An Empirical Study of End-to-End Video-Language Transformers with Masked Visual Modeling

Tsu-Jui Fu[†]*, Linjie Li[‡]*, Zhe Gan[‡], Kevin Lin[‡], William Yang Wang[†], Lijuan Wang[‡], Zicheng Liu[‡]
[†]UC Santa Barbara [‡]Microsoft
{tsu-juifu, william}@cs.ucsb.edu
{lindsey.li, zhe.gan, keli, lijuanw, zliu}@microsoft.com

Abstract

Masked visual modeling (MVM) has been recently proven effective for visual pre-training. While similar reconstructive objectives on video inputs (e.g., masked frame modeling) have been explored in video-language (VidL) pre-training, previous studies fail to find a truly effective MVM strategy that can largely benefit the downstream performance. In this work, we systematically examine the potential of MVM in the context of VidL learning. Specifically, we base our study on a fully end-to-end VIdeO-LanguagE Transformer (VIOLET) [15], where the supervision from MVM training can be backpropogated to the video pixel space. In total, eight different reconstructive targets of MVM are explored, from low-level pixel values and oriented gradients to high-level depth maps, optical flow, discrete visual tokens and latent visual features. We conduct comprehensive experiments and provide insights into the factors leading to effective MVM training, resulting in an enhanced model VIOLETv2. Empirically, we show VIOLETv2 pre-trained with MVM objective achieves notable improvements on 13 VidL benchmarks, ranging from video question answering, video captioning, to text-to-video retrieval.[1]

1. Introduction

Video, containing multiple modalities in nature, has been used as an epitome to test how AI systems perceive. Video-language (VidL) research aims at extending this ability to convey perception via language. Popular VidL tasks were introduced, such as text-to-video retrieval [29, 54, 75], video question answering [25, 74], and video captioning [6, 75]. Recent progresses in VidL learning mostly focus on VidL pre-training [49, 58, 83] with video-text matching [39, 79] and masked language modeling [10]. There have also been attempts on similar masked modeling on vision inputs. For example, masked frame modeling [39] aims to recover masked frame representations. However, the pre-extracted video features cannot be refined during pre-training, which may limit its effectiveness. More recently, VIOLET [15] designs an end-to-end video-language transformer and proposes to reconstruct discrete visual tokens for masked frame patches. Though showing some promises in recovering visual semantics, the performance improvements on downstream VidL tasks are still marginal.

Meanwhile, self-supervised visual pre-training has been proven highly effective by reconstructing the masked image patches through raw pixel values [21, 73], discrete visual tokens [3, 81], or visual-semantic features [70, 71]. However, they all only focus on the visual modality. It is unclear which variant of masked visual modeling (MVM) objectives can help VidL learning, especially given that the paired language inputs can already provide high-level semantics.

Motivated by this, we conduct a comprehensive study of MVM for VidL learning. As illustrated in Figure 1, we base our study on the fully end-to-end VIdeO-LangaugeE Transformer (VIOLET) [15], and study a broad spectrum of MVM targets, including RGB pixel values (Pixel), histogram of oriented gradients (HOG), depth maps (Depth), optical flow (Flow), discrete visual tokens (VQ), spatial-focused image features (SIF), temporal-aware video features (TVF), and mulitmodal features (MMF). During pre-training, we mask out some proportions of the video input along both spatial and temporal dimensions, and the model learns to recover the MVM targets for these masked patches. Equipped with another two standard pre-training tasks (*i.e.*, video-text matching and masked language modeling), we empirically verify the effectiveness of different MVM variants on downstream VidL tasks.

Our study reveals that: (i) spatial-focused image features (SIF) is the most effective MVM target on video-text inputs; and (ii) the effects of different MVM targets on downstream VidL tasks are not shared between video-text and image-text inputs. For example, SIF extracted from the same model brings a large drop on downstream VidL

[1]Code has been released at https://github.com/tsujuifu/pytorch_empirical-mvm.

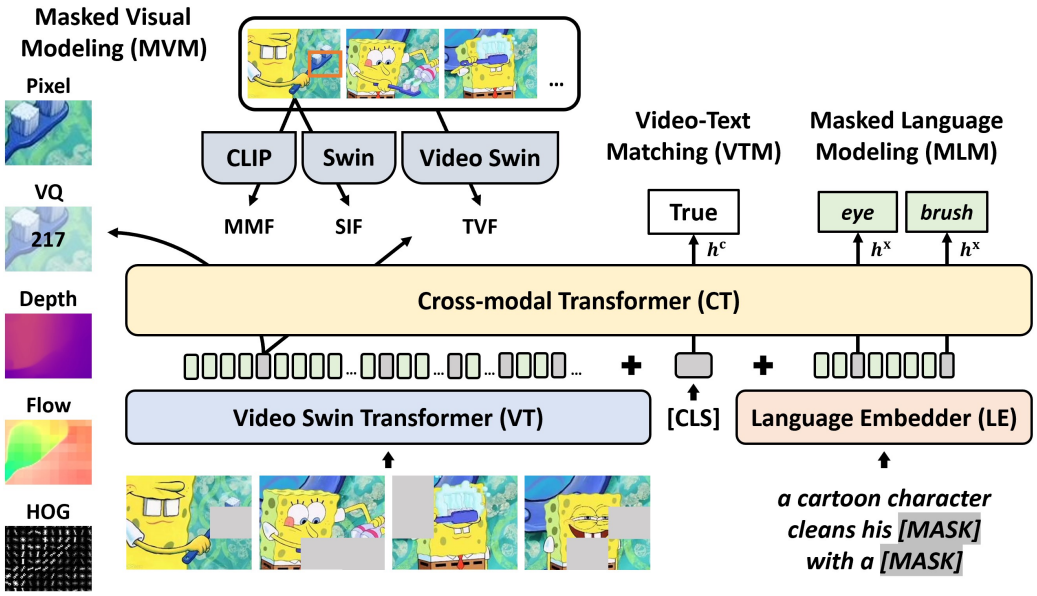

Figure 1. We systematically explore *eight* masked visual modeling (MVM) targets for end-to-end video-language (VidL) pre-training, including RGB pixel values (Pixel), histogram of oriented gradients (HOG), depth maps (Depth), optical flow (Flow), discrete visual tokens (VQ), spatial-focused image features (SIF), temporal-aware video features (TVF), and multimodal features from CLIP (MMF). Besides MVM, we pre-train VIOLET model [15] along with video-text matching (VTM) and masked language modeling (MLM).

performance when pre-trained with image-text pairs. In addition, we conduct comprehensive analyses of the masking strategy and ratio, combination of different MVM targets, to shed light on effective MVM training for VidL learning. We name the enhanced version of the original VIOLET [15] with the best MVM strategy as VIOLETv2.

