Unmasked Teacher: Towards Training-Efficient Video Foundation Models

Kunchang Li[1,2,3*] Yali Wang[1,3†] Yizhuo Li[4,3*] Yi Wang[3] Yinan He[3]
Limin Wang[5,3] Yu Qiao[3,1†]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences [3]Shanghai AI Laboratory [4]The University of Hong Kong
[5]State Key Laboratory for Novel Software Technology, Nanjing University

Code & Models: https://github.com/OpenGVLab/unmasked_teacher

Abstract

*Video Foundation Models (VFMs) have received limited exploration due to high computational costs and data scarcity. Previous VFMs rely on Image Foundation Models (IFMs), which face challenges in transferring to the video domain. Although VideoMAE has trained a robust ViT from limited data, its low-level reconstruction poses convergence difficulties and conflicts with high-level cross-modal alignment. This paper proposes a training-efficient method for temporal-sensitive VFMs that integrates the benefits of existing methods. To increase data efficiency, we mask out most of the low-semantics video tokens, but selectively align the unmasked tokens with IFM, which serves as the **U**n**M**asked **T**eacher (**UMT**). By providing semantic guidance, our method enables faster convergence and multimodal friendliness. With a progressive pre-training framework, our model can handle various tasks including scene-related, temporal-related, and complex video-language understanding. Using only public sources for pre-training in **6 days** on **32 A100** GPUs, our scratch-built ViT-L/16 achieves state-of-the-art performances on various video tasks.*

1. Introduction

Video understanding has emerged as a critical skill for artificial intelligence systems to analyze and comprehend videos effectively. The progress in video understanding is currently driven by the Image Foundation Models (IFMs) [21, 29, 5, 55], which are trained from massive datasets and adapted for different downstream tasks [16, 93, 54, 65]. However, IFMs tend to focus more on scenes and objects, disregarding the essential motion patterns and object interactions required for complex video understanding. The *true* Video Foundation Models (VFMs) are underexplored due to the high computational costs and data scarcity.

[*]Interns at Shanghai AI Laboratory. [†]Corresponding authors.

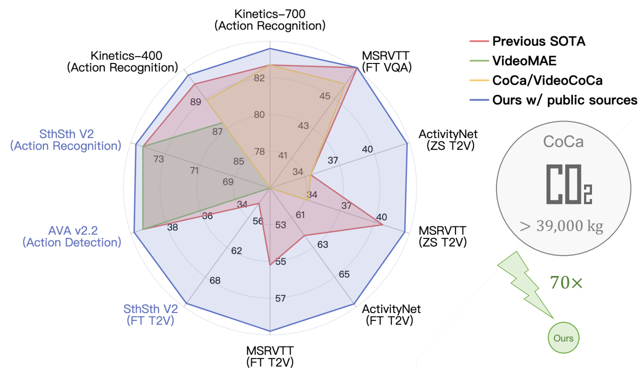


Figure 1: **Comparison with SOTA methods.** "ZS" and "FT" refer to "zero-shot" and "fine-tuned". "T2V" means video-text retrieval. For Kinetics action recognition, [80] and [70] are excluded since they utilize model ensemble. With only public sources (*i.e.*, CLIP[55]) for pre-training, our approach achieves SOTA performances on scene-related, temporal-related and complex video-language benchmarks. Compared with CoCa [85], our method is much more environmentally friendly with **70**× reduction in carbon emissions. Note that the cost of CLIP pre-training is ignored since it is publicly available.

While building VFMs on well-learned IFMs reduces training costs, it poses significant challenges in transferring knowledge from the image domain to the video domain. Firstly, due to limited video data and a substantial domain gap, video post-pretraining may undermine the generality inherited from IFMs [79]. Moreover, the strong spatial initialization offers a shortcut to perceive videos from scenes in single frames (*e.g.*, "grass" in "horse riding"), which constrains VFMs from learning spatiotemporal relationships to recognize and localize temporal-related actions, such as "opening" and "closing" in Figure 2. Lastly, this paradigm is difficult to scale up as it requires well-prepared IFMs.

The recent success of VideoMAE [62, 23] offers a data-efficient way to learn effective spatiotemporal features from

