance gap between ordered and shuffled input video for text-to-video retrieval and VideoQA datasets. For text-to-video retrieval task, SSv2 Template shows the large performance drop after shuffling the input video, which demonstrates that it depends most on the dynamic information thus verifying our assumption. On the contrary, the performance on ActivityNet Caption dataset is barely affected (-0.8 on Mean Recall) since the text almost describes the static objects without relying on temporal information. For video question answering dataset, we observe that the MSVD-QA and ActivityNet-QA are less sensitive to the order of video frames. This is because these two datasets contain more questions requiring frame-region information, such as object categories, scenes, and species. We believe this can be used to evaluate the temporal reliance of the datasets as well as the utilization for temporal cue by models in the future work.

Method	# PT Data	MSRVTT-QA	MSVD-QA
Just Ask [54]	69M	2.9	7.5
LAVENDER [27]	5M	4.5	11.6
MERLOT Reserve [60]	1B	5.8	-
FrozenBiLM [55]	10M	6.4	11.7
HiTeA	5M	**8.6**	**18.2**
BLIP [24]	129M	19.2	35.2
mPLUG [22]	400M	21.1	37.2
HiTeA	5M	**21.7**	**37.4**

Table 10: **Zero-shot evaluation on video question answering.** Accuracy is reported. We gray out those methods additionally supervised pre-training on VQA v2 [12] dataset.

Qualitative Analysis. To verify that our model can capture the motion information with respected to the given text rather than inferring from the static signal, we present the query text in SSv2 Template dataset which has masked all of the object information, and also visualize the query in MSRVTT dataset. As we can see in the Figure 3, the attention map of atomic action "talking" mainly focuses on the mouse of the cartoon characters while the baseline largely focusing on the characters, which indicates that our method can understand the moment better when adopting the temporal-aware pre-training tasks. In another example, the word "spins" can reveal the trajectory of the object showing that our method is able to capture the temporal motion presented in the video.

4.4. Zero-shot Generalizability

To demonstrate the generalizability of proposed video-text pre-trained model, we perform zero-shot evaluation on video-language downstream tasks. We evaluate the zero-shot performance on VideoQA task in Table 10. Our method attains competitive zero-shot performance on MSRVTT-QA and MSVD-QA datasets even without help of audio signal supervision [60] or additional generated video question pairs [54]. In particular, less pre-training data (*i.e*. 5M $<$ 69M) are used while our method can still outperform other SoTA approaches. We also evaluate the zero-shot performance of models supervised on VQA v2 [12]. We can find that our method surpasses the powerful multi-modal SoTA methods (*e.g*. mPLUG [22]) with only 5M pre-training data showing the better generalization ability of **HiTeA**.

		MSRVTT			DiDeMo			LSMDC		
Method	# PT Data	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Frozen [2]	5M	18.7	39.5	51.6	21.1	46.0	56.2	9.3	22.0	30.1
ALPRO [23]	5M	24.1	44.7	55.4	23.8	47.3	57.9	-	-	-
BridgeFormer [11]	5M	26.0	46.4	56.4	25.6	50.6	61.1	12.2	25.9	32.2
Singularity [20]	5M	28.4	50.2	59.5	**36.9**	**61.6**	69.3	-	-	-
Models pre-trained on more data										
VideoCLIP [51]	138M	10.4	22.2	30.0	16.6	46.9	-	-	-	-
VIOLET [10]	183M	25.9	49.5	59.7	23.5	49.8	59.8	-	-	-
Clip4Clip [31]	400M	32.0	57.0	66.9	-	-	-	15.1	28.5	36.4
HiTeA	5M	**29.9**	**54.2**	**62.9**	36.1	60.1	**70.3**	**15.5**	**31.1**	**39.8**

Table 11: **Zero-shot evaluation on text-to-video retrieval.** All results are reported on R@1/R@5/R@10. We gray out methods that use significantly more data for fair comparison.