Our contributions can be summarized as follows. We present an empirical study of masked visual modeling for video-language pre-training, with comprehensive analyses to reveal the ingredients for effective MVM training. VIOLETv2 with the best MVM recipe achieves strong performance on 13 VidL datasets. Concretely, compared to models pre-trained on the same 5M corpus, VIOLETv2 brings mean improvements of +5.4% accuracy on video question answering, +6.6% recall on text-to-video retrieval, and +11.4 CIDEr on video captioning. Direct comparison to VIOLET [15] also shows notable advantages of our model, even when pre-trained with much less data.

2. Related Work

Video-Language Understanding. Joint video-language (VidL) understanding [16, 17, 26, 32, 40, 43, 50] aims at interpreting the physical world via both vision and text perception. Researchers have explored such capability on VidL tasks including text-to-video retrieval [29, 37, 39, 54, 75], video question answering [25, 35, 36, 74], moment retrieval [19, 23, 29, 37], and video captioning [54, 69, 75, 82]. Prior arts before the large-scale pre-training era [13, 18, 32, 33, 36, 80] leverage offline extracted video features [1, 5, 9, 14, 22, 27, 30, 67, 72]. Later on, VidL pre-trained models [39, 49, 58, 83] built on the above pre-extracted features have shown promising results. To enhance the performance, there have been parallel interests in bringing in more modalities from raw video inputs [16, 42, 55] and end-to-end training [2, 34, 48, 79], aiming to elevate video representations.

Masked Visual Modeling (MVM). Aligned with the success of transformer-based [65] language pre-training [31, 44], image-text pre-training [7, 59] and video-text pre-training [28, 76, 77] have shown promising results on diverse vision-language (VL) tasks. Popular VL pre-training tasks include visual-text matching (VTM) and masked language modeling (MLM), which are directly adapted from language pre-training [10]. Similar masked modeling on visual inputs [7, 12, 39] has also been introduced to VL pre-training, but are not as useful. Among the literature of vision pre-training itself, MAE [21, 62] and SimMIM [73] reconstruct the pixels of the masked image patches to enhance visual representation. BEiT [3], iBOT [81], VIMPAC [60], and BEVT [68] adopt a BERT-like pre-training strategy to recover the missing visual tokens. On the other hand, MaskFeat [70] and MVP [71] consider latent features for MVM, including hand-crafted HOG features and image features extracted from pre-trained CLIP models [51]. Unlike previous studies exploring MVM on uni-modal data, in this study, we conduct a comprehensive investigation on how different MVM targets can help VidL learning.

The most relevant study to ours is VIOLET [15], which proposes to augment VidL pre-training with masked visual token modeling, while only showing marginal improvements on downstream performance. In contrast, our comprehensive investigation covers diverse MVM targets and

studies different combinations of masking strategies, which encompasses the design of MVM as well as shows large performance improvements on downstream VidL tasks.

3. Method

We first describe the base model VIOLET in Section 3.1, and then introduce the problem formulation of our investigation in Section 3.2. Section 3.3 discusses eight different target features for masked visual modeling (MVM).

3.1. End-to-End Video-Language Transformer

We conduct our empirical study using an end-to-end VIdeO-LanguagE Transformer (VIOLET) [15], with 3 components: Video Swin Transformer (VT), Language Embedder (LE), and Cross-modal Transformer (CT). VIOLET takes video $\mathcal{V}$ and sentence $\mathcal{X}$ as inputs. Sparse-sampled frames $\{f_1, f_2, ...\}$ from $\mathcal{V}$ are first segmented into a set of video patches, and then processed by VT to compute video features $v = \{v_1, v_2, ...\}$. LE extracts the word embeddings $w = \{w_1, w_2, ...\}$ for each word token $\{x_1, x_2, ...\}$ in $\mathcal{X}$. Then, CT performs cross-modal fusion on top of v and w to produce joint VidL representations $h = [h^v, h^c, h^x]$, where h^v, h^c, h^x denote the hidden representations of video patches, the special `[CLS]` token, and other word tokens.

3.2. Problem Setting

Given a large-scale video-language (VidL) dataset D, we aim to pre-train a VidL transformer to learn effective video-text representations. The learned representations can be transferred to downstream tasks for performance improvement. Different from existing works that focus on MVM for pure vision problems [3, 21, 81], we study MVM as a VidL pre-training task. Given a video-text pair $(\mathcal{V}, \mathcal{X})$ where $\mathcal{V}$ is a sequence of video frames and $\mathcal{X}$ is a sequence of word tokens. As shown in Figure 1, we randomly mask out some portions of the input frames $\mathcal{V}$, and learn to predict the target features corresponding to the masked patches. To output a correct prediction, the model will have to resort to other relevant video frames $\mathcal{V}$ and/or text tokens $\mathcal{X}$. This facilitates cross-modality learning for better VidL understanding.

In addition, we employ the commonly used VidL pre-training objectives, including video-text matching (VTM) and masked language modeling (MLM), where VTM aims to predict whether an input video-text pair is matched or not, while MLM aims to predict the masked word tokens from the surrounding context.[2] Our overall pre-training objective can be written as:

$$\mathcal{L} = \mathcal{L}_{\text{MVM}} + \mathcal{L}_{\text{VTM}} + \mathcal{L}_{\text{MLM}}, \quad (1)$$

where $\mathcal{L}_{\text{MVM}}$, $\mathcal{L}_{\text{VTM}}$, $\mathcal{L}_{\text{MLM}}$ are the MVM, VTM and MLM objectives, respectively.

