This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Video Language Model Pretraining with Spatio-temporal Masking

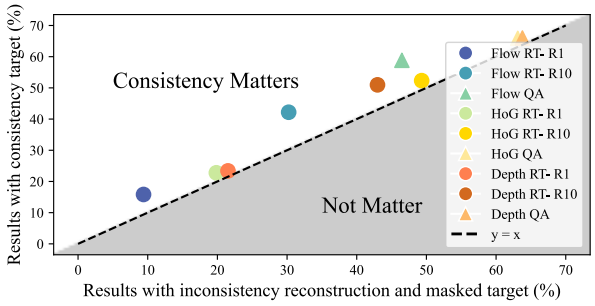
Yue Wu[1,2,4], Zhaobo Qi[3], Junshu Sun[1], Yaowei Wang[2,5], Qingming Huang[2,4], Shuhui Wang[1,2*]
[1] Key Lab of Intell. Info. Process., Inst. of Comput. Tech., CAS
[2] Pengcheng Laboratory [3] Harbin Institute of Technology, Weihai
[4] University of Chinese Academy of Sciences [5] Harbin Institute of Technology, Shenzhen
wuyue221@mails.ucas.ac.cn qizb@hit.edu.cn wangyw@pcl.ac.cn
{sunjunshu21s, wangshuhui}@ict.ac.cn qmhuang@ucas.ac.cn

Abstract

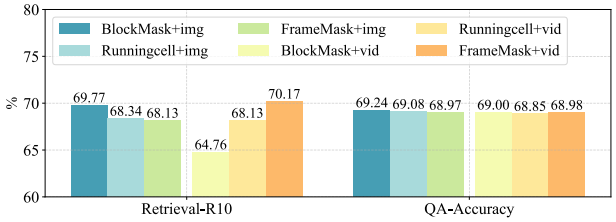
The development of self-supervised video-language models based on mask learning has significantly advanced downstream video tasks. These models leverage masked reconstruction to facilitate joint learning of visual and linguistic information. However, recent study reveals that reconstructing image features yields superior downstream performance compared to video feature reconstruction. We hypothesize that this performance gap stems from the way how masking strategies influence the model's attention to temporal dynamics. To validate this hypothesis, we performed two sets of experiments that demonstrate that alignment between the masked target and the reconstruction target is crucial for self-supervised video-language learning. Based on these findings, we propose a spatio-temporal masking strategy (STM) for video-language model pretraining that operates across adjacent frames, and a decoder leverages semantic information to enhance the spatio-temporal representations of masked tokens. Thanks to the combination of masking strategy and reconstruction decoder, STM enforces the model to learn spatio-temporal feature representation comprehensively. Experiments in three video understanding downstream tasks validate the superiority of our method. Codes are available here.

1. Introduction

Video-language models (VLMs) have made significant strides in advancing video understanding [15, 23, 28, 38, 46, 49, 50]. Through large-scale pre-training on video-text pairs, these models learn to align visual dynamics with semantic concepts, establishing rich cross-modal representations that enhance various downstream tasks such as video retrieval [13, 35, 36], video question answering (video QA) [22, 54], and video captioning [40, 58].



(a) masked targets w./wo. consistent reconstruction targets.



(b) Reconstructing image/video feature with different masking strategies.

Figure 1. Different masked target + reconstruction target results: (a) Retrieval (RT) and video QA (QA) results are improved when the reconstruction targets are consistent with the masked targets. (b) Spatio-temporal-oriented masking strategies benefit downstream tasks when reconstructing video features, but may hurt when reconstructing image features.

Following the success of masked autoencoders [16], the masking and reconstruction strategies have been extended to video-language models to enhance their multimodal representation capabilities [8, 11, 13, 30, 31, 59]. Notably, VIOLETv2 [12] systematically investigates the reconstruction targets for video understanding, *i.e.*, pixel, HoG, depth, flow, image features, video features, and multimodal features. The investigation reveals a counterintuitive finding, *i.e.*, when masking image-level information, the reconstruction of image features yields superior performance com-

*Corresponding author.

pared to video feature reconstruction. This is contradictive to the previous study that reveals the benefit of richer temporal and contextual information from multi-frame inputs over single-frame inputs [26, 33]. We hypothesize that this is caused by the masking strategies of VIOLETv2. These strategies, which rely on random or block-wise masking schemes originally designed for static images, may not adequately capture the spatio-temporal correlations inherent in video data. Meanwhile, the inconsistency between the masking strategy and the spatio-temporal feature reconstruction target further aggravates this issue.

To validate this hypothesis, we conduct two sets of experiments. Following VIOLETv2's original setting, given three basic reconstruction targets including HoG, depth, and flow, we compare the random masking strategy with directly masking the reconstruction targets. As shown in Figure 1a, consistent improvements are observed across all the three targets on several downstream tasks when the masked target matches the reconstruction target. We also evaluate three masking strategies with varying spatio-temporal correlation intensities including BlockMask [12] (spatial-only), RunningCell [39] (low spatio-temporal), and FrameMask [14] (high spatio-temporal), for video and image feature reconstruction. Figure 1b shows that spatio-temporal-oriented masking strategies empower better video feature reconstruction, but may hurt the reconstruction of image features. These results verify our hypothesis that stronger consistency between the masked targets and the reconstruction targets benefits mask-based self-supervised learning.

However, different tasks require distinct visual features. The retrieval task relies on spatio-temporal alignment between modalities, while the QA task necessitates a more refined and comprehensive understanding of spatio-temporal relationships. As evidenced in Figure 1b, incorporating spatio-temporal masks significantly improves the retrieval performances but shows limited improvement in the QA results. To enhance performance across diverse video understanding tasks, a masking strategy for pretraining that facilitate spatio-temporal understanding beyond alignment objectives needs to be further explored.

