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

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

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

Contents

1. Additional Experimental Results

In this section, we provide more experimental results for completeness of our proposed method.

1.1. Transfer to Image-Text Downstream Tasks

Since images can be viewed as the single-frame videos, we evaluate the proposed method on image-text tasks including image-text retrieval and visual question answering.

Image-Text Retrieval We perform the Image-to-Text and Text-to-Image retrieval on COCO datasets, and the results are summarized in Table 1. We can observe that our method surpasses Singularity [19] with same amount of pre-train data, especially 1% improvement on Recall@1 for Text-to-Image retrieval task. Moreover, although some methods [7,24] leverage 4M dataset which contains the COCO dataset as a part of the pre-training dataset, **HiTeA** can still attain comparable results showing the good generalization ability.

*Corresponding Author.

Method	#PT Data	COCO (5K test) TR			IR		
		R1	R5	R10	R1	R5	R10
ViLT [16]	4M	61.5	86.3	92.7	42.7	72.9	83.1
UNITER [7]	4M	65.7	88.6	93.8	52.9	79.9	88.0
OSCAR [27]	4M	70.0	91.1	95.5	54.0	80.8	88.5
ALBEF [24]	4M	73.1	91.4	96.0	56.8	81.5	89.2
BLIP [23]	14M	80.6	95.2	97.6	63.1	85.3	91.1
ALIGN [15]	1.2B	77.0	93.5	96.9	59.9	83.3	89.8
Singularity [19]	5M	71.9	90.8	95.4	54.6	80.0	87.8
HiTeA	5M	72.4	90.9	95.4	55.6	80.6	87.8

Table 1: Comparison to existing methods on image-text retrieval on COCO dataset. We show results for both text retrieval (image-to-text retrieval, TR) and image retrieval (IR).

Visual Question Answering We also evaluate our method on visual question answering task. Table 2 concludes the image question answering results on VQAv2 [12] datasets. We observe that **HiTeA** demonstrates competitive performance on the VQA tasks. It is worthwhile noting that our method achieves the better performance compared to Singularity [19] same pre-training datasets, which indicates the video-text pre-training would boost the performance of image-text downstream tasks. However, we still see a gap with state-of-the-art image-text pre-trained models since our method do not use the in-domain data (*e.g.* COCO) during pre-training, thus leading to the gap with SoTA performance. One future direction is to use more image-text data during video-text pre-training for better generalization.

1.2. Additional Ablation Studies

Impact of positive candidate words size K. We investigate the effect of choosing different positive words size K during cross-modal moment exploration. As depicted in Figure 1, it can be observed that with the increment of K, the performance on each dataset is increasing then start to decrease. In addition, there is a trade-off between the choice of K and performance with respected to different datasets,

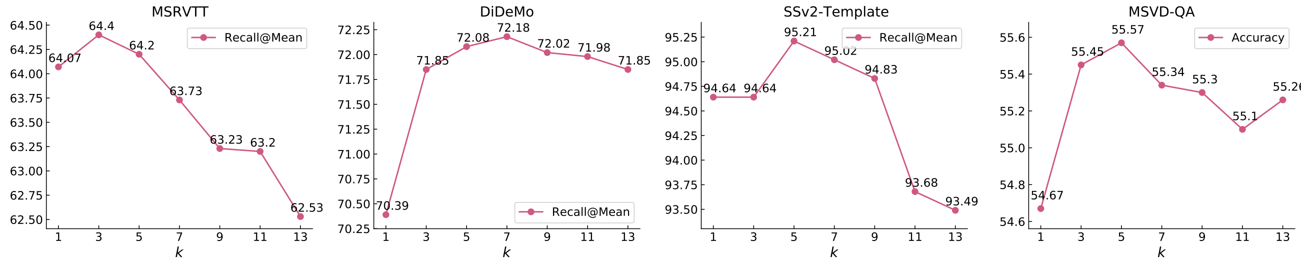

Figure 1: Variations in performance by changing the number of selected positive words K.