[2]Refer to the Appendix for detailed formulation of VTM and MLM.

3.3. Target Features

Masked visual modeling (MVM) is a generic masked feature prediction task, where we mask out some of the visual input patches, and then predict the target features corresponding to the masked ones. Thus, a core design of MVM is the target features, which enables VIOLET learning a desired aspect of visual modeling. While MVM has been explored in pure vision tasks [3, 21, 70], it remains an open question whether MVM can facilitate the interactions between video and language modalities. In this study, we investigate *what design of MVM is effective in the context of video-language pre-training?*

Following [70, 73], we employ a simple linear layer or 2-layer MLP as the prediction head for MVM, to project the hidden video representations (h^v, of hidden size 768) from CT to the same dimension as the MVM targets. The default MVM loss is the l_1 loss, unless specified otherwise. Next, we introduce the considered target features in details.

RGB Pixel Values (Pixel). We treat the normalized RGB pixel values as a candidate target feature. During MVM, VIOLET learns to reconstruct the pixel values of the masked patches. The linear MVM head projects h^v into the same dimension as the raw video frame patch ($H \times W \times 3$).

Histogram of Oriented Gradients (HOG). HOG [8] is a pioneer feature descriptor that describes the gradients of orientations of the image. While HOG has been proven effective for visual pre-training [70], it is unknown whether it can benefit VidL pre-training. We extract HOG features in a dense grid level, and use such feature descriptors as the prediction targets of MVM. The HOG feature map is of the same size as the input video frame, but with channel size 1. The linear MVM prediction head projects h^v to the same dimension as HOG for the video frame patch ($H \times W \times 1$).

Depth Maps (Depth). Since depth maps usually contains finer-grained details of the object shapes and general scene layout of the foreground objects, it is worth exploring whether depth maps can be used to improve the scene/object understanding capability of a VidL pre-trained model. To obtain such MVM target, we employ a pre-trained dense prediction transformer (DPT) [53] to perform monocular depth estimation given an input video frame. The linear prediction head used for Depth is the same as the one for HOG, as both targets are of channel size 1.

Optical Flow (Flow). Optical flow is commonly used in motion analysis and video understanding. Here, we analyze whether apparent velocity of objects can benefit VidL pre-training. We employ a pre-trained recurrent all-pairs field transforms (RAFT) [61] to compute optical flow given the consecutive video frames. We directly use the estimated optical flow values as the prediction target, and supervise the MVM training with l_1 loss. To obtain the MVM predictions, we concatenate the hidden video representations

Pre-training Tasks	MVM Target	TGIF-Frame	DiDeMo-Retrieval			
		Acc.	R1	R5	R10	AveR
VTM+MLM	None	68.1	28.7	57.0	69.7	51.8
+MVM	RGB Pixel Values	68.3 (+0.2)	29.2 (+0.5)	58.6 (+1.6)	70.1 (+0.4)	52.6 (+0.8)
	Histogram of Oriented Gradients [8]	67.3 (-0.8)	26.6 (-2.1)	54.9 (-2.1)	68.1 (-1.6)	49.8 (-2.0)
	Depth Maps (DPT-L [53])	68.0 (-0.1)	27.3 (-1.4)	55.0 (-2.0)	68.3 (-1.4)	50.2 (-1.6)
	Optical Flow (RAFT-L [61])	67.6 (-0.5)	30.3 (+1.6)	58.0 (+1.0)	70.3 (+0.3)	52.9 (+1.1)
	Spatial-focused Image Features (Swin-B [45])	**68.8 (+0.7)**	**35.4 (+6.7)**	**62.4 (+5.2)**	**74.9 (+6.3)**	**57.6 (+5.8)**
	Temporal-aware Video Features (VidSwin-L [46])	68.0 (-0.1)	32.8 (+4.1)	60.5 (+3.5)	73.0 (+3.3)	55.4 (+3.6)
	Discrete Visual Tokens (DALL-E [52])	68.4 (+0.3)	28.1 (-0.6)	56.6 (-0.4)	69.4 (-0.5)	51.3 (-0.5)
	Multimodal Features (CLIP-ViT-B [51])	67.7 (-0.4)	29.8 (+1.1)	57.8 (+0.8)	68.5 (-1.2)	52.1 (+0.3)

Table 1. Comparing target features for MVM applied to video-text data. All variants are pre-trained on WebVid [2] for 5 epochs. Masking is performed randomly (RM) with ratio of 15%. The final pre-training setting is highlighted in gray.

computed by CT on consecutive frames, and employ a linear layer to project the concatenated video representations (of hidden size 768×2) to the same dimension as the estimated optical flow target for a given patch ($H \times W \times 2$).

Discrete Visual Tokens (VQ). In addition to continuous MVM targets, we also consider the discrete variational autoencoder (dVAE) [52, 64] to quantize video inputs. dVAE is learned to tokenize images into discrete visual tokens q from a finite dictionary, and then reconstruct the original visual scene based on q, where q should have a one-to-one correspondence with the input image patches spatially. We first adopt dVAE to tokenize the t^{th} video frame f_t into q_t: $q_t = \text{dVAE}(f_t)$, and then a 2-layer MLP is used to project h_v into the finite VQ vocabularies. As VQ token is discrete, we can model MVM with VQ as a classification problem, and adopt the cross-entropy loss to optimize the MVM training, following [3, 15].

Spatial-focused Image Features (SIF). We investigate whether image features can be useful for improving VidL pre-training. We employ a well-known vision transformer (such as Swin Transformer [45]) to extract the grid features given an input image. We then normalize the extracted grid features and consider them as ground-truth MVM targets. Likewise, we adopt a 2-layer MLP to project h_v to the same dimension as the image feature target.

Temporal-aware Video Features (TVF). We also study the impact of video features to VidL pre-training. We employ pre-trained video transformer (such as Video Swin Transformer [46]) to compute temporal-aware features for this analysis. Given a set of video frames, we use the transformer to extract video features in the form of space-time cubes, and then apply l_1 regression between normalized video features and MVM predictions from a 2-layer MLP head of the masked video patches.

