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.

Cluster-Wise Spatio-Temporal Masking for Efficient Video-Language Pretraining

Weijun Zhuang[1,2] Yuqing Huang[1,2] Weikang Meng[1,2] Xin Li[2*] Ming Liu[1]
Xiaopeng Hong[1] Yaowei Wang[1,2] Wangmeng Zuo[1*]
[1]Harbin Institute of Technology [2]Pengcheng Laboratory
{weijunzhuang.hit, domaingreen2, zacharymengwk, xinlihitsz, cswmzuo}@gmail.com,
csmliu@outlook.com, hongxiaopeng@hit.edu.cn, wangyw@pcl.ac.cn

Abstract

Large-scale video-language pretraining enables strong generalization across multimodal tasks but often incurs prohibitive computational costs. Although recent advances in masked visual modeling help mitigate this issue, they still suffer from two fundamental limitations: severe visual information loss under high masking ratios and temporal information leakage caused by inter-frame correlations. To address these challenges, we propose ClusterSTM, a Cluster-Wise Spatio-Temporal Masking strategy for efficient video-language pretraining. ClusterSTM first performs intra-frame clustering to partition visual tokens into multiple semantically independent clusters, then conducts cluster-wise masking by retaining the token with the highest temporal density within each cluster. Our masking strategy ensure that the retained tokens capture holistic video content while exhibit strong temporal correlation. Additionally, we introduce a video-text relevance reconstruction objective that aligns high-level multimodal semantics beyond conventional visual reconstruction. Extensive experiments across multiple benchmarks demonstrate that ClusterSTM achieves superior performance on video-text retrieval, video question answering, and video captioning tasks, establishing a new state-of-the-art among efficient video-language models. The code is available at https://github.com/illusion-vlm/ClusterSTM.

1. Introduction

Video-language models [5, 45, 48, 49] construct a semantic bridge between visual dynamics and textual descriptions through large-scale pretraining on massive video-text pairs, enabling flexible adaptation across diverse downstream tasks [18, 20, 29, 35, 53, 58, 65]. Although such large-scale pretraining endows these models with powerful generalization capabilities, it inevitably entails substantial computational costs. To mitigate this issue, recent studies [14, 27, 52] have investigated masked visual modeling strategies [16], leveraging its high data efficiency and low memory footprint, which has effectively advanced the research on efficient video-language pretraining.

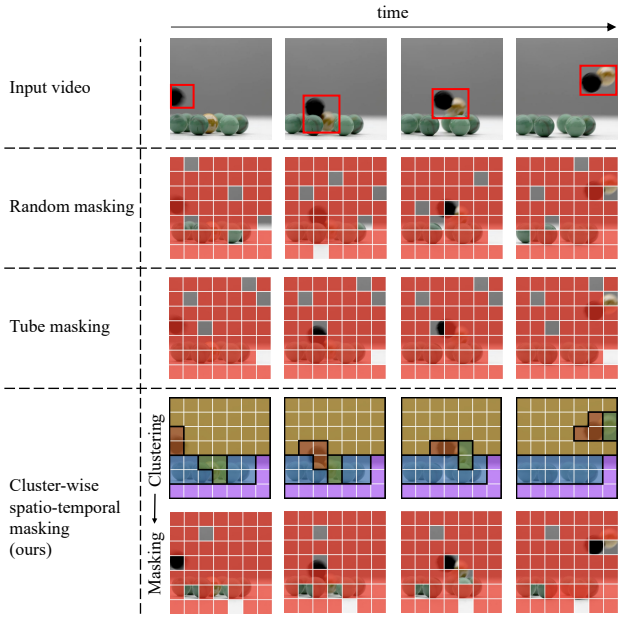

Figure 1. Comparison of different masking strategies. By selecting the token with the highest temporal density within each cluster, the cluster-wise spatio-temporal masking ensures strong temporal correlation among the retained tokens, thereby effectively mitigating the issue of temporal information leakage.

The success of masked image/video modeling in masked autoencoders [16, 43] largely relies on long pretraining schedules (e.g., 2400 epochs on 160K videos) to compensate for the severe loss of visual information caused by high masking ratios (e.g., 0.9). In contrast, large-scale video-

*Corresponding author.

language pretraining cannot afford such lengthy training, as models typically converge within only a few epochs (e.g., 10 epochs on 5M video-text pairs). Consequently, a key challenge for efficient video-language pretraining lies in designing an effective masking strategy that selectively retains visual tokens conducive to video-language alignment. To address this limitation, UMT [27] adopts a semantic masking strategy [17] that selectively retains foreground tokens with high visual semantics for alignment with textual descriptions. However, video-language alignment inherently depends on holistic scene understanding, as accompanying text descriptions usually encompass both foreground and background contents (e.g., *a child is flying a kite on the beach*). Consequently, semantic masking entirely overlooks the contribution of background tokens to vision-language alignment. Meanwhile, STM [52] mitigates the problem of severe visual information loss by significantly lowering the masking ratio (e.g., to 0.3) to preserve a large number of visual tokens; nevertheless, this strategy ultimately fails to achieve efficient pretraining.