We also perform zero-shot evaluation on text-to-video retrieval task. Table 11 summarizes the performance of our model and compared approaches on text-to-video retrieval. We can observe that our model yields more than 3.4% lift in R@1 on MSRVTT dataset [53] while exploiting fewer video-text pairs. Besides, our method surpasses all of the compared models in terms of LSMDC dataset showing the superiority of our method's generalizability.

5. Conclusion

In this work, we introduce **HiTeA**, a novel hierarchical temporal-aware video-language pre-training framework with both understanding and generation capabilities. We vary the video with different views and model cross-modal temporal information between moments and texts as well as temporal contextual relations between multi-modal pairs in a hierarchical way. Specifically, a *cross-modal moment exploration* pre-training task is proposed to explore the fine-grained temporal information between the paired text and video moment by overcoming the partially semantic alignment. Moreover, multi-modal pairs are constructed to learn temporal contextual relations between moments and the event presented by the video with *multi-modal temporal relation exploration* pre-training task. We also demonstrate that our proposed pre-training tasks can consistently boost performance significantly on downstream tasks regardless of the backbone showing the generalization ability. Even pre-trained on less data, **HiTeA** still achieves state-of-the-art performance on a wide range of video-language downstream datasets, which clearly shows the superiority of our method.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 5, 6, 7

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1, 2, 3, 5, 6, 7, 9

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 6

[4] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022. 2, 3, 7, 8

[5] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 1, 5

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 4, 5

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 4

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 1

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 5

[10] 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. 1, 2, 3, 4, 5, 6, 9

[11] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. Bridging video-text retrieval with multiple choice questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16167–16176, 2022. 1, 2, 4, 5, 6, 8, 9

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 8

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 4, 5

[14] Jingjia Huang, Yinan Li, Jiashi Feng, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards a unified video-language alignment and fusion model. *arXiv preprint arXiv:2207.07885*, 2022. 2, 5, 6

[15] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017. 5, 7

[16] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116, 2020. 7

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 5, 6, 7

[19] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for multimodal video question answering. *International Journal of Computer Vision*, 129(11):3027–3050, 2021. 7

[20] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022. 1, 2, 3, 5, 6, 7, 8, 9

[21] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 2, 3, 4, 5, 6, 8

[22] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 1, 3, 8

[23] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4953–4963, 2022. 1, 2, 3, 4, 5, 6, 8, 9

[24] 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. 3, 8

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021. 1

[26] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 1, 2, 3

[27] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *arXiv preprint arXiv:2206.07160*, 2022. 1, 2, 3, 4, 5, 6, 8

[28] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 5

[29] 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. 6

[30] 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. 2, 6

[31] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3, 5, 7, 9

[32] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*, 2022. 3, 5

[33] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2017. 1, 5

[34] 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, 5

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

[36] 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. 6

[37] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 5, 6

[38] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 5, 6, 7

[39] Karsten Roth, Oriol Vinyals, and Zeynep Akata. Integrating language guidance into vision-based deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16177–16189, 2022. 2

[40] 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. 6

[41] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 5

[42] 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. 2

[43] Mingkang Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021. 6

[44] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 1, 5

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

[46] 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. 5, 6

[47] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1

[48] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021. 1, 3, 5, 7

[49] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. *arXiv preprint arXiv:2207.05342*, 2022. 7

[50] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 1, 5, 7

[51] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 9

[52] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, Guohai Xu, Ji Zhang, Songfang Huang, Fei Huang, and Jingren Zhou. mplug-2: A modularized multi-modal foundation model across text, image and video. In *Proceedings of International Conference of Machine Learning (ICML)*, 2023. 2

[53] 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, 6, 7, 8, 9

[54] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 6, 8

[55] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022. 8

[56] Qinghao Ye, Xiyue Shen, Yuan Gao, Zirui Wang, Qi Bi, Ping Li, and Guang Yang. Temporal cue guided video highlight detection with low-rank audio-visual fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7950–7959, 2021. 4

[57] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2

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

[59] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 1, 5, 7

[60] 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. 8

[61] 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. 1, 2, 3, 5, 6

[62] 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. 2