Built upon these experimental insights, we propose a paradigm of video-language model training with consistent **S**patio-**T**emporal **M**asking (STM) and reconstruction targets. The masking strategy first independently calculates both the temporal and spatial contributions of the visual tokens. It masks visual tokens with rich global spatio-temporal information across consecutive frames at a predefined ratio. The decoder then processes these masked visual tokens through semantic-level spatio-temporal enhancement mechanisms to guide their reconstruction. To compel the network to learn these masked global spatio-temporal features, we employ Masked Visual Modeling loss combined with three conventional pretraining objectives to facilitate comprehensive global understanding and cross-modal alignment. Our contribution is threefold:

- We observe the inconsistency between the masked targets and the reconstruction targets, which affects the pretraining performance in video understanding tasks.
- We propose a masking strategies and a decoder to enforce spatio-temporal reconstruction and semantic consistency for video understanding.
- Experiments on multiple datasets demonstrate the promising performance of STM on text-video retrieval, video QA and captioning.

2. Related Work

Video-Language Understanding. Video-language understanding focuses on comprehending the relationship inside and between video content and textual semantics, serving as a foundation for downstream video analytic tasks. Recent advances in this field can be categorized into temporal modeling, knowledge distillation, and self-supervised learning. For temporal modeling, methods such as HiTeA [15, 37, 59] leverage hierarchical structures to capture multi-scale temporal dependencies. For knowledge distillation, models like InternVideo [49, 50], OmniViD [46], and Chat-UniVi [23] leverage large language model or large multi-modality models to incorporate extensive visual and linguistic knowledge. For self-supervised learning, various approaches [8, 12, 26, 30, 31, 33] leverages Masked AutoEncoder (MAE) [16] techniques for training. These methods improve the performance by reconstructing patches through four pretraining objectives, which include visual-text contrastive learning, visual-text matching, masked language modeling, and masked video modeling. Following the mask-based learning paradigm, we investigate the spatio-temporal masking and reconstruction that emphasize the representation of spatio-temporal information in videos.

Mask Visual Learning. MAE [16] introduces self-supervised learning by reconstructing masked image patches, where the key challenges lie in designing effective masking strategies and reconstruction targets. For masking strategies, approaches range from simple random masking to more sophisticated designs. Basic strategies include random patch masking for general understanding, and block-wise or tube-wise masking for simple spatial-temporal correlations. Advanced strategies for videos focus on temporal dynamics, such as frame-wise random masking in OmniMAE [46], motion-guided masking [1, 17], and running cell masking [39] for focusing on the spatio-temporal continuity. For reconstruction targets, approaches vary from low-level reconstruction to high-level feature reconstruction. While MAE focuses on pixel reconstruction, MVP [53] and BEiT [5] predict token-based features, MaskFeat [52] reconstructs HoG features for low-level visual information,

and MME [42] reconstructs optical flow for motion understanding. Notably, an empirical study, *i.e.*, VIOLETv2 [12], comprehensively evaluates seven different reconstruction targets across various multimodal tasks. It compares pixels, depth, optical flow, HOG, image, video, and multimodal features, illustrating how different targets can influence model performance across multimodal tasks. Different from VIOLETv2 [12], we investigate the consistency between masking strategies and reconstruction targets, aiming to enhance the ability to extract spatio-temporal information for video-language understanding.

3. Method

In this section, we introduce our video-language model leveraging **S**patio-**T**emporal **M**aking (**STM**). The framework consists of a spatio-temporal masking strategy and a decoder that effectively reconstructs video representations, as shown in Figure 2.

3.1. STM Overview

As discussed in Section 1, maintaining consistency between masked targets and reconstruction targets enhances the representational capability of pretraining models. For video understanding tasks, spatio-temporal features in videos play a crucial role. Therefore, to improve the model's capability in video understanding tasks, we focus on spatio-temporal feature reconstruction while maintaining the consistency between the masked targets and the reconstruction targets through spatio-temporal masking.

Architecture. Given a video-text pair consisting of a video $\boldsymbol{v} = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_t, ..., \boldsymbol{v}_T\}$ with T frames and a text sentence $\boldsymbol{z}$, we first employ a tokenizer to transform the video data into a sequence of tokens, denoted as $\boldsymbol{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_t, ..., \boldsymbol{x}_T\}$. The token sequence is then divided into two sets based on their locations, including visible tokens $\boldsymbol{x}^{\text{vis}}$ and invisible tokens $\boldsymbol{x}^{\text{inv}}$, through our spatio-temporal masking strategy (detailed in Section 3.2). The invisible tokens are masked by zero padding, which will be reconstructed in the decoder. The visible tokens $\boldsymbol{x}^{\text{vis}}$ are then processed through a video encoder to obtain visible video features $\tilde{\boldsymbol{x}}^{\text{vis}}$. We employ the vanilla ViT [10] as our video encoder. Next, a spatio-temporal decoder (detailed in Section 3.2) combines both the visible and invisible tokens $\tilde{\boldsymbol{x}} = \{\tilde{\boldsymbol{x}}^{\text{vis}}, \boldsymbol{x}^{\text{inv}}\}$ to reconstruct the full video feature $\hat{\boldsymbol{x}} = \{\hat{\boldsymbol{x}}^{\text{vis}}, \hat{\boldsymbol{x}}^{\text{inv}}\}$. For the text modality, we employ random word masking and a cross-modal decoder following UMT [30] for text reconstruction. Finally, we implement the following pretraining objectives to ensure that the model focuses on spatio-temporal features.