Another challenge hindering the application of masked video modeling in large-scale video-language pretraining is the problem of information leakage caused by the inherent temporal correlation in video data. Unlike static images, video data contain an additional temporal dimension and exhibit unique properties of temporal correlation. Such temporal correlation may increase the risk of information leakage in the masking and reconstruction pipeline. As illustrated in Figure 1, random masking allows masked tokens to be easily reconstructed by referencing corresponding unmasked tokens from adjacent frames, which undermining the effectiveness of representation learning. The key idea for addressing this problem is to retain tokens that are temporally correlated. Tube masking [43] achieves this by applying an identical masking map across all frames, thereby preserving tokens at the same spatial locations throughout the video. However, this approach relies on the assumption of minimal motion occurring between adjacent frames, which limits its effectiveness in scenarios characterized by complex motion dynamics.

Building on the above analysis, a masking strategy suitable for efficient video-language pretraining should ensure that the retained tokens not only capture the holistic video content, including both foreground and background information, but also exhibit strong temporal correlation. To this end, we propose a novel masking strategy termed **ClusterSTM** (**Cluster**-Wise **S**patio-**T**emporal **M**asking), which effectively addresses the aforementioned challenges while maintaining a high masking ratio. Our masking strategy begins by performing intra-frame clustering over all visual tokens within each video frame, grouping them into multiple semantically independent clusters. Next, for each visual token in a video frame, we compute a temporal density based on its semantic distances to all tokens in adjacent frames. The temporal density of a target token reflects its temporal correlation with tokens in neighboring frames—intuitively, the more semantically similar tokens exist in adjacent frames, the higher the temporal density of the target token. Finally, we perform temporal-density-based cluster-wise masking: within each cluster, we retain the token with the highest temporal density and discard remaining tokens.

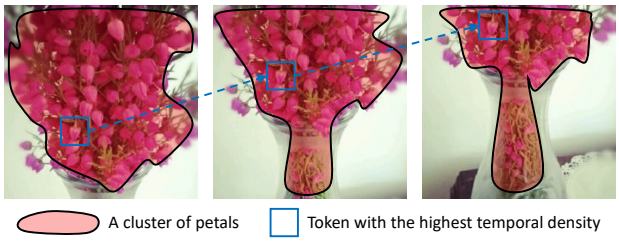

Figure 2. The token with the highest temporal density in a frame typically maintains the highest temporal density in subsequent frames, regardless of its spatial displacement, thereby ensuring that the retained tokens exhibit strong temporal correlation.

The cluster-wise masking mechanism in ClusterSTM retains one token for each semantically independent cluster, enabling the preserved tokens in a video frame to more comprehensively represent the overall visual content and thereby mitigating visual information loss. Meanwhile, the temporal-density-based token selection strategy ensures that the retained tokens exhibit strong temporal semantic correlation (as illustrated in Figure 2), effectively alleviating temporal information leakage. The key contributions of this paper are threefold:

- We present ClusterSTM, a novel masking strategy designed for efficient video-language pretraining. Through intra-frame clustering and subsequent cluster-wise masking, our approach effectively mitigates the severe loss of visual information encountered in prior approaches.
- We introduce a temporal-density-based cluster-wise masking mechanism that computes the temporal semantic density of each visual token and retains token with the highest density within each cluster, thereby ensuring that the preserved tokens exhibit strong temporal correlation.
- We conduct extensive experiments on multiple datasets, demonstrating that ClusterSTM consistently achieves superior performances across various downstream tasks, including text-video retrieval, video question answering, and video captioning.

2. Related Work

Video-Language Pretraining. Video-language pretraining refers to the process of training models that jointly learn representations from both video and text data, with the goal