Multimodal Features (MMF). We further study if the features learned via multimodal pre-training can benefit VidL pre-training. We utilize the vision branch of the ViT-Base backbone [11] in CLIP [51] to extract such multimodal features, and use the normalized features as the prediction targets in MVM pre-training. Again, we apply l_1 regression between the MVM predictions made via a 2-layer MLP head and the MMF targets for the masked patches.

In the following sections, we conduct comprehensive investigation over MVM targets described above, and perform detailed analysis on MVM strategies. To avoid confusion, we denote the strongest model with the most effective MVM training as VIOLETv2.

4. Study: Target Features for MVM

Settings. We conduct pre-training on WebVid-2.5M [2] for 5 epochs, and report accuracy on TGIF-Frame [25] for video question answering and R1/R5/R10/AveR on DiDeMo [24] for text-to-video retrieval.[3] We initialize our Video Swin Transformer (VT) with VideoSwin-Base [46], pre-trained on Kinetics-600 [27]. Language Embedder (LE) and Cross-modal Transformer (CT) are initialized from pre-trained BERT-Base [10]. During pre-training, we sparsely sample 4 video frames and randomly crop them into 224x224 to split into patches with $H = W = 32$. For all downstream tasks, we adopt the same video frame size and patch size but 5 sparse-sampled frames. We keep the training recipe (*e.g.*, optimizer settings, masking ratio, training schedule, *etc.*) consistent across all targets, which we find generally good in practice.[4] For MVM targets that involve a teacher model, we use official models released by the authors. We compare models pre-trained with 8 different MVM variants to the baseline pre-trained with only VTM and MLM. Our goal is to find the best MVM target features that can provide the largest performance improvement over this baseline. Results are summarized in Table 1. We first categorize the MVM targets into 4 groups, and discuss their performance in details.

One-stage Visual Targets. We include *Pixel* and *HOG*, as they do not require training a deep neural network in advance to extract these features. Compared to the baseline without MVM objective, regressing the explicit RGB colors contributes to a relatively small gain of +0.2% on TGIF-Frame and +0.8% on AveR for DiDeMo Retrieval. In

[3] We base our ablation experiments on these two representative datasets for fast iteration, our main results are reported on 13 benchmarks in Section 6. Details about downstream adaptation are included in the Appendix.

[4] Refer to the Appendix for more on training details.

MVM Targets	TGIF-Frame	DiDeMo-Retrieval			
	Acc.	R1	R5	R10	AveR
Pixel	68.3	29.2	58.6	70.1	52.6
Flow	67.6	30.3	58.0	70.3	52.9
SIF	**68.8**	**35.4**	**62.4**	**74.9**	**57.6**
SIF + Pixel	**68.8**	31.8	60.4	73.0	55.1
SIF + Flow	68.7	34.4	61.5	72.8	56.3

Table 2. Combining MVM targets. All variants are pre-trained on WebVid [2] for 5 epochs, using RM with 15% as the masking strategy. We highlight the final setting in gray.

Image Features Model	Train Data	IN-1K ACC@1	TGIF-Frame Acc.	DiDeMo-Retrieval R1	R5	R10	AveR
ResNet-50 [22]	IN-1K	76.1	67.3	29.1	58.1	69.3	52.2
Swin-T [45]	IN-1K	81.2	**68.9**	33.8	**63.6**	74.2	57.2
Swin-B	IN-1K	83.5	68.3	34.9	63.4	73.9	57.4
Swin-B	IN-22K	85.2	68.8	**35.4**	62.4	**74.9**	**57.6**
Swin-L	IN-22K	**86.3**	68.2	33.2	62.4	72.6	56.1

Table 3. Comparing different image feature targets for MVM. All variants are pre-trained on WebVid [2] with VTM+MLM+MVM (SIF) for 5 epochs, using RM with 15% as the masking strategy. The final pre-training setting is highlighted in gray.

contrast, HOG renders degradation on downstream video-language (VidL) performance (-0.8% on TGIF-Frame and -2.0% on DiDeMo-Retrieval). We hypothesize that this is due to the missing color information in HOG features, which is critical in VidL understanding.

Supervised Pseudo-label Targets. We include *Depth Maps (Depth)* and *Optical Flow (Flow)*. Intuitively, Depth and Flow can be considered as continuous pseudo "labels", which are made by models trained to perform depth and optical flow estimation [53, 61]. Depth does not improve over baseline with VTM+MLM. The nature of depth maps are to separate the foreground from the background, thus may guide the model to ignore information from the background, even when they are relevant for solving downstream VidL tasks (-0.1% on TGIF-Frame, -1.6% on DiDeMo Retrieval). Flow only focuses on the moving part between frames, while ignores the spatial details of static components, thus fail on more spatially-focused TGIF-Frame task (-0.5%). We also find that the optical flow estimation model easily fails with sparse sampling strategy, which is widely adopted in VidL pre-training.[5]

Supervised Visual Feature Targets. We include continuous features extracted from the last layers of image classification model [45] (*i.e.*, *Spatial-focused Image Features (SIF)*) and action recognition model [46] (*i.e.*, *Temporal-aware Video Features (TVF)*). We consider regressing supervised features from Swin-B or VidSwin-L[6] as a type of knowledge distillation from unimodal models to our model. SIF achieves significant improvement over baseline (+0.7% on TGIF-Frame and +5.8% on AveR for DiDeMo-Retrieval). In contrast, TVF fails to improve TGIF-Frame accuracy (-0.1%), though it brings notable improvement on retrieval performance (+3.6% on AveR). By distilling the knowledge from Swin-B, we enforce the model to focus more on spatial details of each frame, which we hypothesize is the main reason behind the large performance improvement. As previous study [4] pointed out, existing VidL benchmarks largely test on spatial understanding about the key frame of the video, with only a fractional of examples actually testing on temporal reasoning over multiple frames.