Masked Visual Modeling Objective. To enable our encoder to learn spatio-temporal video features, we reconstruct video features aligned with STM strategy. Following VIOLETv2 [12], we implement the masked visual modeling (MVM) loss $\mathcal{L}_{\text{MVM}}$ using a supervised learning approach where a well-trained video model serves as the teacher model f_{tea}, providing pseudo-labels as video feature reconstruction targets. Specifically, $\mathcal{L}_{\text{MVM}} = \text{L2}(\hat{\boldsymbol{x}}^{\text{inv}}, f_{\text{tea}}(\boldsymbol{x}^{\text{inv}}))$ measures the L2 distance between reconstructed features and teacher model's output. We employ ActionCLIP [48] as teacher model, leveraging its strong spatio-temporal modeling and multimodal representation ability.

Total Pre-training Objectives. Following methods like VIOLETv2 [11, 12, 30, 59], we adopt four standard pretraining objectives, *i.e.*, video-text matching (VTM) and video-text contrastive learning (VTC) for cross-modal feature alignment, MVM to evaluate video feature reconstruction accuracy, and masked language modeling (MLM) to assess text reconstruction correctness. The overall pretraining objective is formulated as the sum of these four components:

$$\mathcal{L} = \mathcal{L}_{\text{VTC}} + \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{MVM}} + \mathcal{L}_{\text{MLM}}. \tag{1}$$

3.2. Spatio-temporal Masking

The principle of our masking strategy is to identify video patches that capture the key spatio-temporal information. We calculate the spatial and temporal contribution quantification scores separately. The spatial contribution quantification measures the contribution of spatial features within each patch at the image level, while the temporal contribution quantification analyzes the temporal information between adjacent frames. These two contribution scores are then combined to generate the final mask.

Spatial Contribution. To obtain the spatial contribution of image information, the most straightforward approach is to use the attention mechanism within the visual encoder [30]. Tokens with higher weights often correspond to regions with rich structure information that the model focuses on. Therefore, we utilize attention in the visual encoder to compute the spatial contribution within the image. Specifically, given the visual encoder of the well-trained teacher model f_{tea}, the contribution score $\boldsymbol{s}_t^{\text{sp}} \in \mathbb{R}^{1\times N}$ is quantified between the class token $[cls] \in \mathbb{R}^{1\times D}$ and the tokens $\boldsymbol{x}_t \in \mathbb{R}^{N\times D}$. N is the number of tokens in $\boldsymbol{x}_t$, while D is the hidden dimension. The calculation is as follows:

$$\boldsymbol{s}_t^{\text{sp}} = \text{softmax}(\Phi_{\text{Q}}([cls]) \cdot \Phi_{\text{K}}(\boldsymbol{x}_t^{\top})/\sqrt{D}), \tag{2}$$

where the Φ is a linear layer. Guided by the global semantic feature $[cls]$, the attention mechanism of the visual encoder can focus on the most important spatial regions. Subsequently, we concatenate the spatial contribution score of all frames to the overall $\boldsymbol{s}^{\text{sp}} \in \mathbb{R}^{T\times N}$.

Temporal Contribution. To identify the temporal contribution of video, one might consider utilizing contribution

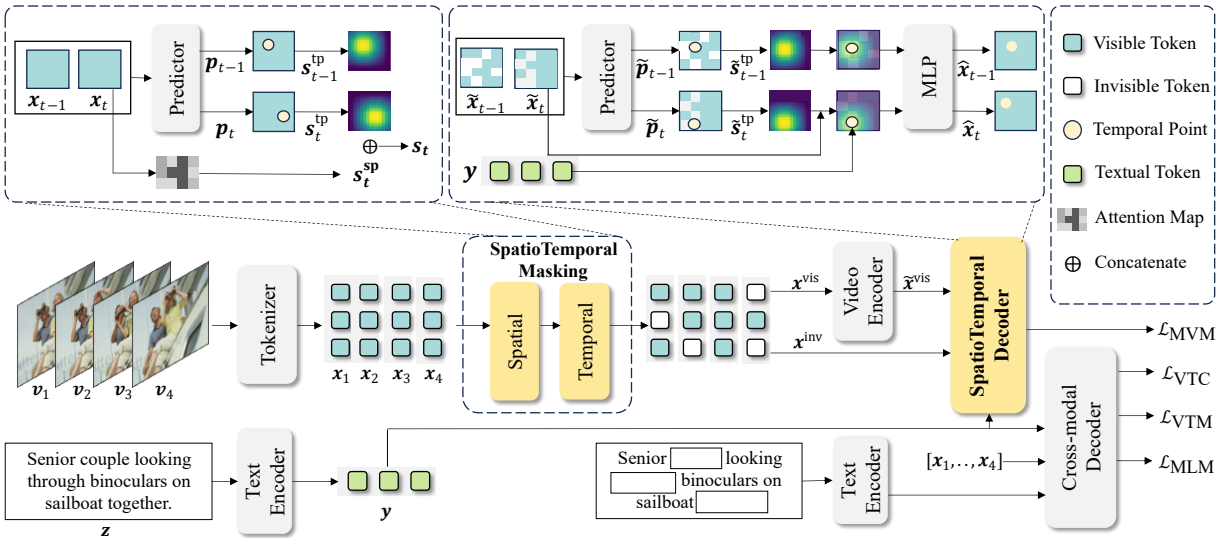


Figure 2. The STM pipeline consists of the spatio-temporal masking strategy and the decoder. The spatio-temporal masking strategy masks the video based on spatial and temporal contribution. The Spatio-Temporal Decoder reconstructs spatio-temporal features by leveraging the relationships between adjacent frames and applying semantic enhancement. The reconstructed features are then used for MVM loss computation. Additionally, the commonly used losses, *i.e.*, VTC, VTM and MLM, are also incorporated into the training of STM.

scores from the temporal encoder of f_{tea}. However, this encoder focuses on frame-level information, neglecting token-level temporal dependencies. Previous studies [19, 20] mask patches that have significant inter-frame differences. However, these approaches may be unreliable as high visual variation does not necessarily correlate with temporal significance. Inspired by methods that extract spatio-temporal information via adjacent frame comparison [43, 63], we propose to predict the temporal information location and generate the heatmap by aligning information between adjacent frames.