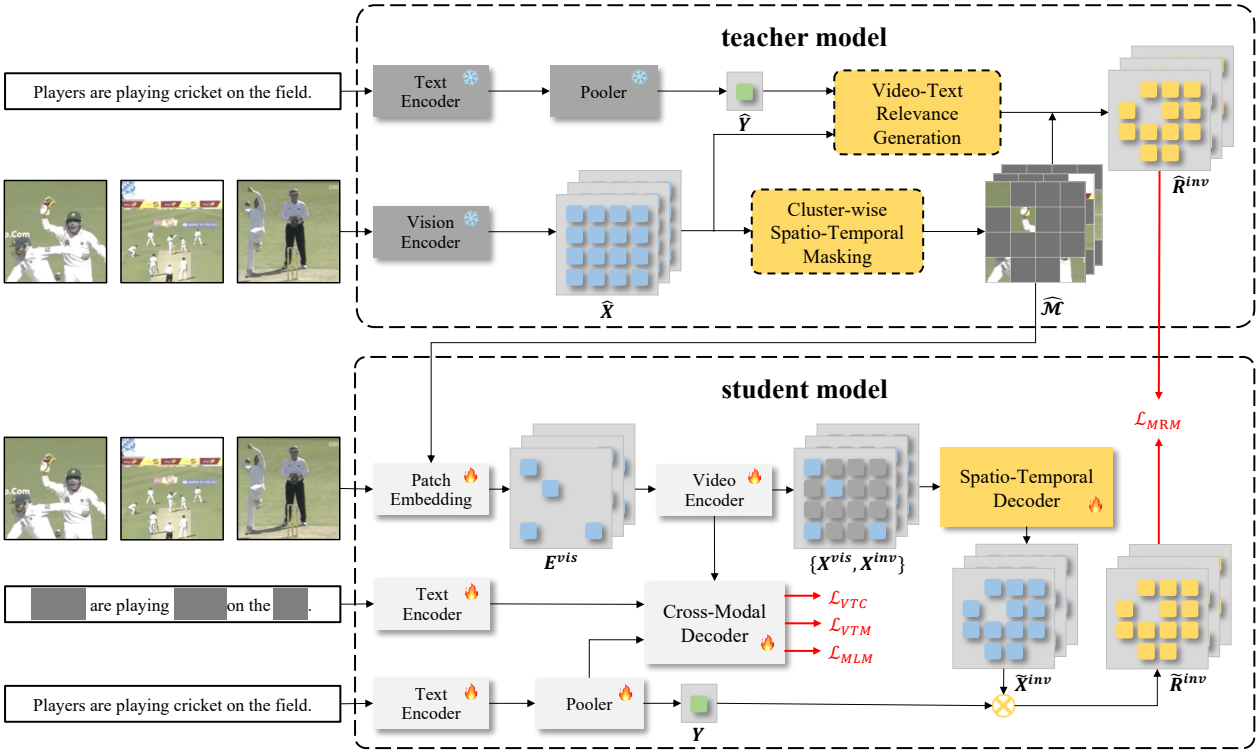

Figure 3. The ClusterSTM pipeline consists of two main components: the Cluster-wise Spatio-Temporal Masking strategy and the Video-Text Relevance Generation process. The Cluster-wise Spatio-Temporal Masking strategy first performs intra-frame clustering, followed by Temporal-Density-based Cluster-wise Masking. In this way, the retained tokens not only comprehensively capture the holistic content of each frame but also exhibit strong temporal semantic consistency. The Video-Text Relevance Generation process then produces fine-grained video-text relevance matrices, which serve as reconstruction targets for MRM loss computation.

of capturing the complex relationships between visual content and linguistic semantics. Recent advances in this field can be broadly categorized into two families: post-training from image-language models [20, 48, 49] and training from scratch [5, 46, 59]. The former initializes the model using the weights from well-learned image-language foundation models [25, 38], whereas the latter starts from random initialization. Despite this difference, both approaches rely on massive video-text pairs for pretraining and therefore face substantial memory and computational requirements. Recent works [23, 27, 28, 31, 36, 37, 52] have explored efficient video-language pretraining from different perspectives. Among them, Masked Autoencoder-based methods have demonstrated impressive efficiency gains by masking a large portion of tokens during pretraining. Although these methods have demonstrated impressive efficiency gains, they still suffer from severe visual information loss under high masking ratios as well as temporal information leakage caused by inherent temporal correlations in videos. Our method provides targeted solutions to these limitations by performing cluster-wise spatio-temporal masking.

Masked Visual Modeling. Following the success of BERT [10] in natural language processing, a growing line of work has explored masked visual modeling (MVM) for learning visual representations with Vision Transformers [11]. MAE [16] masks a subset of image patches and reconstructs their normalized pixel values, while BEiT [4] learns to predict discrete visual tokens obtained from a pre-trained tokenizer. Subsequent approaches have investigated a wide range of target signals for the mask-prediction task, including HOG features [51], deep features [2, 64], and frequency components [32, 54]. For spatio-temporal learning, several methods [12, 43, 47] designs specific MVM strategies tailored to the temporal characteristics of video data. However, despite their effectiveness, these approaches still fail to adequately address the issue of temporal information leakage caused by temporal correlation in video data. Our method tackles this challenge by leveraging temporal density to guide token selection within each cluster to ensure that the retained tokens exhibit strong temporal correlation, thereby effectively mitigates temporal leakage.

3. Method

This section introduces the core components of our proposed framework ClusterSTM, designed for efficient video-

language pretraining. Specifically, we detail the overall model architecture in Section 3.1, the cluster-wise spatio-temporal masking mechanism in Section 3.2, and the video-text relevance reconstruction in Section 3.3.

3.1. Model Overview

The overall architecture of our model is illustrated in Figure 3. We adopt a well-trained vision-language foundation model as the teacher model to generate spatio-temporal masks and video-text relevance matrices. Following UMT [27], the student model consists of a video encoder built upon a vanilla ViT [11] and a text encoder initialized with weights from BERT [10].

Given a video-text pair consisting of a video $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_t, ..., \mathbf{v}_T\}$ with T frames and a text sentence $\mathbf{S}$, the teacher model outputs the corresponding video tokens $\widehat{\mathbf{X}} = \{\widehat{\mathbf{x}}_1, ..., \widehat{\mathbf{x}}_t, ..., \widehat{\mathbf{x}}_T\} \in \mathbb{R}^{T\times N\times D}$, where N denotes the number of tokens per frame, and the pooled text feature $\widehat{\mathbf{Y}} \in \mathbb{R}^{1\times D}$. We then apply the cluster-wise spatio-temporal masking strategy to $\widehat{\mathbf{X}}$ to generate the spatio-temporal masks $\widehat{\mathcal{M}} \in \mathbb{R}^{T\times N}$ (see Section 3.2). Subsequently, the video-text relevance generation process takes $\widehat{\mathbf{X}}$ and $\widehat{\mathbf{Y}}$ as inputs and produces fine-grained video-text relevance matrices $\widehat{\mathbf{R}} = \{\widehat{\mathbf{r}}_1, ..., \widehat{\mathbf{r}}_t, ..., \widehat{\mathbf{r}}_T\} \in \mathbb{R}^{T\times N}$, which serve as the reconstruction targets (see Section 3.3).