Self-supervised Multimodal Feature Targets. We use *Discrete Visual Tokens (VQ)* from DALL-E [52] and continuous *Multimodal Features (MMF)* extracted from CLIP [51]. Both are pre-trained on large-scale image-text datasets, usually much more expensive than all other targets. Both targets improve the performance by a slight margin on only one task. VQ that can capture patch-level semantics, benefits TGIF-Frame (+0.3%) which mostly focuses on scene understanding. While MMF from CLIP, contrastively pre-trained to measure the high-level similarity between the entire image and text sentence, is helpful for DiDeMo-Retrieval (+0.3% on AveR).

Summary. We hypothesize that many factors could lead to the low performance of an MVM target, such as its own characteristics (*e.g.*, local vs. global features); the target model; the loss design; or the mismatch between pre-train objectives and downstream focus. We try our best to compare them rigorously with controlled experiments to find the best setting. Based on our experiments, regressing RGB values (Pixel) and distilling features from Swin-B [45] (SIF) are the only two that produce consistent gains over the baseline on both downstream tasks. MVM with SIF achieves the best performance, with a gain of +0.7% on TGIF-Frame and +5.8% on AveR for DiDeMo-Retrieval over the baseline. Therefore, we use SIF as the default target for MVM in the following sections, unless specified otherwise.

5. Analyses of MVM

Combining MVM Targets. As different MVM targets focus on different aspects of visual modeling, a naive way to enable models with different visual capabilities is to combine them together. Specifically, the model pre-training can be supervised by more than one MVM loss, which is simply added together to be backpropagated. In Table 2, we find there is no merit in combining different MVM targets, leading to worse downstream performance than using SIF alone. When combining the best two targets found in Table 1: Pixel+SIF, it performs better than Pixel only, but does not improve over using SIF alone. We hypothesize that the explicit details of pixel values may conflict with the high-level visual semantics summarized in the grid features from the image classifier. We further try to combine SIF with Flow in the hope of enforcing both temporal and spatial reasoning over video inputs. In addition, Flow is a better candidate than other targets, as it demonstrates some advantages

[5] Please find visualization examples in the Appendix.
[6] VidSwin-L is trained on Kinetics-400 [27] with 83.1% accuracy.

Masking Strategy	Time Cost hours	TGIF-Frame Acc.	DiDeMo-Retrieval R1	R5	R10	AveR
RM	8.0	68.8	35.4	62.4	74.9	57.6
BM	8.0	**69.0**	35.9	63.3	74.6	57.9
AM	34.5	68.4	31.5	59.9	72.0	54.7
RM+BM	8.0	68.7	**36.4**	**64.2**	74.4	**58.3**
RM+AM	20.5	68.8	33.7	63.2	73.5	56.8
BM+AM	20.5	68.9	35.6	61.9	74.4	57.3
RM+BM+AM	17.0	68.6	34.7	62.0	**74.8**	57.2

Table 4. Impact of **masking strategy of MVM**. All variants are pre-trained on WebVid [2] with VTM+MLM+MVM (SIF) for 5 epochs. The masking ratio is set as 15% for all masking strategies. The final pre-training setting is highlighted in gray.

p_m	TGIF-Frame Acc.	DiDeMo-Retrieval R1	R5	R10	AveR
15%	68.8	35.4	62.4	74.9	57.6
30%	68.8	**36.2**	**64.0**	**74.5**	**58.2**
45%	**68.9**	35.6	61.9	74.4	57.3
60%	68.1	34.1	63.9	74.6	57.5
75%	68.3	35.4	62.4	74.2	57.3

Table 5. Impact of **masking ratio of MVM**. All variants are pre-trained on WebVid [2] with VTM+MLM+MVM (SIF) for 5 epochs, using RM as the masking strategy. The final pre-training setting is highlighted in gray.

on retrieval performance in Table 1, and it is a different type of target from SIF, compared to temporal-aware video features. The results are consistent, with improvements over optical flow only; while the performance drops, compared to SIF alone. Though our results are not encouraging, we believe how to effectively combine different MVM targets is an interesting direction for future study.

MVM Target Extractors *vs.* Downstream Performance. In Table 3, we explore different image classification models as the MVM target extractor for SIF, and investigate whether stronger image classification model enables better VidL performance. We compare ResNet-50 [22], Swin-Tiny/Base/Large [45], trained on ImageNet-1K (IN1K) or ImageNet-22K (IN-22K) [9], and summarize the observations below:

- ResNet-50 performs lower than Swin variants. Two potential reasons are (i) ResNet-50 architecture is very different from VidSwin (*i.e., with different inductive bias*); and (ii) the much lower ImageNet performance ($\sim$76 vs. $>$81) suggest the ResNet-50 features are not as strong.
- When the target model shares *similar inductive biases* to the video encoder (*i.e.*, Swin-T/B/L), the downstream performance is *not directly proportional to* ImageNet accuracy, and is overall better than that of Res50. This suggests that the architecture design of both target model and video encoder should be similar.
- A key difference between different Swin targets is the feature dimension (768/1024/1568 for Swin-T/B/L), while the video tokens from CT are of size 768. Although we project them into the same dimension as the targets, the mismatch may lead to slightly lower performance (with Swin-L especially).

In short, we believe a SIF target model should share similar inductive biases as the video encoder.

Masking Strategy. We investigate the effect of different masking strategies in Table 4, including random masking (RM), blockwise masking (BM), attended masking (AM), and their combinations.

- **Random Masking (RM).** Following the conventional practice in MLM, we randomly select a certain percentage p_m of video frame patches from the whole video inputs to be masked. In Table 5, we explore different masking ratios (p_m), and empirically find $p_m = 30\%$ gives the best downstream performance.
- **Blockwise Masking (BM).** To make MVM relying less on similar neighbor patches, we adopt blockwise masking [3, 60] that masks blocks of video patches along spatial-temporal dimension rather than independently masking randomly sampled patches for each frame. Specifically, we randomly sample an (H', W', T') as a masking block, where all $H' \times W'$ visual patches in the following T' consecutive frames will be masked; we repeat this process until $> p_m$ of video patches are masked to perform MVM pre-training.
- **Attended Masking (AM).** Attended masking tries to put more weights on the more important elements based on the attention weights computed by Cross-modal Transformer (CT). A similar idea has been explored in [79] for MLM. Here, we extend AM to both visual and textual modalities. We first keep the video-text inputs intact, feed them into CT to compute the attention weights, to decide which portions in video and text are more important. We then select the top p_m of most-attended patches/tokens to be masked in video-text inputs for MVM and MLM.