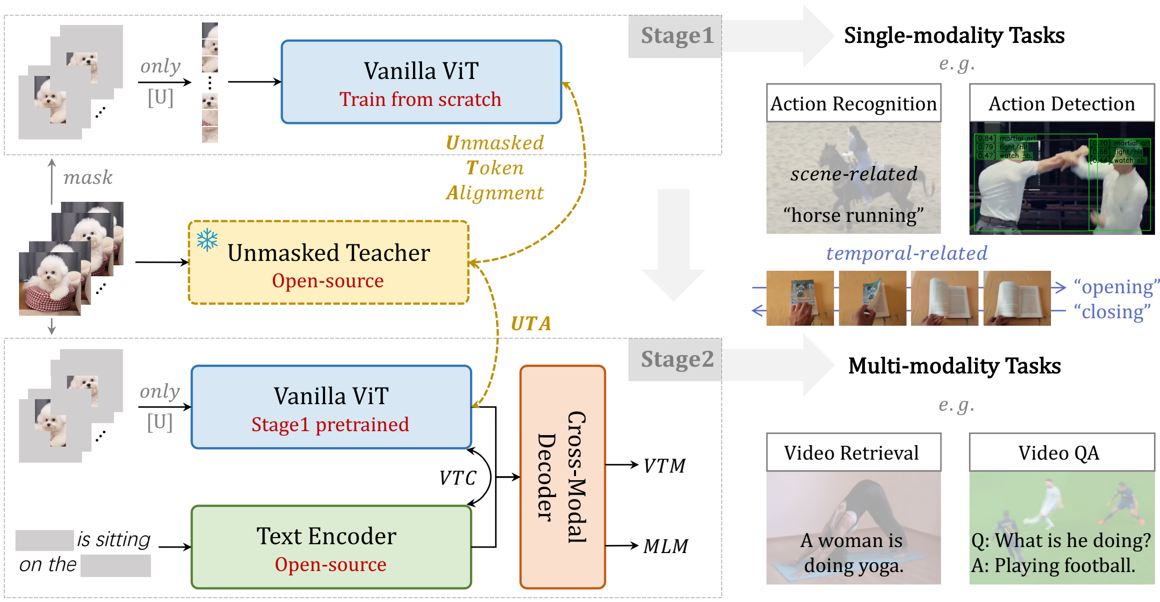


Figure 2: **Training-efficient framework for video foundation models.** For general video understanding, we propose the *progressive pre-training* with the unmasked teacher, which is *simple, scalable and reproducible*. The resulting models can not only handle scene-related and temporal-related actions well, but also conduct complex video-language understanding.

scratch, which handles complex temporal action recognition and detection tasks impressively. Nonetheless, its strong data efficiency and spatiotemporal modeling are traded by long pre-training (*e.g.*, 2400 epochs on 160k videos). Besides, it is not well-suited for video-language tasks since the low-level pixel reconstruction task conflicts with high-level cross-modal alignment [60]. Additionally, the extra decoder that handles masked and unmasked tokens causes high memory costs due to global self-attention, making scaling up this paradigm also challenging.

In this paper, we present a training-efficient method for temporal-sensitive VFMs by integrating the benefits of previous methods. Rather than directly adapting public IFM, *e.g.*, CLIP [55], we utilize them as **U**n**M**asked **T**eacher (**UMT**) to train vanilla ViT from scratch. We mask out most of the video tokens with low semantics and only align the unmasked tokens with a linear projection to the corresponding ones from the teacher. This approach not only inherits data efficiency from VideoMAE but also makes the learned video encoder multimodal-friendly (validated in Table 1). Moreover, training with only unmasked tokens without a decoder further saves GPU memory compared to VideoMAE, and the guidance from the teacher's semantically rich representation leads to faster convergence. Notably, the resulting model can handle both scene-related [33, 49] and temporal-related actions [27, 28] exceptionally well, while the alignment to CLIP features enables the model to be compatible with cross-modal learning.

To address various video tasks, we propose a progressive pre-training framework in Figure 2. In Stage 1, we only use video data for masked video modeling, resulting in a model that excels at video-only tasks. In Stage 2, we employ public vision-language data for multi-modality learning. This allows the model to conduct complex video-language tasks, such as video-text retrieval [77, 57] and video question answering [87, 75]. We use the UMT in both stages, significantly reducing the training sources and speeding up convergence. Thanks to readily-available image and language foundation models [55, 52, 46, 92, 15], our simple framework is easily scalable for video foundation models.

We conduct extensive experiments to verify the effectiveness and efficiency of our approach. As shown in Figure 1, with public sources (data/models) for pre-training, our method achieves state-of-the-art performances on various video tasks, including action recognition [33, 9, 10, 49, 27] (**90.6%** top-1 accuracy on K400), spatiotemporal localization [28] (**39.8** mAP on AVA), video-text retrieval [77, 1, 34, 57, 12] (**58.8** R@1 on MSRVTT) and video question-answering [87, 75, 86] (**47.1%** accuracy on MSRVTT). It is worth emphasizing that our method is much more environmentally friendly compared to CoCa [85], which uses 2,048 CloudTPUv4 chips for 5 days. In contrast, our pre-training requires **32 A100(80G)** GPUs within **6 days**, leading to a remarkable **70**× reduction in carbon emissions.

2. Related Works

Video foundation models. The present Video Foundation Models (VFMs) are primarily based on well-prepared Image Foundation Models (IFMs) [85, 80, 2, 91, 40, 25, 42, 64, 79]. However, the strong spatial pre-training restricts their ability to learn spatiotemporal representations. Despite the impressive results demonstrated by Florence [88], CoCa [85], MTV [80], and UniFormerV2 [40] on video-only tasks [33, 9, 10], these models struggle to handle temporal-related actions [27, 58] and localize actions [28, 32]. As for video-language tasks, there have been promising explorations on model architecture [37, 63, 36] and learning paradigms [76, 89, 25, 42, 64]. Recently, InternVideo [70] introduces general VFMs through generative and discriminative learning. However, the dependence on CLIP pre-training and tremendous training costs make it difficult to scale up. In this paper, we propose an easily scalable framework for VFMs that is much more training-efficient.