Method	#PT Data	test-dev	test-std
ClipBERT [20]	0.2M	69.08	69.43
ViLT [16]	4M	70.94	-
VL-BART [8]	0.2M	-	71.30
LXMERT [36]	4M	72.42	72.54
UNITER [7]	4M	72.70	72.91
UNIMO [26]	4M	73.79	74.02
OSCAR [27]	4M	73.16	73.44
ALBEF [24]	4M	74.54	74.70
BLIP [23]	14M	77.54	77.62
Singularity [19]	5M	70.30	70.53
HiTeA	5M	**74.06**	**74.28**

Table 2: Comparison to existing methods on VQA.

and $K = 5$ gives relative good results among these datasets. It also suggests that the small K would give more deterministic results since the model would only select the word with the largest similarity, thus more focusing on the single action or object. Then, as number of positive words increased, more accurate words are selected to align with the short-view of video. However, the model no longer benefits from cross-modal moment exploration when K is large enough (*i.e.*, $K = 11$ or $K = 13$) due to the increased noise in the selected candidate words.

Temporal evaluation of loss terms. To further validate the temporal dependency for the proposed method, we adopt the shuffling test for models with different loss terms, as shown in Table 3. Table 3 shows that our loss terms contribute more significantly when the dataset requires more temporal understanding. In concrete, $\mathcal{L}_{\text{CME}}$ and $\mathcal{L}_{\text{MTRE}}$ consistently improve the performances of Original and Gap on more temporal relied datasets (*i.e.* SSv2-Template and SSv2-Label). For example, model with two loss terms largely surpasses the baseline model in the metric of Gap by achieving 4.4 and 0.5 improvement on SSv2-Template and SSv2-Label, respectively.

Generalization to other vision backbone. Here, we demonstrate the generalization ability of our proposed methods by ablating that the proposed pre-training tasks on the plain backbone. Table 4 shows that our proposed method is generalizable to different vision backbones. In details, we instantiate the video encoder with TimeSformer [3] pre-trained on ImageNet-21K [34]. It can be observed both CME and MTRE consistently improve the model performance across the video backbones considered showing the generalization of proposed hierarchical temporal-aware pre-training framework. It is worth noting that, TimeSformer generates long video tokens compared to that of Multi-scale ViT [28], which brings extra memory cost for the multi-modal encoder and decoder since the computation of self-attention is quadratic. This makes TimeSformer expensive to scale to more input frames with longer sequences. Besides, we provide the computation analysis of them. The FLOPs of TimeSformer is 98.0 GFlops while that of MViT is 55.7 GFlops, which demonstrates that MViT outperforms TimeSformer in terms of efficiency without compromising video frame representation. As a consequence, we choose MViT as our default backbone.

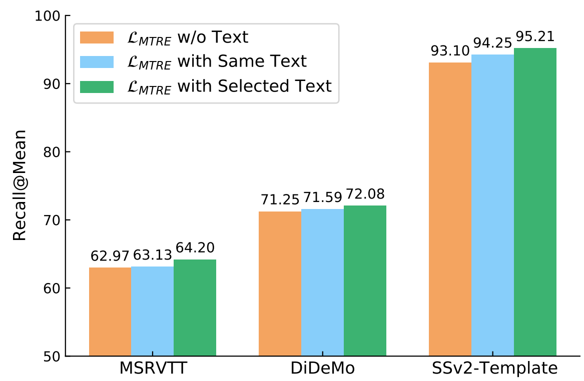

Figure 2: Variations in performance by adopting language during Multi-modal Temporal Relation Exploration (MTRE). We report the Mean Recall of Recall@1, Recall@5, and Recall@10.