To combine different masking strategies, we randomly apply one masking method for each video-text pair in a batch. Results in Table 4 suggest that TGIF-Frame can slightly benefit from BM, and combining BM with RM leads to the best retrieval performance on DiDeMo. As video usually presents analogous visual patterns in spatial-temporal neighbors (*i.e.*, nearby patches within current frame or neighboring frames), when masking patches independently (*i.e.*, RM), these neighbors can make the masked patches easy to recover, and may lead to spurious success in MVM evaluation. By masking a block (*i.e.*, BM) instead of individual patches, the model cannot merely rely on similar neighboring visual cues but requires actual visual reasoning to recover a group of missing patterns. Combining BM with RM leads to more diverse dropout patterns in video inputs, which is in analogy to data augmentation.

In addition, AM and combinations with AM are not effective for both downstream tasks. It is also worth noting that AM greatly increase the training time (4 times more than RM/BM), due to the additional forward pass needed

Pre-training Tasks	MVM Target	TGIF-Frame	DiDeMo-Retrieval			
		Acc.	R1	R5	R10	AveR
ITM+MLM	None	**69.8**	**36.4**	64.3	74.7	**58.4**
+MVM	RGB Pixel Values	69.7 (-0.1)	35.8 (-0.6)	**64.4 (+0.1)**	74.9 (+0.2)	**58.4**
	Histogram of Oriented Gradients [8]	**69.8**	34.9 (-1.5)	**64.4 (+0.1)**	75.1 (+0.4)	58.1 (-0.3)
	Depth Maps (DPT-L [53])	69.6 (-0.2)	32.3 (-4.1)	63.8 (-0.5)	74.2 (-0.5)	56.9 (-1.5)
	Spatial-focused Image Features (Swin-B [45])	69.7 (-0.1)	31.6 (-4.8)	60.5 (-3.8)	72.5 (-2.2)	54.9 (-3.5)
	Discrete Visual Tokens (DALL-E [52])	**69.8**	34.4 (-2.0)	62.6 (-1.7)	75.1 (+0.4)	57.4 (-1.0)
	Multimodal Features (CLIP-ViT-B [51])	**69.8**	33.6 (-2.8)	62.9 (-1.4)	**75.6 (+0.9)**	57.4 (-1.0)

Table 6. Comparing target features for MVM applied to image-text data. All variants are pre-trained on CC3M [57] for 5 epochs. Masking is performed randomly (RM) with ratio of 15%.

Pre-training Tasks	MVM Target		TGIF-Frame	DiDeMo-Retrieval			
	WebVid2.5M	CC3M	Acc.	R1	R5	R10	AveR
VTM+MLM	None	None	69.7	36.7	66.5	76.6	59.9
+MVM	Spatial-focused Image Features (Swin-B [45])	None	71.1	38.8	**69.6**	**80.0**	**62.8**
	Spatial-focused Image Features (Swin-B)	Pixel	**71.3**	**39.7**	69.3	78.4	62.5

Table 7. Combining MVM target features for both video-text and image-text data. All variants are pre-trained on WebVid2.5M [2] +CC3M [57] for 5 epochs. The final pre-training setting is highlighted in gray.

to compute the attention weights. In our implementation, we optimize the three losses altogether in the same forward-backward pass. Hence, the performance drop with AM may be due to the important elements (*e.g.*, visual patches containing the main object or content words) are more likely to be masked together and leaving the less relevant elements (*e.g.*, scene background or stop words) intact, which will especially make the learning of video-text matching harder.

Applying MVM to Image-Text Data. As image can be considered as a special case of video with temporal size 1, video-language (VidL) pre-training can take advantages of image-text data, which has been proven successful in [2,34]. The current trend in VidL pre-training is to leverage both video-text data and image-text data. Therefore, we repeat the experiments in Section 4 and examine which MVM targets work the best on downstream VidL tasks, when pre-trained on image-text data only. We remove optical flow and temporal-aware video features from this study, as the inputs are static images. In Table 6, we pre-train our model on CC3M [57] for 5 epochs and report results on TGIF-Frame and DiDeMo-Retrieval. The performance trend with different MVM targets are not consistent with that observed on video-text data. Pixel is able to largely preserve the baseline (VTM+MLM) performance, while other MVM targets lead to different degrees of performance drop, especially on retrieval. Without visual implications from neighbor frames as video, MVM is more challenging to learn on image data. On the other hand, MVM over an image may easily fit in static visual representation, which could hurt video temporal reasoning and not benefit downstream VidL learning.

Combining Video-Text Data with Image-Text Data. We further follow recent VidL literature [2, 38] to use both video-text data and image-text data for pre-training, and investigate different ways to combine MVM targets on image and video data in Table 7. Note that we adopt the best training strategy found in the above investigations, that is, using spatial-focused image feature (SIF) as MVM target for video inputs, and using blockwise masking (BM) + random masking (RM) with masking ratio of 30% as the masking strategy. As the best MVM target (Pixel) on image data does not show improvement over the baseline without MVM objective in Table 6, we explore with/without MVM objective on images in this combined pre-training. For the baseline with VTM+MLM only, we simply remove the MVM objective on both image and video data, while keeping the rest training settings. Under the strict fair comparison, we observe adding MVM objectives contributes to $>$+0.4% gains on TGIF-Frame and $>$+2.6% increase on AveR for DiDeMo-Retrieval. Comparing with or without MVM objective on images, adding MVM on image-text brings minor performance difference (+0.2% on TGIF-Frame and degrades by -0.3% on AveR for DiDeMo-Retrieval) over MVM on video-text only. Therefore, in our final setting, we only apply MVM objective on video data.