Specifically, we take adjacent frame tokens $\boldsymbol{x}_{t-1} \in \mathbb{R}^{P_1 \times P_2 \times D}$ and $\boldsymbol{x}_t \in \mathbb{R}^{P_1 \times P_2 \times D}$, where P_1 and P_2 denote the number of tokens along the width and height, respectively, and $P_1 \times P_2 = N$. These tokens are concatenated along the hidden dimension and fed into a lightweight prediction network. The network predicts two temporal points $\boldsymbol{p}_{t-1} \in \mathbb{R}^{1\times 2}$ and $\boldsymbol{p}_t \in \mathbb{R}^{1\times 2}$ as spatial coordinates:

$$\boldsymbol{p}_{t-1}, \boldsymbol{p}_t = \text{Predictor}([\boldsymbol{x}_{t-1} \oplus \boldsymbol{x}_t]), \tag{3}$$

where $\text{Predictor}(\cdot)$ consists of a sequence of convolution, normalization, pooling layers, and MLPs. The operator $[\cdot \oplus \cdot]$ denotes concatenation. The predicted temporal point $\boldsymbol{p}_t$ indicates the locations that present meaningful temporal information between frames.

Furthermore, tokens closer to these temporal points generally contain related temporal information, while distant tokens are less temporally relevant. Based on this, we employ the distance-based weighting method [43] to generate the temporal heatmap. The weighting function is formulated as:

$$\boldsymbol{s}^{\text{tp}}_{t,i} = \begin{cases} \eta, & \text{if } d \le \delta, \\ \max(0, \eta - \beta \times (d - \delta)), & \text{if } d > \delta, \end{cases} \tag{4}$$

where $\boldsymbol{s}^{\text{tp}}_t \in \mathbb{R}^{1\times N}$, $d = \text{dist}(\boldsymbol{p}, \boldsymbol{x}_{t,i})$ denotes the distance from the temporal point $\boldsymbol{p}_t$ to the i-th token $\boldsymbol{x}_{t,i}$ of $\boldsymbol{x}_t$. δ represents a distance threshold, η represents the maximum weight, and β represents the rate of weight decay with distance. This formulation ensures higher weights for the regions closer to the predicted temporal points and lower weights for the far regions. Note that for each pair of consecutive frames, two heatmaps are generated. We retain the original heatmaps for the first and last frames, while for the intermediate frames, we use the average of the heatmaps from the adjacent pairs. Finally, we concatenate all the scores of tokens to get the temporal score $\boldsymbol{s}^{\text{tp}} \in \mathbb{R}^{T\times N}$.

Masking. To efficiently identify key tokens, we compute the final contribution score $\boldsymbol{s}$ by summing the spatial similarity score $\boldsymbol{s}^{\text{sp}}$ and temporal similarity score $\boldsymbol{s}^{\text{tp}}$. Based on a masking ratio r, we rank the tokens in $\boldsymbol{s}$ and select the top-r proportion of tokens to generate invisible tokens $\boldsymbol{x}^{\text{inv}}$, while the remaining tokens are designated as visible tokens $\boldsymbol{x}^{\text{vis}}$. This approach enables us to effectively locate and mask the most informative regions while maintaining computational efficiency.

3.3. Spatio-temporal Decoder

The proposed spatio-temporal masking in Section 3.2 poses a challenge regarding to reconstruction where the decoder has to reconstruct complete video features from limited cues. To address this challenge, we introduce a text-guided spatio-temporal decoder that leverages the remaining weak information to synthesize video features with high quality.

Specifically, given the combined tokens $\tilde{\boldsymbol{x}} = \{\tilde{\boldsymbol{x}}^{\text{vis}}, \boldsymbol{x}^{\text{inv}}\}$ with the visible tokens $\boldsymbol{x}^{\text{vis}}$ further encoded with the video encoder, we input $\tilde{\boldsymbol{x}}_{t-1} = \{\tilde{\boldsymbol{x}}^{\text{vis}}_{t-1}, \boldsymbol{x}^{\text{inv}}_{t-1}\}$ and $\tilde{\boldsymbol{x}}_{t} = \{\tilde{\boldsymbol{x}}^{\text{vis}}_{t}, \boldsymbol{x}^{\text{inv}}_{t}\}$ of two consecutive frames into a new $\text{Predictor}(\cdot)$ network, as formulated in Eq. 3 and Eq. 4, to generate a temporal heatmap $\tilde{\boldsymbol{s}}$. $\text{Predictor}(\cdot)$ re-evaluates the contribution of the tokens that are not masked and gives rise to higher weights to the significantly contributed tokens. As the masked regions also gain higher weights from the $\text{Predictor}(\cdot)$ during masking, we hypothesize that the tokens of higher weights in the heatmap $\tilde{\boldsymbol{s}}$ demonstrate stronger correlations with the masked regions. Therefore, the heatmap $\tilde{\boldsymbol{s}}$ can be employed to enhance the remaining weak spatio-temporal information. For the masked regions, these invisible tokens now have no meaningful information and need to be reconstructed with additional guidance. Since the textual descriptions inherently contain spatio-temporal information in linguistic form, we leverage text features $\boldsymbol{y}$ to enhance the spatio-temporal representation of the invisible tokens $\boldsymbol{x}^{\text{inv}}$, particularly at positions with high weights. Specifically, we integrate semantic information with the overall tokens $\tilde{\boldsymbol{x}}$, and input these enhanced tokens into an MLP for reconstruction as follows:

$$\hat{\boldsymbol{x}} = \text{MLP}([\tilde{\boldsymbol{s}} * ((\tilde{\boldsymbol{x}}^{\text{vis}} + pos^{\text{vis}}) \oplus (\boldsymbol{y} + \boldsymbol{x}^{\text{inv}} + pos^{\text{inv}}))]), \quad (5)$$

where pos denotes the position embeddings that identify the token locations. Combining the position embedding with tokens gives rise to position-aware features and empowers better video reconstruction.