Influence of language for MTRE. We investigate the influence of language for multi-modal temporal relation explo-

Method	MSRVTT [42]			SSv2-Template [19]			SSv2-Label [19]		
	Original ↑	Shuffled ↓	Gap ↑	Original ↑	Shuffled ↓	Gap ↑	Original ↑	Shuffled ↓	Gap ↑
$\mathcal{L}_{\text{base}}$	61.7	60.8	0.9	93.5	75.1	18.4	74.6	71.9	2.7
$\mathcal{L}_{\text{base}} + \mathcal{L}_{\text{CME}}$	63.7	62.9	0.8	94.4	72.6	21.8	74.8	71.8	3.0
$\mathcal{L}_{\text{base}} + \mathcal{L}_{\text{MTRE}}$	63.0	62.6	0.4	94.1	73.0	21.1	75.8	72.2	**3.6**
$\mathcal{L}_{\text{base}} + \mathcal{L}_{\text{CME}} + \mathcal{L}_{\text{MTRE}}$	64.2	63.3	**0.9**	95.2	72.4	**22.8**	76.7	73.5	3.2

Table 3: Evaluation of proposed methods for temporal dependency with temporal shuffling test. 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.

Method	MSRVTT	DiDeMo	SSv2-Template
TimeSformer ($\mathcal{L}_{\text{base}}$)	57.30	62.38	92.91
+ $\mathcal{L}_{\text{MTRE}}$	59.23	63.18	93.68
+ $\mathcal{L}_{\text{CME}}$	59.03	63.78	93.30
+ $\mathcal{L}_{\text{CME}} + \mathcal{L}_{\text{MTRE}}$	**59.93**	**65.34**	**94.25**

Table 4: Effectiveness of the proposed methods on different video backbone. We use TimeSformer [3] pre-trained on ImageNet-21K [34] to verify the generalization ability of our proposed method. For text-to-video retrieval, the Mean Recall of Recall@1, Recall@5, and Recall@10 is reported. For video question answering task, we report the Top-1 accuracy.

ration. Instead of utilizing the language signals, we directly adopt the video representation v_{cls} from the video encoder during the learning. The results are sketched in Figure 2. It can be observed that the model trained with multi-modal pairs attains better performance than the model without text. In concrete, it achieves 2.1% gains on SSv2-Template which mainly depends on the understanding of actions, which indicates that our method can better understanding the actions via multi-modal temporal relation exploration. Besides, we also notice that performance of the model trained with correct multi-modal pairs surpasses that of model trained by multi-modal pairs with same text, which indicates that improper video-text pair yields noisy multi-modal representation thus degrading the performance of the model.

Impact of the number of frames for fine-tuning. To further explore the capabilities of our model, we conducted experiments varying the number of frames used in downstream tasks. As depicted in Figure 3, our findings illustrate a marked improvement in performance when utilizing 2 to 8 frames, indicating that the model benefits from greater temporal intricacy. However, once the number of frames surpasses a certain threshold, performance levels off, suggesting that sufficient temporal details do not enhance performance further.

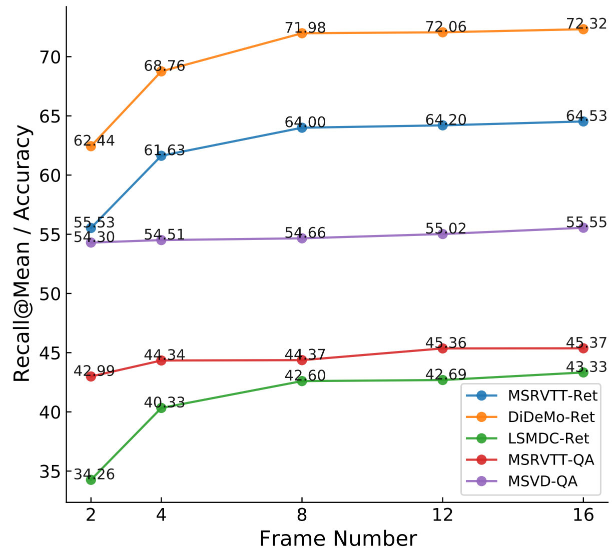

Figure 3: Impact of frame numbers during downstream fine-tuning. For text-to-video retrieval, the performance is shown as avg recall, i.e., average of R@1,5,10. For VideoQA task, the accuracy is reported.