For the student model, we first feed the video $\mathbf{V}$ into a patch embedding module to obtain video embeddings $\mathbf{E} \in \mathbb{R}^{T\times N\times D}$. Guided by the spatio-temporal masks $\widehat{\mathcal{M}}$ generated by the teacher model, we select the visible video embeddings $\mathbf{E}^{vis}$ and input them into the video encoder to produce the visible video tokens $\mathbf{X}^{vis}$. We then combine the visible and masked (invisible) tokens, denoted as $\{\mathbf{X}^{vis}, \mathbf{X}^{inv}\}$, and feed them into a spatio-temporal decoder, composed of multiple transformer layers similar to those in the video encoder, to reconstruct the full sequence of video tokens $\widetilde{\mathbf{X}}$. For the text sentence $\mathbf{S}$, we pass it through the text encoder to obtain the pooled text feature $\mathbf{Y}$. Finally, we compute the reconstructed video-text relevance matrices $\widetilde{\mathbf{R}}$ by multiplying the reconstructed invisible video tokens $\widetilde{\mathbf{X}}^{inv}$ with the text features $\mathbf{Y}$.

3.2. Cluster-Wise Spatio-Temporal Masking

Our masking strategy aims to ensure that the retained visual tokens not only effectively represent the holistic visual content of each frame but also preserve those tokens with strong semantic correlation temporally, therefore effectively addresses the severe visual information loss and temporal information leakage issues encountered when applying masked video modeling to large-scale video-language pretraining. In the following, we detail our proposed cluster-wise spatio-temporal masking method.

Intra-Frame Clustering. To ensure that the retained visual tokens effectively represent the holistic visual content of each frame, we need to distinguish different semantic regions within the frame. Therefore, we perform intra-frame clustering on each video frame. Specifically, the number of clusters for each frame is defined as $N_c = N \times (1 - r)$, where r denotes the masking ratio and N is the number of visual tokens per frame. Given the output visual tokens $\widehat{\mathbf{x}}_\mathbf{t}$ at frame t from teacher model, we employ an off-the-shelf clustering algorithm, the Density Peaks Clustering (DPC) algorithm [39], to partition $\widehat{\mathbf{x}}_t$ into N_c clusters:

$$\mathbb{C}_t = DPC(\widehat{\mathbf{x}}_t, N_c), \tag{1}$$

Temporal-Density-based Cluster-Wise Masking. Although randomly selecting one token from each semantically independent cluster can effectively preserve the holistic visual content of a video frame, this simple strategy fails to mitigate temporal information leakage. To address this issue, we propose a fine-grained masking mechanism based on temporal density. Specifically, we first define the semantic distance between two vectors as:

$$d(\mathbf{u}_i, \mathbf{u}_j) = 1 - \frac{\mathbf{u}_i \cdot \mathbf{u}_j}{||\mathbf{u}_i||_2 ||\mathbf{u}_j||_2}, \tag{2}$$

To capture temporal semantic consistency, we refine the local density computation in DPC [39] and define a temporal density as follows:

$$\rho(\widehat{\mathbf{x}}_{t,n}) = \sum_{\substack{i=1 \\ i\neq t}}^{T} \sum_{j=1}^{N} \exp^{-d(\widehat{\mathbf{x}}_{t,n}, \widehat{\mathbf{x}}_{i,j})/d_c}, \tag{3}$$

where $\widehat{\mathbf{x}}_{t,n} \in \mathbb{R}^{1\times D}$ is the n-th video token of frame t, and $d_c \in \mathbb{R}^+$ is a cutoff distance. Finally, for frame t, we retain the token with the highest temporal density within each cluster and discard the remaining tokens, thereby generating the spatio-temporal mask $\widehat{\mathcal{M}}_t$.

Although the spatial position of the token with the highest temporal semantic density may shift across subsequent frames, it typically maintains the highest temporal density and thus remains consistently preserved. As a result, our masking strategy ensures that the retained tokens exhibit strong temporal correlation, effectively mitigating the issue of temporal information leakage.

3.3. Video-Text Relevance Reconstruction

Previous video-language pretraining approaches [14, 42] have shown that reconstructing high-level spatial features is more advantageous for cross-modal alignment than reconstructing low-level pixels. STM [52] further demonstrates the effectiveness of reconstructing even higher-level spatio-temporal representations. However, these methods restrict the reconstruction targets to the visual modality alone, overlooking the collaborative role of the text modality in cross-modal alignment. To address this limitation, we introduce a

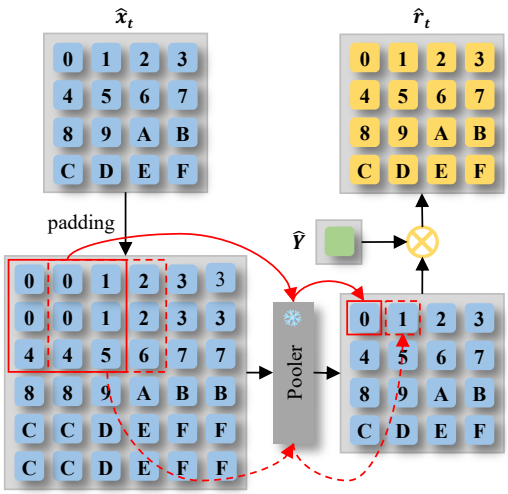

Figure 4. A schematic illustration of the Video-Text Relevance Generation process. The module first aggregates each target token with its neighboring tokens through a pooling operator to obtain an enhanced token. This enhanced token is then multiplied with the text feature to produce a high-quality video-text relevance matrix.

reconstruction target that operates at a higher semantic level than visual features and inherently carries multimodal properties, termed video-text relevance.