4. Experiments

4.1. Experiment Setup

Pre-training Datasets. Following methods [12, 30, 59], we conduct pre-training on two large-scale datasets, *i.e*, WebVid-2M [4] and Google Conceptual Captions 3M (CC3M) [41]. WebVid-2M contains 2.5M video-text pairs scraped from the web, while CC3M consists of 3M image-text pairs collected from web pages. The combination of these datasets provides rich visual-linguistic alignment signals from both static and dynamic visual content.

Downstream Datasets. To comprehensively evaluate our pre-trained model, we conduct experiments on three representative video-language tasks, including text-to-video retrieval, video question answering (VideoQA), and video captioning. Specifically, VideoQA can be categorized as multiple-choice (MC) and open-ended (OE) settings.

- **Video-text retrieval**: MSRVTT [56], DiDeMo [2] , and ActivityNet Caption [24].
- **VideoQA (MC)**: TGIF-Action, TGIF-Transition [21], and MSRVTT-MC [60].
- **VideoQA (OE)**: TGIF-Frame [21], MSRVTT-QA, MSVD-QA [55], and ActivityNet-QA [61].
- **Video captioning**: MSRVTT [56] and MSVD [6].

Implementation Details. We implement STM using PyTorch and conduct experiments on 8 NVIDIA A100 GPUs. For the model architecture, we adopt ViT-B/16 [10] and BERT-base [9] as our visual and textual encoders, respectively. To establish a robust foundation for spatio-temporal understanding, we initialize the visual encoder with weights from UMT [30], which is pre-trained on Kinetics-710 [29]. For the masking strategy, we apply a 30% masking rate to both image tokens and video tokens. The model is trained on the combined WebVid-2M and CC3M datasets for 10 epochs, with batch sizes of 256 for video and 1,024 for image inputs, using the AdamW optimizer [34] with a base learning rate of 1e-4. To enable gradient propagation through the predictor network, we employ the Gumbel-Softmax technique when selecting tokens with the highest scores for masking.

Table 1. Zero-shot text-to-video retrieval Recall@1 results on MSRVTT, DiDeMo, Activitynet (ANet) and MSVD.

Method	#Pairs	MSRVTT	DiDeMo	ANet	MSVD
Frozen [3]	5M	18.7	20.2	-	-
VIOLETv2 [12]	5M	21.9	24.4	-	-
Singularity [26]	5M	28.4	**36.9**	**30.8**	-
UMT [30]	5M	29.6	33.4	28.3	36.2
STM (ours)	5M	**29.8**	34.4	30.6	**38.7**
Models pre-trained on more data					
VIOLET [11]	138M	25.9	23.5	-	-
OmniVL [44]	17M	34.6	33.3	-	-
VINDLU[8]	25M	32.0	36.9	30.9	-
CLIP4Clip [35]	400M	30.6	-	-	36.2
InternVideo [49]	646M	40.7	31.5	30.7	43.4

4.2. Comparison

We evaluate STM against recent state-of-the-art video-language models across three downstream tasks under both zero-shot and fine-tuning settings.

Zero-shot Text-to-video Retrieval. As shown in Table 1, STM achieves superior performance compared to state-of-the-art models trained on datasets of comparable size, surpassing them by 0.2% and 2.5% on MSRVTT and MSVD, respectively. While models like InternVideo [49] and OmniVL [44] achieve leading results on MSRVTT and DiDeMo through training on massive datasets (646M and

Table 2. Text-to-video retrieval on MSRVTT, DiDeMo and AcitivityNet. #Pairs denotes the number of pre-training pairs. Models pre-trained with large-scale pairs are noted in gray.

Method	#Pairs	MSRVTT			DiDeMo			ActivityNet		
		R1	R5	R10	R1	R5	R10	R1	R5	R10
ClipBERT [25]	0.2M	22.0	46.8	59.9	20.4	48.0	60.8	21.3	49.0	63.5
Frozen [3]	5M	31.0	59.5	70.5	31.0	59.8	72.4	-	-	-
ALPRO [27]	5M	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-
BridgeFormer [13]	5M	37.6	64.8	75.1	37.0	62.2	73.9	-	-	-
Singularity [26]	5M	36.8	65.9	75.5	47.4	75.2	84.0	43.0	70.6	81.3
LAVENDER [31]	5M	37.8	63.8	75.0	47.4	74.7	82.4	-	-	-
VIOLETv2 [12]	5M	37.2	64.8	75.8	47.9	76.5	84.1	-	-	-
HiTeA [59]	5M	44.4	69.3	78.9	51.8	79.1	85.3	45.1	73.5	84.2
VINDLU [8]	5M	43.8	70.3	79.5	54.6	81.3	89.0	51.1	79.2	88.4
UMT [30]	5M	46.3	72.7	82.0	54.8	83.0	89.0	52.1	80.5	89.6
STM (ours)	5M	**48.5**	**74.6**	**83.8**	**56.9**	**84.1**	**89.7**	**53.6**	**82.1**	**90.4**
Models pre-trained on more data										
VIOLET [11]	186M	34.5	63.0	73.4	32.6	62.8	74.7	-	-	-
All-in-one [45]	138M	37.9	68.1	77.1	32.7	61.4	73.5	22.4	53.7	67.7
Clip4Clip [35]	400M	42.1	71.9	81.4	43.4	70.2	80.6	40.5	72.4	-
UCOFIA [51]	400M	49.4	72.1	-	46.5	74.8	-	45.7	76.0	-
UMT-L [30]	25M	58.8	81.0	87.1	70.4	90.1	93.5	66.8	89.1	94.9
VAST [7]	152M	63.9	-	-	72.0	-	-	70.5	-	-

Table 3. Video QA results on TGIF, MSRVTT and ActivityNet.