2. Discussion

2.1. Qualitative Analysis

We sample some videos and corresponding texts and compute similarities between words and videos in Figure 4. As we can see in the figure, our model can effectively capture the moments such as "spreading", "moving", and "preparing" etc. in the video, which is essential for understanding videos. Besides, we can notice that the video would also attend to the object that appeared in the video, showing the capability for modeling fine-grained moment information.

2.2. Connection to Other Fine-Grained Methods

Some efforts [18,33,44] have been made to learn the fine-grained correlation and alignment between two modalities by leveraging the token-wise similarities in vision-language

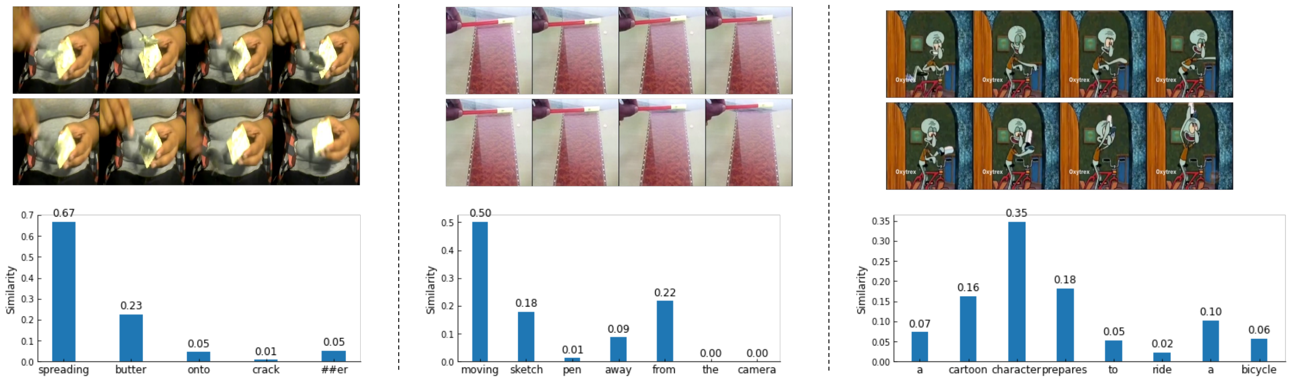

Figure 4: Examples of similarities between words and videos generated by our method. Our method captures the atomic actions in the videos as well as the object information with the help of cross-modal moment exploration.

pre-training. FILIP [44] and TERAN [33] aggregate the maximum token similarity scores and assign the optimal patch-word transport matrix. SCAN [18] utilizes the similarity scores to attend each tokens for soft fine-grained alignment. These approaches are originally tailored for image-text pre-training, which aims to locate the fine-grained static object. However, different from image-text pre-training, video-text pre-training needs to understand the correlation between words and moments, which not only contains static objects but also consists of atomic actions. Our proposed cross-modal moment exploration leverages the short-view of video to reflect the moment information and discover the relationship between short-view videos and words, which results in fine-grained moment representations for video-language pre-training.

Besides, some other efforts [5, 39, 43] are proposed to model the fine-grained text information by leveraging the whole video. For example, HGR [5] constructs the text semantic graph for hierarchical alignment, while T2VLAD [39] and TACo [43] weight the importance of words in the text through multiple video experts with VLAD and IDF of words respectively. However, all of these methods are aim to filter useful text tokens for the whole video without considering detailed temporal information in the video. Our cross-modal exploration task captures **both fine-grained text and temporal information** for modeling atomic actions and moments, which the contribution of CME task lies in the **temporal aspect** and is crucial for understanding the temporal details revealed in untrimmed videos for video-language pre-training.