6. Main Results

To this end, we combine the most effective MVM strategies to pre-train VIOLETv2 and evaluate on 13 video-language (VidL) tasks. Table 8 shows the comparison to prior arts on **video question answering (QA)** and **video captioning**. We observe that VIOLETv2 is effective in learning transferable knowledge for the downstream tasks. For example, considering pre-training data at a similar scale (*i.e.,* $\leq$ 5M, the top rows of Table 8), VIOLETv2 achieves better results than prior arts, including ALPRO [38], ClipBERT [34], and SwinBERT [41], across all considered video QA and video captioning benchmarks. Specifically, when pre-training with the exact same data (*i.e.*, WebVid2.5M [2] + CC3M [57]). VIOLETv2 surpasses ALPRO by 2.4% accuracy on MSRVTT-QA and 8.4% accuracy on MSVD-QA, respectively. We also compare with other models pre-trained on significantly larger scale of video-text pairs. As shown in the bottom rows of Table 8, although

Method	# Pretrain videos/images	TGIF [25]			MSRVTT [75]		LSMDC [63]		MSVD [6]	Captioning	
		Act.	Trans.	Frame	MC [78]	QA [74]	MC	FiB	QA [74]	MSRVTT	MSVD
ClipBERT [34]	0.2M	82.8	87.8	60.3	88.2	37.4	-	-	-	-	-
ALPRO [38]	5M	-	-	-	-	42.1	-	-	46.3	-	-
SwinBERT [41]	-	-	-	-	-	-	-	-	-	53.8	120.6
Models pre-trained on more data											
JustAsk [76]	69M	-	-	-	-	41.5	-	-	46.3	-	-
MERLOT [79]	180M	94.0	96.2	69.5	90.9	43.1	81.7	52.9	-	-	-
All-in-one [66]	283M	95.5	94.7	66.3	92.3	46.8	84.4	-	48.3	-	-
MV-GPT [56]	53M	-	-	-	-	41.7	-	-	-	60.0	-
VIOLET [15]	186M	92.5	95.7	68.9	91.9	43.9	82.8	53.7	47.9	-	-
VIOLETv2	5M[7]	**94.8**	**99.0**	**72.8**	**97.6**	**44.5**	**84.4**	**56.9**	**54.7**	**58.0**	**139.2**

Table 8. Comparison with SOTA on **video question answering** (accuracy) and **video captioning** (CIDEr). VIOLETv2 is pre-trained on WebVid2.5M [2]+CC3M [57] with VTM+MLM+MVM (SIF on videos) for 10 epochs. We gray out methods that use significantly more pre-training data.

Method	# Pretrain videos/images	MSRVTT [75]			DiDeMo [24]			LSMDC [54]		
		R1	R5	R10	R1	R5	R10	R1	R5	R10
ClipBERT [34]	0.2M	22.0	46.8	59.9	20.4	48.0	60.8	-	-	-
Frozen [2]	5M	31.0	59.5	70.5	31.0	59.8	72.4	15.0	30.8	39.8
ALPRO [38]	5M	33.9	60.7	73.2	35.9	67.5	78.8	-	-	-
BridgeFormer [20]	5M	**37.6**	**64.8**	75.1	37.0	62.2	73.9	17.9	35.4	44.5
Models pre-trained on more data										
HERO [39]	136M	16.8	43.4	57.7	-	-	-	-	-	-
All-in-one [66]	138M	37.9	68.1	77.1	32.7	61.4	73.5	-	-	-
Clip4Clip [47]	400M	42.1	71.9	81.4	43.4	70.2	80.6	21.6	41.8	49.8
VIOLET [15]	186M	34.5	63.0	73.4	32.6	62.8	74.7	16.1	36.6	41.2
VIOLETv2	5M	37.2	**64.8**	**75.8**	**47.9**	**76.5**	**84.1**	**24.0**	**43.5**	**54.1**

Table 9. Comparison with SOTA on **text-to-video retrieval** tasks (R1/5/10). We gray out methods that use significantly more pre-training data.

we use less pre-training data than others, VIOLETv2 still achieves comparable or better performance.

We observe similar findings on video captioning. On MSRVTT captioning, VIOLETv2 is only 2 points behind MV-GPT [56] pre-trained with 53M video-text pairs, which is 10 times larger than ours (5M). In addition, MV-GPT leverages ASR transcripts to enhance the captioning performance, while our captioning model takes only video frames as inputs and outputs the video caption.[8] We believe augmenting VIOLETv2 with additional modalities, such as audio or ASR transcripts, can further improve captioning performance, which we leave as future work.

Table 9 presents the comparison on **text-to-video retrieval**. When pre-training with the same datasets (*i.e.*, WebVid2.5M [2] + CC3M [57]), VIOLETv2 shows across-the-board improvements with all metrics considered on DiDeMo and LSMDC. It is worth noting that our method performs comparably to BridgeFormer [20] on MSRVTT-Retrieval. BridgeFormer adopts a noun/verb masking strategy during pre-training, which is specially aligned to the simple sentences in MSRVTT. However, it cannot show similar effects on DiDeMo and LSMDC due to more complex texts with multiple nouns/verbs. In contrast, the studied MVM can achieve a comprehensive enhancement in VidL learning and lead to notable improvements (+10.9% R1 on DiDeMo and +6.1% R1 on LSMDC).

[7]The SIF target model Swin-B is trained on IN-22K.

[8]Details about downstream finetuning on captioning are in Appendix.

Direct Comparison to VIOLET [15]. Across Table 8 and 9, it is worth noting that VIOLETv2 outperforms VIOLET with notable margins, even when VIOLET is pre-trained with significantly more data (about 37 times more). Specifically, VIOLETv2 yields an average gain of +3.4% across 8 video QA datasets, and an absolute gain of +8.6% on R1 across all three retrieval benchmarks. These results suggest the importance of an appropriate MVM setting, which is the core belief in our study.