Method	#Pairs	TGIF			MSRVTT		ANet
		Act.	Trans	Frame	MC	QA	QA
ClipBERT [25]	0.2M	82.8	87.8	60.3	88.2	37.4	-
ALPRO [27]	5M	-	-	-	-	42.1	-
Clover [18]	5M	94.9	98.0	71.4	95.0	43.9	-
Singularity [26]	5M	-	-	-	92.0	42.7	41.8
VIOLETv2 [12]	5M	94.8	99.0	72.8	**97.6**	44.5	-
LAVENDER [31]	5M	96.6	**99.1**	72.2	96.6	44.2	-
HiTeA [59]	5M	96.8	98.8	72.5	97.2	45.4	**45.1**
UMT [30]	5M	-	-	-	95.9	44.3	43.5
STM (ours)	5M	**96.9**	99.0	**73.0**	96.8	**45.9**	44.9
Models pre-trained on more data							
JustAsk [57]	69M	-	-	-	-	41.5	38.9
MERLOT [62]	180M	94.0	96.2	69.5	90.9	43.1	41.4
All-in-one [45]	283M	95.5	94.7	66.3	92.3	46.8	-
MV-GPT [40]	53M	-	-	-	-	41.7	39.1
VIOLET [11]	186M	92.5	95.7	68.9	91.9	43.9	38.9
UMT-L [11]	25M	-	-	-	97.3	47.1	55.2

17M pairs respectively), STM demonstrates competitive performance despite using only 5M training pairs. This remarkable efficiency in leveraging limited training data underscores STM's robustness and learning capabilities.

Text-to-video Retrieval. Table 2 demonstrates the superior performance of STM model. Specifically, STM achieves state-of-the-art results with Recall@1 scores of 48.5% (+2.2%), 56.9% (+2.1%), and 53.5% (+1.5%) on MSRVTT, DiDeMo and ActivityNet datasets, respectively. Notably, when compared to models trained on larger-scale datasets, STM also exhibits comparable or even superior retrieval accuracy while using significantly less training data.

Video Question Answering. As shown in Table 3, although STM does not surpass the UMT-L model trained on large-scale data, it demonstrates competitive performance on video QA tasks across TGIF, MSRVTT, ANet and MSVD datasets. While HiTeA [59] achieves marginally better results on certain metrics due to its hierarchical event alignment approach, our method focuses on extracting global spatio-temporal information.

Video Captioning. For video captioning, shown in the right part of Table 4, STM achieves impressive CIDEr scores of 63.9 on MSRVTT and 145.4 on MSVD, producing captions that are both relevant and detailed. The robust generalization and efficient use of training data of STM underscore its potential as a powerful and data-efficient approach to diverse video understanding tasks, making it especially valuable in data-constrained settings.

4.3. Ablation Studies

To verify the contribution of each component, we conduct ablation studies on each component. All ablation experiments are trained on the WebVid-2M dataset for five epochs. Subsequently, we evaluate the model on three downstream tasks using the MSRVTT dataset.

Masking Strategies. Table 5 presents the results of various commonly used masking strategies on the MSRVTT,

Table 4. Video captioning results under CIDEr.

Method	#Pairs	MSRVTT	MSVD
SwinBERT [32]	5M	53.8	120.6
VIOLETv2 [12]	5M	58.0	139.2
LAVENDER [31]	5M	58.0	142.9
HiTeA [59]	5M	62.5	145.1
STM (ours)	5M	**63.9**	**145.4**
MV-GPT [40]	53M	60.0	-

Table 5. Different masking strategies results on MSRVTT dataset.

Masking Strategy	Retrieval	QA	Captioning
Random [16]	32.1	35.4	51.5
Tube-wise [47]	33.7	35.9	52.1
Running cell [39]	34.0	36.8	52.7
Frame-wise [14]	35.1	37.2	53.3
Spatial	37.4	39.0	53.5
Temporal	37.0	39.6	54.0
Spatio-temporal (ours)	**39.7**	**41.2**	**55.9**

Table 6. Results of different decoders on MSRVTT dataset. CM-Transformer denotes the cross-modality transformer.

Decoder Method	Retrieval	QA	Captioning
MLP [30]	35.2	38.7	51.0
Transformer	37.2	39.8	52.9
CM Transformer [12]	37.9	40.1	53.4
Spatio-temporal wo. text	38.9	39.0	53.1
Spatio-temporal w. text (ours)	**39.7**	**41.2**	**55.9**

i.e., random masking [16], tube-wise masking [47], running cell masking [39], frame-wise masking [14], spatial masking, temporal masking and our spatio-temporal masking. Compared to spatial-only masking, incorporating temporal masking leads to significant accuracy gains, highlighting the effectiveness of STM. We attribute this improvement to the spatio-temporal masking strategy to reconstruct these masked spatio-temporal features. During the pertaining process, the encoder is compelled to learn crucial spatio-temporal information in videos, thereby enhancing the model's ability to comprehend video contextual information.