2.3. Limitations and Boarder Impact

Despite the effectiveness of the proposed method on various downstream tasks, our method still has some limitations that would make for promising directions for future work. (1) Currently, we only pre-train our model on 5M data with the base-size encoders, and the scalability of the model is not explored which deserves more in-depth investigation in the future. (2) Our method shares similar risks like other pre-training methods that the pre-training data might consist bias and unsafe content which requires further analysis before the deployment.

Method	# of Parameter
ClipBERT [20]	137M
Frozen [2]	232M
BridgeFormer [11]	152M
All-in-one [38]	110M
VIOLET [10]	198M
ALPRO [22]	231M
Singularity [19]	209M
LAVENDER [10]	198M
HiTeA	297M

Table 5: Comparison to other models in the number of parameters.

3. Implementation Details

3.1. Number of Parameters

We include some of previous models with their parameter counts (which were reported in the original paper or calculated by follow-up work), and we compare them with **HiTeA** in Table 5. Compared with other models, our model is of comparable model size and requires less pair of video-text pre-training data to achieve better performance in terms of both video-language understanding and generation.

3.2. Model Architecture

As sketched in Figure 5, our model consists two uni-modal encoders for text and video respectively, a multi-

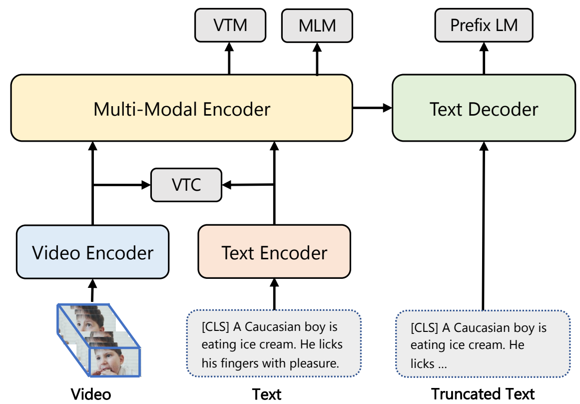

Figure 5: Architecture of the proposed **HiTeA** and other pre-training objectives.

modal encoder for video-text interaction, and a text decoder for generation. In concrete, given an arbitrary view of video $V \in \mathbb{R}^{T\times H\times W}$ is encoder into a sequence of embeddings: $\{v_{\text{cls}}, v_1, \cdots, v_M\} \in \mathbb{R}^{(M+1)\times D}$, where M is the number of flattened patches for video V, and v_{cls} is the embedding of the visual [CLS] token and used to provide global representation of the video. The text encoder transforms the text into a sequence of embeddings: $\{w_{\text{cls}}, w_1, \cdots, w_N\} \in \mathbb{R}^{(N+1)\times D}$, where N is the number of words in the text. To efficiently encode multi-modal information while preserving unimodal information, we fuse the video and text features from uni-modal encoders following [21]. The output of the multi-modal encoder $\{\boldsymbol{v}_{\text{cls}}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_M, \boldsymbol{w}_{\text{cls}}, \boldsymbol{w}_1, \cdots, \boldsymbol{w}_N\} \in \mathbb{R}^{(M+N+2)\times D}$ is fed into a transformer decoder for sequence to sequence generation, which equips **HiTeA** with the capabilities of both multi-modal understanding and generation.

3.3. Pre-training Objectives

During pre-training, we also perform four pre-training tasks including Video-Text Contrastive Learning ($\mathcal{L}_{\text{VTC}}$), Video-Text Matching ($\mathcal{L}_{\text{VTM}}$), Masked Language Modeling ($\mathcal{L}_{\text{MLM}}$), and Prefix Language Modeling ($\mathcal{L}_{\text{PrefixLM}}$). The VTC task first is applied to align the unimodal representation of video and text. And the multi-modal representation can be learned by VTM and MLM tasks. Upon on the video-language representations obtained from multi-modal encoder, the decoder is trained by PrefixLM loss with text completion task.