7. Conclusion

We initiate the first empirical study on adopting masked visual modeling (MVM) for video-language (VidL) learning. We explore diverse MVM objectives upon end-to-end VIdeO-LanguagE Transformer (VIOLETv2), including low-level pixel space, high-level visual semantics, and extracted latent features. Our results show that VIOLETv2 pre-trained on 5M video/image-text data with MVM objective achieves strong performance on 3 popular VidL tasks across 13 VidL benchmarks. Our comprehensive analyses on different combinations of MVM targets, various SIF target extractors, and varying masking strategies/ratios shed light on effective MVM design. We believe our study can guide future research on large-scale VidL pre-training and wish to study how MVM can generalize to larger-scale data. In addition, we vision that with the emergence of video/VidL foundation models in future works, better choices of MVM targets can be explored.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 4, 5, 6, 7, 8

[3] Hangbo Bao, Li Dong, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *International Conference for Learning Representations (ICLR)*, 2022. 1, 2, 3, 4, 6

[4] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "Video" in Video-Language Understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 5

[5] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[6] David L. Chen and William B. Dolan. Collecting Highly Parallel Data for Paraphrase Evaluation. In *ACL*, 2011. 1, 8

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[8] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 3, 4, 7

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a Large-Scale Hierarchical Image Database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2, 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 1, 2, 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference for Learning Representations (ICLR)*, 2021. 4

[12] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An Empirical Study of Training End-to-End Vision-and-Language Transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[13] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[15] 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. In *arXiv:2111.1268*, 2021. 1, 2, 3, 4, 8

[16] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[17] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. Vision-Language Pre-training: Basics, Recent Advances, and Future Trends. In *Foundations and Trends in Computer Graphics and Vision*, 2022. 2

[18] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-Appearance Co-Memory Networks for Video Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[19] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal Activity Localization via Language Query. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[20] Yuying Ge, Yixiao Ge, Xihui Liu, Dian Li, Ying Shan, Xiaohu Qie, and Ping Luo. BridgeFormer: Bridging Video-text Retrieval with Multiple Choice Questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 5, 6

[23] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing Moments in Video with Natural Language. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[24] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing Moments in Video with Natural Language. In *International Conference on Computer Vision (ICCV)*, 2017. 4, 8

[25] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 4, 8

[26] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and Conquer: Question-Guided Spatio-Temporal Contextual Attention for Video Question Answering. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[27] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola,

Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. In *arXiv:1705.06950*, 2017. 2, 4, 5

[28] Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. Self-supervised Pre-training and Contrastive Representation Learning for Multiple-choice Video QA. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[29] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-Captioning Events in Videos. In *International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. In *International Journal of Computer Vision (IJCV)*, 2017. 2

[31] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference for Learning Representations (ICLR)*, 2020. 2

[32] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical Conditional Relation Networks for Video Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[33] Jie Lei, Tamara L Berg, and Mohit Bansal. QVHighlights: Detecting Moments and Highlights in Videos via Natural Language Queries. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2

[34] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. M3L: Language-based Video Editing via Multi-Modal Multi-Level Transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 7, 8

[35] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: Localized, Compositional Video Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 2

[36] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. TVQA+: Spatio-Temporal Grounding for Video Question Answering. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

[37] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. TVR: A Large-Scale Dataset for Video-Subtitle Moment Retrieval. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[38] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven C.H. Hoi. Align and Prompt: Video-and-Language Pre-training with Entity Prompts. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 8

[39] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 1, 2, 8

[40] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, Tamara Lee Berg, Mohit Bansal, Jingjing Liu, Lijuan Wang, and Zicheng Liu. VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2

[41] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7, 8

[42] Song Liu, Haoqi Fan, Shengsheng Qian, Yiru Chen, Wenkui Ding, and Zhongyuan Wang. HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval. In *arXiv:2103.15049*, 2021. 2

[43] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use What You Have: Video Retrieval Using Representations From Collaborative Experts. In *British Machine Vision Conference (BMVC)*, 2020. 2

[44] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv:1907.11692*, 2019. 2

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *International Conference on Computer Vision (ICCV)*, 2021. 4, 5, 6, 7

[46] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4, 5

[47] 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. In *arXiv:2104.08860*, 2021. 8

[48] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[49] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[50] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *International Conference for Learning Representations (ICLR)*, 2021. 2

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 2021. 2, 4, 5, 7

[52] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *International Conference on Machine Learning (ICML)*, 2021. 4, 5, 7

[53] Rene Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *International Conference on Computer Vision (ICCV)*, 2021. 3, 4, 5, 7

[54] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A Dataset for Movie Description. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2, 8

[55] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. AVLnet: Learning Audio-Visual Language Representations from Instructional Videos. In *INTERSPEECH*, 2021. 2

[56] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end Generative Pretraining for Multimodal Video Captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[57] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 7, 8

[58] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A Joint Model for Video and Language Representation Learning. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[59] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. 2

[60] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. VIMPAC: Video Pre-Training via Masked Token Prediction and Contrastive Learning. In *arXiv:2106.11250*, 2021. 2, 6

[61] Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 4, 5

[62] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *arXiv:2203.12602*, 2022. 2

[63] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning Language-Visual Embedding for Movie Understanding with Natural-Language. In *arXiv:1609.08124*, 2016. 8

[64] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 4

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 2

[66] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in One: Exploring Unified Video-Language Pre-training. In *arXiv:2203.07303*, 2022. 8

[67] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[68] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[69] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[70] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and andChristoph Feichtenhofer. Masked Feature Prediction for Self-Supervised Visual Pre-Training. In *arXiv:2112.09133*, 2022. 1, 2, 3

[71] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. MVP: Multimodality-guided Visual Pre-training. In *arXiv:2203.05175*, 2022. 1, 2

[72] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[73] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. SimMIM: A Simple Framework for Masked Image Modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3

[74] 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 *ACM Multimedia (ACMMM)*, 2017. 1, 2, 8

[75] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 8

[76] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just Ask: Learning to Answer Questions from Millions of Narrated Videos. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 8

[77] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. BERT Representations for Video Question Answering. In *Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[78] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A Joint Sequence Fusion Model for Video Question Answering and Retrieval. In *European Conference on Computer Vision (ECCV)*, 2018. 8

[79] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal Neural Script Knowledge Models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 6, 8

[80] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-Modal and Hierarchical Modeling of Video and Text. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[81] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. In *International Conference on Learning Representations (ICLR)*, 2012. 1, 2, 3

[82] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards Automatic Learning of Procedures from Web Instructional Videos. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2

[83] Linchao Zhu and Yi Yang. ActBERT: Learning Global-Local Video-Text Representations. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2