Decoder Design. Table 6 compares the performance of four commonly used decoder methods on the MSRVTT dataset. The 2-layered MLP decoder [12], despite efficient, lacks the capacity to model complex dependencies. The Transformer decoder captures temporal patterns well, but may miss fine-grained spatial details. The cross-modality Transformer, which takes both text and video features as input, improves the cross-modal alignment but still struggles to capture the intricate spatio-temporal interactions. In contrast, our Spatio-temporal Decoder with text-guided effectively models both spatio-temporal and semantics dependencies, enhancing the model's ability to handle dynamic cross-modality content.

Table 7. Results of different loss functions on the MSRVTT.

$\mathcal{L}_{\text{VTM}}$	$\mathcal{L}_{\text{VTC}}$	$\mathcal{L}_{\text{MLM}}$	$\mathcal{L}_{\text{MVM}}$	Retrieval	QA	Captioning
✓				35.0	36.7	49.4
✓	✓			35.8	37.3	51.4
✓	✓	✓		36.1	38.2	52.9
✓	✓	✓	✓	**39.7**	**41.2**	**55.9**

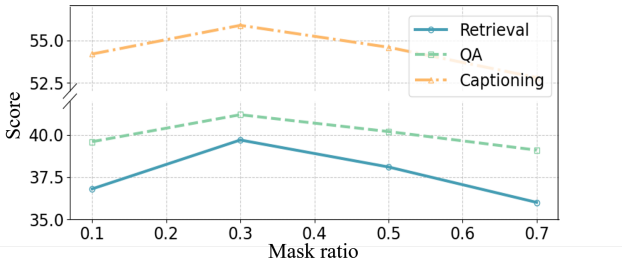


Figure 3. Different masking ratio results on MSRVTT dataset. Masking 30% tokens works best.

Loss Functions. Table 7 highlights the contributions of loss functions employed on MSRVTT. Each loss function plays a crucial role in enhancing the overall performance of the model, demonstrating the effectiveness of diverse training strategies. Incorporating $\mathcal{L}_{\text{MVM}}$ leads to the best performance in all tasks, highlighting the contribution of video feature reconstruction to represent video content in a fine-grained manner.

Masking Ratios. We investigate the impact of varying mask ratios on STM performance, as shown in Figure 3. Performance improves with increasing mask ratio, reaches the peak at 30%, then declines as the ratio rises further. This trend likely arises because a low mask ratio retains too much information, potentially causing information leakage and hindering the optimization of learning objective. In contrast, a high mask ratio may result in excessive information loss, making it challenging for the model to reconstruct patterns effectively, thus impairing the feature representation.

4.4. Analysis

To further investigate the underlying reasons for the effectiveness of STM, we conduct additional experiments.

Quality of Temporal Point Prediction. We employ a prediction network to forecast the locations of temporal points for masking. To assess the accuracy of these predictions, we visualize the masking of temporal point locations throughout the training process in Figure 4. For visualization, we replace the Gumbel-Softmax operation with a top-k selection to showcase the raw masking heatmap. During initialization (epoch 0), the masking is applied from the center of the image, which effectively covers most of the impor-

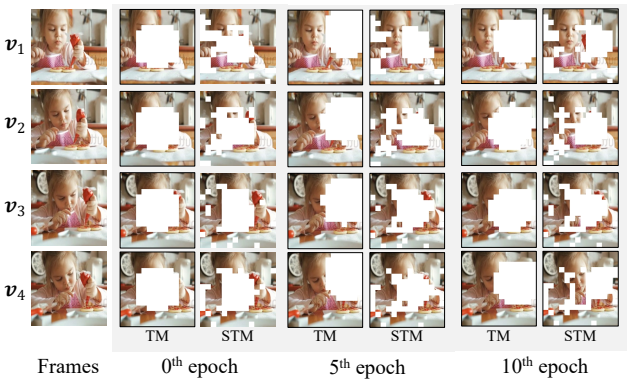


Figure 4. Visualization of the training process with temporal masking (TM) and spatio-temporal masking (STM), under the caption "A little girl cooks gingerbread in the kitchen." As training progresses, the STM mask positions gradually move closer to the critical interaction areas.

Table 8. Results of the consistent masked target and MVM target on MSRVTT dataset. "S/ST features" refer to reconstructing spatial/spatio-temporal features.

Masked Target	MVM Target	Retrieval	QA	Captioning
Random	S features	35.6	36.2	52.9
S masking	S features	37.3	38.4	53.1
ST masking	S features	36.7	37.9	52.3
ST masking	ST features	**39.7**	**41.2**	**55.9**

tant key information. Introducing spatial masking further enhances the capture of spatial information. As the training progresses, the temporal point predictions become increasingly accurate. By epoch 5, the model can reliably identify precise action locations (near the girl's hand). By epoch 10, the accuracy further improves, with the spatial masking mechanism enriching the masked content. Although the prediction network is learned in an unsupervised manner, the combination of center-initialized temporal points and spatial masking enables STM to effectively locate the spatiotemporal points of interest more accurately.

Consistent Masked Target and Reconstruction Target. The effectiveness of STM relies on the consistency between spatio-temporal masking and target reconstruction. To validate this consistency hypothesis, we remove temporal components from spatio-temporal masking and decoder, focusing solely on spatial feature reconstruction. These spatial features are supervised using CLIP [35] as a teacher model. Table 8 demonstrates that this modified approach outperformed the original image feature reconstruction. However, when we apply spatio-temporal masking to spatial feature reconstruction, we observe performance degradation. This decline can be attributed to the inconsistency between reconstruction target and masked target, resulting in confused attention patterns within the model.