Video-Text Contrast (VTC) Following [22, 38], we align the unimodal encoders via this task. Specially, the softmax-normalized video-to-text and text-to-video similarities are computed, and we employ memory queues in MoCo [6] to increase the number of negative samples during learning. Formally, the video-text contrastive loss is calculated as:

$$\begin{aligned}
\mathcal{L}_{v2t} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s(\mathcal{V}_i,\mathcal{T}_i))}{\sum_{j=1}^{B}\exp(s(\mathcal{V}_i,\mathcal{T}_j))}, \\
\mathcal{L}_{t2v} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s(\mathcal{V}_i,\mathcal{T}_i))}{\sum_{j=1}^{B}\exp(s(\mathcal{V}_j,\mathcal{T}_i))}, \\
\mathcal{L}_{\text{VTC}} &= \frac{1}{2}(\mathcal{L}_{v2t}+\mathcal{L}_{t2v}),
\end{aligned} \tag{1}$$

where $\mathcal{V}_i$ and $\mathcal{T}_j$ are the projected representations of v_{cls} and w_{cls} for i-th video-text pair in the batch.

Video-Text Matching (VTM) This task aims to predict whether a video and a text is paired or not based on the multi-modal representation. As suggested in [22, 24], hard negative video-text pairs are selected based on the similarity of video and text during contrastive learning. Formally, the video-text matching loss is calculated as:

$$\mathcal{L}_{\text{VTM}} = -\mathbb{E}_{(\mathcal{W},\mathcal{V})}\log p(y|\mathcal{W},\mathcal{V}), \tag{2}$$

where $\mathcal{W}$ denotes the word tokens, and $\mathcal{V}$ denotes the video features of long-view video.

Masked Language Modeling (MLM) The setup of this pre-training task is same as that used in BERT [9], where 15% of tokens in the text are randomly masked, and the model needs to predict the masked tokens based on the multi-modal representation. Formally, the masked language modeling loss is calculated as:

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{(\mathcal{W},\mathcal{V})}\log p(w_i|\mathcal{W}_{\backslash i},\mathcal{V}), \tag{3}$$

where w_i denotes the masked word token.

Prefix Language Modeling (PrefixLM) This pretext task requires model to complete the truncated texts based on given videos and prefix sequence of truncated texts [21, 23]. The model can be trained by maximizing the likelihood of the truncated text in an auto-regressive manner. Formally, the prefix language modeling loss is calculated as:

$$\mathcal{L}_{\text{PrefixLM}} = -\mathbb{E}_{(\mathcal{W},\mathcal{V})}\left[\sum_{l=L_p}^{L}\log p(w_l|\mathcal{W}_{[L_p,l)},\mathcal{W}_{<L_p},\mathcal{V})\right], \tag{4}$$

where L denotes the total number of words in the text, and L_p is the length of a prefix sequence of tokens which is randomly selected.

3.4. Downstream Task Implementation Details

We evaluate **HiTeA** on various downstream video-language tasks, including Text-to-Video Retrieval, Open-ended VideoQA, Multiple Choice VideoQA, and Video Captioning. The fine-tuning procedures are described as follows:

Dataset	Optimizer	Learning Rate	Weight Decay	LR Schedule	Batch Size $\times$ # GPUs	Epochs
MSRVTT-Ret [42]	AdamW	2e-5	0.02	Cosine Decay	24×8	10
DiDeMo [1]	AdamW	1e-5	0.02	Cosine Decay	24×8	20
LSMDC [35]	AdamW	2e-5	0.02	Cosine Decay	24×8	10
Activity Caption [17]	AdamW	2e-5	0.02	Cosine Decay	24×8	20
SSv2-Template [19]	AdamW	5e-5	0.02	Cosine Decay	24×8	20
SSv2-Label [19]	AdamW	2e-5	0.02	Cosine Decay	24×8	20
MSRVTT-QA [41]	AdamW	2e-5	0.02	Cosine Decay	16×8	8
MSVD-QA [41]	AdamW	2e-5	0.02	Cosine Decay	16×8	8
TGIF-FrameQA [14]	AdamW	2e-5	0.02	Cosine Decay	16×8	8
LSMDC-FIB [32]	AdamW	2e-5	0.02	Cosine Decay	16×8	8
ActivityNet-QA [46]	AdamW	2e-5	0.02	Cosine Decay	16×8	8
TGIF-Action [14]	AdamW	3e-5	0.02	Cosine Decay	16×8	56
TGIF-Transition [14]	AdamW	3e-5	0.02	Cosine Decay	16×8	30
LSMDC-MC [37]	AdamW	2e-5	0.02	Cosine Decay	16×8	10
NExT-QA [40]	AdamW	2e-5	0.02	Cosine Decay	16×8	10
MSRVTT-Caption [42]	AdamW	2e-5	0.02	Cosine Decay	24×8	10
MSVD-Caption [4]	AdamW	2e-5	0.02	Cosine Decay	24×8	10

Table 6: End-to-end fine-tuning configurations for video-language downstream tasks.

- For retrieval tasks, we jointly optimize the VTC loss and VTM loss for video-text alignment during fine-tuning. During inference, we first select top-k candidates by computing the dot-product similarity between the video and text features, and then reranking the selected candidates based on their VTM scores. k is set to 128 by default.
- For open-ended VideoQA, we first generate video features and text features with two unimodal encoders, and then fuse them with multi-modal encoder. The output of multi-modal features are fed to text decoder for answer generation. We use the language modeling loss to optimize the model. During inference, the answer would be generated by the text decoder.
- For multiple choice VideoQA, we treat the problem as the text-to-video retrieval task where the correct answer should have the highest matching probability. During training, we compute the VTM scores for each candidate answer and video, then optimize the model with cross entropy loss. During the inference, the answer with highest VTM score is the prediction answer.
- For Video Captioning, we use the video features from video encoder and directly feed it into text decoder for caption generation. The language modeling loss is utilized for model optimization.

For all above video-language downstream tasks, we resize video frames to 224×224. During fine-tuning, following [19,22], we randomly sample 12 frames for text-to-video retrieval, 16 frames for video question answering and video captions, and we perform temporal downsampling with factor 2 for video before feeding into the network. Here, we sample video frames from the whole video instead of treating videos into different views. During inference, we adopt uniform sampling for video frames. We use RandomCrop with minimum ratio 0.5 and HorizontalFlip with 0.5 probability for data augmentation. The hyperparameters that we used for fine-tuning on downstream tasks are summarized in Table 6. For the video caption task, we use a prefix prompt "A video of" to improve the quality of generated captions.

3.5. Datasets Description

In this section, we describe all of the downstream video-language datasets used during evaluation. The details of the datasets are represented below:

Text-to-Video Retrieval. We evaluate **HiTeA** on 6 popular text-to-video retrieval datasets including MSRVTT [42], DiDeMo [1], LSMDC [35], ActivityNet Caption [17], SSv2 Template [19], and SSv2 Label [19]. Details of these datasets: **MSRVTT** [42] contains 10K YouTube sourced videos with 200K text descriptions. Following [13, 25, 30], we train the video on 9K videos and evaluate on the rest 1K video. **DiDeMo** [1] contains of 10K videos from Flickr and 4 descriptions for each video. Following [22, 25, 31], we concatenate all of the given descriptions from the same video as a paragraph, and evaluate the paragraph-to-video retrieval performance. The number of video in training set is 8K, leaving 1K for validation set and 1K for test set. **LSMDC** [35] consists of 118K video clips from 202 movies, and each clip is accompanied with a caption from video scripts. It has 101K video clips for training and 1K clips for testing. We use the standard splits from [35]. **Activi-**