Video-Text Relevance Generation. Our process for generating high-quality video-text relevance matrices is illustrated in Figure 4. Since the teacher model is trained with vision-language contrastive learning, it primarily attends to global semantic information. Consequently, directly interacting each visual token with the text feature tends to produce low-quality relevance matrices. To enhance the quality of the relevance representations, we instead aggregate the neighboring tokens surrounding each target token to form a locally enhanced, information-rich token before interacting it with the text feature.

Specifically, we first apply border padding to the visual tokens $\widehat{\mathbf{x}}_t$. Then, we extract local groups of tokens through a sliding window and feed them into the teacher model's pooling module to generate fused visual tokens. Finally, each fused token is multiplied with the text feature to obtain the video-text relevance matrix $\widehat{\mathbf{r}}_t$. We adopt SigLIP [63] as the teacher model, as its flexible attentive pooling mechanism allows an arbitrary number of visual tokens to be processed, making it highly compatible with the design of our relevance generation process.

Masked Relevance Modeling. We enhance the model's ability to learn joint visual-textual representations by reconstructing the video-text relevance. The masked relevance modeling loss uses the teacher-generated relevance matrices $\widehat{\mathbf{R}}^{inv}$ as the reconstruction targets to supervise the reconstruction of $\widetilde{\mathbf{R}}^{inv}$ output by student model:

$$\mathcal{L}_{MRM} = L2(\widetilde{\mathbf{R}}^{inv}, \widehat{\mathbf{R}}^{inv}), \tag{4}$$

Table 1. Zero-shot text-to-video retrieval Recall@1 results on MSRVTT, DiDeMo, ActivityNet (ANet) and MSVD. #Pairs denotes the number of pretraining pairs, and #F indicates the number of video frames sampled during evaluation. The best results are indicated in **bold**. Models pretrained with massive datasets are noted in gray.

Method	#Pairs	#F	MSRVTT	DiDeMo	ANet	MSVD
Frozen [3]	5M	4	18.7	20.2	-	-
VIOLETv2 [14]	5M	5	21.9	24.4	-	-
Singularity [23]	5M	12	28.4	**36.9**	30.8	-
UMT [27]	5M	4	29.6	33.4	28.3	36.2
STM [52]	5M	4	29.8	34.4	30.6	38.7
ClusterSTM	5M	4	**31.2**	36.5	**31.4**	**40.3**
Models pretrained on more data						
VIOLET [13]	138M	5	25.9	23.5	-	-
OmniVL [44]	17M	8	34.6	33.3	-	-
VINDLU [8]	25M	4	32.0	36.9	30.9	-
CLIP4Clip [34]	400M	12	30.6	-	-	36.2
InternVideo [48]	646M	8	40.7	31.5	30.7	43.4

where $L2(\cdot)$ denotes the L2 distance.

Pretraining Objectives. In addition to the masked relevance modeling loss, following prior work [14, 27, 52], we incorporate video-text contrastive learning (VTC) and video-text matching (VTM) loss for cross-modal feature alignment, as well as masked language modeling (MLM) loss to supervise text reconstruction. The overall training objective during pretraining is:

$$\mathcal{L} = \mathcal{L}_{MRM} + \mathcal{L}_{VTC} + \mathcal{L}_{VTM} + \mathcal{L}_{MLM}. \tag{5}$$

4. Experiments

4.1. Experiment Setup

Pretraining Datasets. Following prior methods [14, 27, 52, 59], we pretrain our model on two large-scale datasets, WebVid-2M [3] and CC3M [41]. WebVid-2M is a web-curated video-text corpus comprising roughly 2.5M video-caption pairs collected from stock footage platforms, whereas CC3M contains 3M web-sourced image-text pairs with diverse, automatically filtered descriptions. The combination of these complementary video and image datasets provide broad and diverse visual-language supervision, enabling the model to jointly benefit from dynamic motion cues and rich static visual semantics.

Implementation Details. Our model is trained on 8 NVIDIA A100 GPUs using PyTorch. We adopt ViT-B/16 [11] without a class token as video encoder and initialize it with the pretrained weights from UMT [27], which is pretrained on Kinetics-710 [26]. We initialize the text encoder with weights from BERT-base [10].We adopt SigLIP-base [63] as the teacher model to leverage its flexible attention pooling mechanism to generate the relevance matrix.

Table 2. Text-to-video retrieval on MSRVTT, DiDeMo and AcitivityNet. #Pairs denotes the number of pretraining pairs. The best results are indicated in **bold**. Models pretrained with massive datasets are noted in gray.