Different Teacher Model. We further examine how different teacher models influence STM. Besides our chosen ActionCLIP [48], we also evaluate VideoMAEv2 [47] and InternVideo2 [50] as alternative teacher models, given their enhanced spatio-temporal perception capabilities for action recognition compared to ActionCLIP. The comparative results are presented in Table 9. Notably, employing VideoMAEv2 as the teacher model resulted in lower performance. We hypothesize that video understanding tasks depend more significantly on multimodal alignment abilities. Although VideoMAEv2 demonstrates superior action recognition performance compared to ActionCLIP, the latter's exceptional multimodal capabilities contribute to better overall performance in STM. Consequently, InternVideo2 also achieves improved results due to its extensive training with cross-modal data. Regarding to the higher computational resources consumed by InterVideo2, the ActionCLIP can achieve an optimal balance between performance efficacy and computational efficiency.

Table 9. Results using different teacher models on the MSRVTT.

Teacher Model	Params	Retrieval	QA	Captioning
VideoMAEv2 [47]	86M	39.0	39.1	52.8
InternVideo2 [50]	1B	**41.7**	**43.0**	**57.1**
ActionCLIP [48] (ours)	105M	39.7	41.2	55.9

5. Conclusion

In this paper, we demonstrate that inconsistency between reconstruction targets and masked targets may affect the performance of pretraining models. While maintaining consistency between the two targets allows the model to focus more effectively on the desired fine-grained features. Spatiotemporal features are crucial for understanding the context of videos. Therefore, we design a spatiotemporal masking strategy and a consistent spatiotemporal reconstruction decoder for the VLM. Our experiments across three downstream tasks in video understanding demonstrate improvements in all cases. We hope that these findings will inspire new ideas and approaches for other tasks when using mask learning. Large-scale pre-trained models and knowledge reasoning will be incorporated to improve VLM's video comprehension abilities in the future.

ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program of China under Grant 2023YFC2508704, in part by the National Natural Science Foundation of China: 62236008, 62022083, U21B2038, 62306092, 62441232, and in part by the Fundamental Research Funds for the Central Universities. The authors would like to thank Shufan Shen and the anonymous reviewers for their constructive comments and suggestions that improved this manuscript.

References

[1] Mona Ahmadian, Frank Guerin, and Andrew Gilbert. Mofo: Motion focused self-supervision for video understanding. *arXiv preprint arXiv:2308.12447*, 2023. 2

[2] 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

[3] 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. 5, 6

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 5

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[6] David Chen and WilliamB. Dolan. Collecting highly parallel data for paraphrase evaluation. *Meeting of the Association for Computational Linguistics,Meeting of the Association for Computational Linguistics*, 2011. 5

[7] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866, 2023. 6

[8] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10739–10750, 2023. 1, 2, 5, 6

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, 2019. 5

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 5

[11] 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, 3, 5, 6

[12] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. An empirical study of end-to-end video-language transformers with masked visual modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22898–22909, 2023. 1, 2, 3, 5, 6, 7

[13] 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, 6

[14] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10406–10417, 2023. 2, 7

[15] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13504–13514, 2024. 1, 2

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1, 2, 7

[17] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13493–13504, 2023. 2

[18] Jingjia Huang, Yinan Li, Jiashi Feng, Xinglong Wu, Xiaoshuai Sun, and Rongrong Ji. Clover: Towards a unified video-language alignment and fusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14856–14866, 2023. 6

[19] S Hwang, J Yoon, Y Lee, and SJ Hwang. Efficient video representation learning via masked video modeling with motion-centric token selection., 2022. 4

[20] Sunil Hwang, Jaehong Yoon, Youngwan Lee, and Sung Ju Hwang. Everest: Efficient masked video autoencoder by removing redundant spatiotemporal tokens. In *International Conference on Machine Learning*, 2024. 4

[21] 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

[22] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1

[23] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710, 2024. 1, 2

[24] 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

[25] 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. 6

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

[27] 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. 6

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *arXiv preprint arXiv:2211.09552*, 2022. 5

[30] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960, 2023. 1, 2, 3, 5, 6, 7

[31] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23119–23129, 2023. 1, 2, 6, 7

[32] 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

[33] Yuanze Lin, Chen Wei, Huiyu Wang, Alan Yuille, and Cihang Xie. Smaug: Sparse masked autoencoder for efficient video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2459–2469, 2023. 2

[34] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017. 5

[35] 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. 1, 5, 6, 8

[36] 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. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 638–647, 2022. 1

[37] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591, 2023. 2

[38] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15579–15591, 2023. 1

[39] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Xiang Wang, Yuehuan Wang, Yiliang Lv, Changxin Gao, and Nong Sang. Mar: Masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia*, 26:218–233, 2023. 2, 7

[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. 1, 6, 7

[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] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, and Chuang Gan. Masked motion encoding for self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2235–2245, 2023. 3

[43] Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, and Yu-Gang Jiang. Implicit temporal modeling with learnable alignment for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19936–19947, 2023. 4

[44] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems*, 35:5696–5710, 2022. 5

[45] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023. 6

[46] Junke Wang, Dongdong Chen, Chong Luo, Bo He, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Omnivid: A generative framework for universal video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18209–18220, 2024. 1, 2

[47] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 7, 8

[48] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3, 8

[49] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun

Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1, 2, 5

[50] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 1, 2, 8

[51] Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2816–2827, 2023. 6

[52] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 2

[53] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *European conference on computer vision*, pages 337–353. Springer, 2022. 2

[54] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[55] 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. 5

[56] 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. 5

[57] 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

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

[59] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15405–15416, 2023. 1, 2, 3, 5, 6, 7

[60] 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

[61] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, page 9127–9134, 2019. 5

[62] 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. 6

[63] Yizhou Zhao, Zhenyang Li, Xun Guo, and Yan Lu. Alignment-guided temporal attention for video action recognition. *Advances in Neural Information Processing Systems*, 35:13627–13639, 2022. 4