tyNet Caption [17] is built on 20K YouTube videos with 100K captions. We use the train split with 10K videos for training, and report the performance on the val1 split with 4.9K videos. **SSv2-Template** and **SSv2-Label** [19] contain 169K videos for training and 2K videos for testing. The text queries in SSv2-Template are templates without object information (*e.g*. "Throwing [something] in the air and catching it"). By contrast, SSv2-Label contains annotated text queries with specific object information (*e.g*. "Throwing keys in the air and catching it"). Therefore, SSv2-Template mainly focuses on temporal understanding of actions, while SSv2-Label needs a more comprehensive understanding of both appearance and temporal dynamic.

Multiple-choice Video QA. Five datasets are evaluated for multiple-choice video question answering tasks. **TGIF-Action** and **TGIF-Transition** [14] are adopted to evaluate model's capability to recognize the repeated actions and state transitions in short GIFs. Each video and question is equipped with 5 candidate answers. We concatenate the question and answer as the text and use the highest similarity among the video and candidate texts. TGIF-Action contains 18K GIFs for training and 2K for testing. TGIF-Transitions has 47K GIF-question pairs for training and 6K for testing. **MSRVTT-MC** [45] and **LSMDC-MC** [37] are originally retrieval task, but reformulated as the multiple choice video QA task. It requires the model to find the optimal caption that describes the video out of 5 candidate texts. **NExT-QA** [40] is explicitly designed for temporal and causal understanding. Questions in the dataset are categorized into three types: Descriptive, Temporal, and Causal. Each question in the dataset are paired with 5 candidate answers. Therefore, this dataset is able to evaluate model's ability in video question answering in different aspects.

Open-ended Video QA. For open-ended video QA, we evaluate the model on five datasets. **MSRVTT-QA** is composed of 243K open-ended questions over 10K videos, while **MSVD-QA** [41] consists 2K videos with 47K questions. **TGIF-Frames** [14] collects the answerable with just a single frame in the video, and is divided into training set with 35K questions and test set with 14K questions.. For LSMDC-FiB [32], the model needs to predict a correct word for the blank with a given video and a sentence with blank. It contains 297K sentences for training and 30K sentences for testing. **ActivityNet-QA** [46] .

Video Captioning. We use MSRVTT [42] and MSVD [4] for video captioning evaluation. As described before, **MSRVTT** is composed of 10K videos with 20 captions per video, and **MSVD** contains 2K videos with around 40 captions per video. We follow the standard splits from [25, 29]. During inference, we generate the caption with beam search until the model outputs a [SEP] that indicates the end of sentence or when it reaches the maximum generation step 40.

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. 6

[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. 4

[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. 2, 3

[4] 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. 6, 7

[5] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 4

[6] 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. 5

[7] 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, 2

[8] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 2

[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. 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. 4

[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. 4

[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. 1

[13] 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. 6

[14] 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. 6, 7

[15] 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. 1

[16] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1, 2

[17] 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. 6, 7

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 3, 4

[19] 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, 4, 6, 7

[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, 4

[21] 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. 5

[22] 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. 4, 5, 6

[23] 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. 1, 2, 5

[24] 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, 2, 5

[25] 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. 6, 7

[26] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 2

[27] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 2

[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. 2

[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. 7

[30] 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. 6

[31] 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. 6

[32] 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. 6, 7

[33] Nicola Messina, Giuseppe Amato, Andrea Esuli, Fabrizio Falchi, Claudio Gennaro, and Stéphane Marchand-Maillet. Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(4):1–23, 2021. 3, 4

[34] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 2, 3

[35] 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. 6

[36] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[37] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 6, 7

[38] 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. 4, 5

[39] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2021. 4

[40] 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. 6, 7

[41] 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. 6, 7

[42] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3, 6, 7

[43] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11562–11572, 2021. 4

[44] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 3, 4

[45] 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. 7

[46] 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. 6, 7