Method	#Pairs	MSRVTT			DiDeMo			ActivityNet		
		R1	R5	R10	R1	R5	R10	R1	R5	R10
ClipBERT [22]	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 [24]	5M	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-
BridgeFormer [15]	5M	37.6	64.8	75.1	37.0	62.2	73.9	-	-	-
Singularity [23]	5M	36.8	65.9	75.5	47.4	75.2	84.0	43.0	70.6	81.3
LAVENDER [28]	5M	37.8	63.8	75.0	47.4	74.7	82.4	-	-	-
VIOLETv2 [14]	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 [27]	5M	46.3	72.7	82.0	54.8	83.0	89.0	52.1	80.5	89.6
STM [52]	5M	48.5	74.6	83.8	56.9	84.1	89.7	53.6	82.1	90.4
ClusterSTM	5M	**49.7**	**76.1**	**85.3**	**58.5**	**85.0**	**90.2**	**54.9**	**82.9**	**90.7**
Models pretrained on more data										
VIOLET [13]	138M	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 [34]	400M	42.1	71.9	81.4	43.4	70.2	80.6	40.5	72.4	-
UCOFIA [50]	400M	49.4	72.1	-	46.5	74.8	-	45.7	76.0	-
UMT-L [27]	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	-	-

For the masking strategy, we apply a 75% masking ratio to image tokens and a 90% masking ratio to video tokens. We apply random cluster-wise spatial masking to image tokens and cluster-wise spatio-temporal masking to video tokens. The cutoff distance d_c used in the temporal density computation is set to the top 20% quantiles of the semantic distances. The model is trained on the combined WebVid-2M and CC3M datasets for 10 epochs, with batch size of 1024 for video and 4096 for image, using the AdamW optimizer [33] with a base learning rate of 1e-4.

4.2. Quantitative Results and Comparison

To comprehensively evaluate our method, we benchmark it against recent state-of-the-art video-language models across three representative video-language tasks, including **video-text retrieval** (MSRVTT [56], DiDeMo [1], MSVD [6], and ActivityNet [21]), **video question answering** (TGIF-Action [19], TGIF-Transition [19], TGIF-Frame [19], MSRVTT-MC [60], MSRVTT-QA [60], ActivityNet-QA [61], and MSVD-QA [55]), and **video captioning** (MSRVTT [56] and MSVD [6]).

Zero-Shot Text-to-Video Retrieval. Table 1 reports the zero-shot text-to-video retrieval performance of our method across multiple datasets. Under a comparable pretraining data scale, ClusterSTM achieves remarkable improvements over state-of-the-art models, surpassing them by $\mathbf{1.4}\%$ on MSRVTT, $\mathbf{0.8}\%$ on ActivityNet, and $\mathbf{1.6}\%$ on MSVD. Although ClusterSTM attains performance comparable to Singularity [23] on DiDeMo, this comparison is skewed by the fact that Singularity samples 12 frames during evaluation, whereas ClusterSTM uses only 4 frames. When the number of campled frames is matched to 4, ClusterSTM outperforms STM [52] on DiDeMo by $\mathbf{2.1}\%$. Overall, these results demonstrate the strong generalization ability and efficiency of ClusterSTM across diverse retrieval benchmarks.

Text-to-Video Retrieval. Table 2 presents the finetuning results for text-to-video retrieval, where ClusterSTM significantly outperforms previous methods pretrained with a comparable data scale. Specifically, ClusterSTM achieves state-of-the-art results with Recall@1 scores of $49.7\%(+\mathbf{1.2}\%)$, $58.5\%(+\mathbf{1.6}\%)$, and $54.9\%(+\mathbf{1.3}\%)$ on MSRVTT, DiDeMo and ActivityNet datasets, respectively. Notably, despite using substantially less pretraining data, ClusterSTM achieves comparable or even superior performance to methods pretrained on massive datasets.

Video Question Answering. As shown in Table 3, ClusterSTM achieves significant performance improvements in the finetuning results for the video question answering task. Specifically, ClusterSTM outperforms STM [52] by $\mathbf{0.6}\%$ on TGIF-Frame, $\mathbf{0.8}\%$ on MSRVTT-QA, and $\mathbf{0.6}\%$ on ActivityNet-QA, which demonstrates its powerful capability of complex multimodal reasoning.

Video Captioning. Table 4 presents the experimental results on the video captioning task. ClusterSTM achieves state-of-the-art performance, with CIDEr scores

Table 3. Video QA results on TGIF, MSRVTT and ActivityNet. The best results are indicated in **bold**. Models pretrained with massive datasets are noted in gray.

Method	#Pairs	TGIF			MSRVTT		ANet-QA	MSVD-QA
		Act.	Trans	Frame	MC	QA		
ClipBERT [22]	0.2M	82.8	87.8	60.3	88.2	37.4	-	-
ALPRO [24]	5M	-	-	-	-	42.1	-	45.9
Singularity [23]	5M	-	-	-	92.0	42.7	41.8	-
VIOLETv2 [14]	5M	94.8	99.0	72.8	**97.6**	44.5	-	-
LAVENDER [28]	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 [27]	5M	-	-	-	95.9	44.3	43.5	49.1
STM [52]	5M	96.9	99.0	73.0	96.8	45.9	44.9	-
ClusterSTM	5M	**96.9**	**99.1**	**73.6**	97.1	**46.7**	**45.5**	**51.3**
Models pretrained on more data								
JustAsk [57]	69M	-	-	-	-	41.5	38.9	47.5
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	-	47.9
MV-GPT [40]	53M	-	-	-	-	41.7	39.1	-
VIOLET [13]	138M	92.5	95.7	68.9	91.9	43.9	38.9	47.9
UMT-L [27]	25M	-	-	-	97.3	47.1	47.9	55.2

Table 4. Video captioning results under CIDEr. The best results are indicated in **bold**. Models pretrained with massive datasets are noted in gray.

Method	#Pairs	MSRVTT	MSVD
SwinBERT [30]	5M	53.8	120.6
VIOLETv2 [14]	5M	58.0	139.2
LAVENDER [28]	5M	58.0	139.2
HiTeA [59]	5M	62.5	145.1
STM [52]	5M	63.9	145.4
ClusterSTM	5M	**64.4**	**145.6**
Models pretrained on more data			
MV-GPT [40]	53M	60.0	-

of 64.4(+**0.5**) on MSRVTT and 145.6(+**0.2**) on MSVD, demonstrating its strong capability in video understanding.

4.3. Ablation Studies

To evaluate the contribution of each component, we conduct ablation studies on each component. The ablation experiments are trained on a subset composed of half of the WebVid-2M and CC3M datasets, for a total of 5 epochs. We then evaluate the model on the MSRVTT dataset across three downstream tasks. Except for Table 9, we pretrain UMT [27] for 10 epochs on the full 5M corpus.

Masking Strategies. Table 5 presents the experimental results of different masking strategies on the MSRVTT dataset across three downstream tasks. Notebly, cluster-wise-S denotes randomly selecting one token to retain within each cluster, while cluster-wise-ST denotes retaining the token with the highest temporal density within each cluster-corresponding to our proposed method. Cluster-wise-S performs intra-frame clustering followed by cluster-wise masking, which substantially preserves the complete visual content of video frames. This enables the model to efficiently leverage the retained tokens for representation learning, leading to significant performance gains compared with frame-wise masking, random masking, and tube-wise masking. Furthermore, cluster-wise-ST addresses the issue of temporal information leakage by ensuring strong temporal correlation among the retained tokens, thereby further enhancing the model's ability to comprehend video contextual information.

Table 5. Experimental results of different masking strategies on MSRVTT dataset across three downstream tasks. The best results are indicated in **bold**.

Masking Strategy	Retrieval	QA	Captioning
frame-wise	33.9	34.6	50.8
random	37.5	38.5	52.2
tube-wise	38.1	39.7	53.3
cluster-wise-S	39.8	41.4	55.4
cluster-wise-ST	**41.3**	**42.7**	**56.2**

Masking Ratios. We investigate the impact of different masking ratios for image and video tokens, as shown in

Table 6. Experimental results of different masking ratios for image and video tokens on MSRVTT dataset across three downstream tasks. The best results are indicated in **bold**.

Image(%)	Video(%)	Retrieval	QA	Captioning
50	80	40.1	41.8	55.6
50	90	41.1	42.4	56.1
75	80	40.9	42.1	56.0
75	90	**41.3**	**42.7**	**56.2**
90	95	39.5	41.2	55.3

Table 7. Experimental results of different reconstruction targets on MSRVTT dataset across three downstream tasks. The best results are indicated in **bold**.

Reconstrction Target	Retrieval	QA	Captioning
pixels	36.3	37.4	52.1
HOG	38.5	39.5	53.7
features	40.2	41.5	55.3
video-text relevance	**41.3**	**42.7**	**56.2**

Table 8. Effect of different pretraining objectives on MSRVTT dataset across three downstream tasks. The best results are indicated in **bold**.

VTC	VTM	MLM	MRM	Retrieval	QA	Captioning
✓	-	-	-	36.8	37.8	51.2
✓	✓	-	-	38.2	38.9	52.7
✓	✓	✓	-	38.9	39.6	54.0
✓	✓	✓	✓	**41.3**	**42.7**	**56.2**

Table 6. Our model achieves the best performance when using masking ratios of 75% for image tokens and 90% for video tokens. Compared with UMT [27](with masking ratios of 50% for image tokens and 80% for video tokens) and STM [52](with a masking ratio of 30% for both image and video tokens), our model benefits from a more balanced masking strategy, allowing it to make more comprehensive use of the data and achieve more efficient pretraining.

Reconstruction Targets. As shown in Figure 7, we compare the experimental results obtained using different reconstruction targets. HOG denotes the Histogram of Oriented Gradients [9], and the features target is supervised by SigLIP [63]. The model performs the worst when the reconstruction target is low-level pixels. As the semantic level of the reconstruction target increases, the model performance improves. Benefiting from its high-level semantic attributes and multimodal nature, which better align with the cross-modal alignment task, the video-text relevance target enables the model to achieve the best performance.

Pretraining Objectives. Table 8 presents the experimental results illustrating the impact of different pretraining objectives on model performance. It can be observed that each pretraining objective contributes positively to performance improvement. In particular, MRM yields the most significant performance gain, demonstrating that its capability to align high-level multimodal semantic information greatly enhances the model's ability for joint visual-textual representation learning.

Table 9. Effectiveness of cluster-wise spatio-temporal masking strategy. The best results are indicated in **bold**.

Method	MSRVTT	DiDeMo	ANet	MSVD
UMT [27]	29.6	33.4	28.3	36.2
+ ClusterSTM	**30.4**	**34.7**	**30.2**	**38.5**

Effectiveness of Cluster-Wise Spatio-Temporal Masking strategy. To further validate the effectiveness of the cluster-wise spatio-temporal masking strategy, we replace the semantic masking in UMT [27] with our proposed cluster-wise spatio-temporal masking while keeping all other model components (with CLIP [38] as the teacher model) unchanged. We report the results under the zero-shot text-to-video retrieval setting, as shown in Figure 9. It can be observed that replacing the masking strategy leads to a significant performance improvement for UMT. Although a high masking ratio causes most visual tokens to be dropped, our masking strategy ensures that the retained tokens comprehensively capture the overall visual content, which is crucial for effective cross-modal alignment.

5. Conclusion

In this work, we introduced ClusterSTM, a novel Cluster-Wise Spatio-Temporal Masking strategy tailored for efficient video-language pretraining. By coupling intra-frame clustering with temporal-density-based cluster-wise masking, our approach simultaneously alleviates severe visual information loss and mitigates temporal information leakage—two crucial bottlenecks in masked video modeling. In addition, we proposed a video-text relevance reconstruction objective that enhances high-level multimodal alignment beyond conventional visual reconstruction. Through extensive experiments on multiple video-language benchmarks, ClusterSTM consistently achieved superior performance and strong generalization capability compared to existing approaches. We believe that the proposed cluster-wise spatio-temporal masking paradigm provides a new perspective for efficient multimodal representation learning and can inspire future work toward scalable, temporally coherent video-language foundation models.

Acknowledgments This work was supported by the National Natural Science Foundation of China (Nos. 62476148, 62501191), and the Guangdong Basic and Applied Basic Research Foundation (No. 2024A1515011292).

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] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International conference on machine learning*, pages 1298–1312. PMLR, 2022. 3

[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] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

[5] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025. 1, 3

[6] 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] 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. 5, 6

[9] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 886–893. Ieee, 2005. 8

[10] 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 American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 3, 4, 5

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4, 5

[12] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 3

[13] 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. 5, 6, 7

[14] 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, 4, 5, 6, 7

[15] 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. 6

[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, 3

[17] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and Sun-Yuan Kung. Milan: Masked image pretraining on language assisted representation. *arXiv preprint arXiv:2208.06049*, 2022. 2

[18] Yuqing Huang, Xin Li, Zikun Zhou, Yaowei Wang, Zhenyu He, and Ming-Hsuan Yang. Rtracker: Recoverable tracking via pn tree structured memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19038–19047, 2024. 1

[19] 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

[20] 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, 3

[21] 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

[22] 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, 7

[23] Jie Lei, Tamara Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507, 2023. 3, 5, 6, 7

[24] 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, 7

[25] 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. 3

[26] 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

[27] 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, 4, 5, 6, 7, 8

[28] 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. 3, 6, 7

[29] Xin Li, Yuqing Huang, Zhenyu He, Yaowei Wang, Huchuan Lu, and Ming-Hsuan Yang. Citetracker: Correlating image and text for visual tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9974–9983, 2023. 1

[30] 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

[31] 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. 3

[32] Hao Liu, Xinghua Jiang, Xin Li, Antai Guo, Yiqing Hu, Deqiang Jiang, and Bo Ren. The devil is in the frequency: Geminated gestalt autoencoder for self-supervised visual pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1649–1656, 2023. 3

[33] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5 (5):5, 2017. 6

[34] 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. 5, 6

[35] 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

[36] Weikang Meng, Yadan Luo, Liangyu Huo, Yaowei Wang, Xin Li, and Zheng Zhang. Nalaformer: Norm-aware linear attention for transformer models. *arXiv e-prints*, pages arXiv–2506, 2025. 3

[37] Weikang Meng, Yadan Luo, Xin Li, Dongmei Jiang, and Zheng Zhang. Polaformer: Polarity-aware linear attention for vision transformers. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3, 8

[39] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *science*, 344(6191):1492–1496, 2014. 4

[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. 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 the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 5

[42] Fangxun Shu, Biaolong Chen, Yue Liao, Shuwen Xiao, Wenyu Sun, Xiaobo Li, Yousong Zhu, Jinqiao Wang, and Si Liu. Masked contrastive pre-training for efficient video-text retrieval. *arXiv preprint arXiv:2212.00986*, 2022. 4

[43] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 1, 2, 3

[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. 1, 6, 7

[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. 3

[47] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14733–14743, 2022. 3

[48] 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, 3, 5

[49] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 1, 3

[50] 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

[51] 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. 3

[52] Yue Wu, Zhaobo Qi, Junshu Sun, Yaowei Wang, Qingming Huang, and Shuhui Wang. Video language model pretraining with spatio-temporal masking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8557–8567, 2025. 1, 2, 3, 4, 5, 6, 7, 8

[53] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 1

[54] Jiahao Xie, Wei Li, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Masked frequency modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2206.07706*, 2022. 3

[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. 6

[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. 6

[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. 7

[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. 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. 6

[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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9127–9134, 2019. 6

[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. 7

[63] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 5, 8

[64] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 3

[65] Weijun Zhuang, Bowen Dong, Zhilin Zhu, Zhijun Li, Jie Liu, Yaowei Wang, Xiaopeng Hong, Xin Li, and Wangmeng Zuo. Spatial-temporal saliency guided unbiased contrastive learning for video scene graph generation. *IEEE Transactions on Multimedia*, 2025. 1