Masked vision modeling. Inspired by the success of masked language modeling [46, 18], masked vision modeling has been proposed for vision transformers [21]. BeiT [6] is the first to propose a BERT-like mask-then-predict framework to recover the discrete tokens [56], while MAE [29] designs masked autoencoders to reconstruct normalized pixel values, which reduces memory consumption by processing only unmasked tokens in the encoder. Later works can be roughly divided into BeiT-style [20, 94, 71, 3, 53] and MAE-style [74, 13, 26, 31] with various target supervision, such as HOG descriptors [71] and momentum features [61]. For spatiotemporal learning, BEVT [68] and VideoMAE [62, 23] can be seen as extensions of BeiT and MAE, respectively. Recent works also indicate that CLIP features provide good guidance for mask modeling [72, 30, 53, 52, 78], but all of them actually perform worse than CLIP itself with elaborate fine-tuning [19]. In contrast, we demonstrate that in the video domain, our model with CLIP supervision clearly outperforms the teacher.

3. Method

In this section, we introduce our **U**n**M**asked **T**eacher (**UMT**) for masked video modeling and the progressive pre-training framework for temporal-sensitive video foundation models, as illustrated in Figure 2.

3.1. Unmasked Teacher

As discussed in the introduction, directly adapting the public Image Foundation Model (IFM) to Video Foundation Model (VFM) is challenging [50, 40], thus we propose using IFM as a teacher to train a VFM from scratch. Given the limited data scale, we leverage mask modeling [29] to make good use of the video data. However, unlike VideoMAE [62], we selectively align the unmasked tokens with the teacher, removing an extra decoder for efficient training.

Architecture. We choose CLIP-ViT [55] as an unmasked teacher due to its rich semantics that are learned with language guidance, which is beneficial for our following multi-modality learning. To fully impart the teacher's knowledge, we maintain its spatial architecture to process each video frame individually. For our backbone, we apply the vanilla ViT without a class token. We employ spatiotemporal attention [7] to encourage all the unmasked tokens to interact with each other. For better alignment with the spatial teacher, we do not use temporal downsampling, thus the tokens can be aligned frame by frame.

Masking. Following VideoMAE, we use a high masking ratio (*e.g.*, 80%) to cut down video redundancies. However, the aggressive random masking may only retain the background tokens, which contain insignificant information and hinder the teacher's knowledge transfer. To enhance target effectiveness, we apply the semantic masking [30] frame by frame, where the tokens with important clues are maintained at higher probabilities. Specifically, given the class token $\mathbf{z}_{cls}\in\mathbb{R}^{1\times C}$ and the spatial tokens $\mathbf{Z}\in\mathbb{R}^{L\times C}$ in the t-th frame of CLIP-ViT ($L{=}H\times W$ is the token number and C is the token dimension), we calculate the attention score in the last self-attention [21] layer:

$$\mathbf{A} = \sum_{n=1}^{N} \mathbf{A}_n(Q_n(\mathbf{z_{cls}}), K_n(\mathbf{Z}))/N, \tag{1}$$

$$\mathbf{A}_n(\mathbf{q},\mathbf{k}) = \mathrm{softmax}(\mathbf{q}\mathbf{k}^T/\sqrt{C/N}), \tag{2}$$

where N is the head number, and $Q_n(\cdot)$ and $K_n(\cdot)$ are the linear projections in the n-th head. The $\mathbf{A}\in\mathbb{R}^{1\times L}$ represents the semantic importance of each token, and we select the unmasked tokens by a multinomial distribution based on $\mathbf{A}$ to retain the informative objects in each frame. Moreover, we sparsely sample frames from the raw videos [67], which provides a more complicated action context due to the large frame stride. The strategy encourages the model to reason long-term spatiotemporal relationships among objects.

Target. For the teacher model, we input all L spatial tokens along with the class token, frame by frame. In contrast, for the student model, we only input the unmasked tokens, which are equal to $L(1-r)T$ tokens, where r is the masking ratio and T is the frame number. To distill the rich semantics more effectively, we process the output teacher tokens using the pre-trained visual projection, which is designed to establish meaningful connections between visual and text embeddings. Additionally, we add a simple linear projection for the student model to align the token dimension. We select the corresponding unmasked token from the student and teacher, and compute the mean squared error (MSE) between the normalized pairs. Compared to low-level pixel reconstruction, token alignment requires a high-level understanding, which is beneficial for multi-modality learning.

3.2. Progressive Pre-training

For general video understanding, it is vital for the foundation model to handle video-language tasks. However, directly training such a model from scratch is inefficient. For example, CoCa [85] utilizes 4.8B data to train 5 days on 2,048 CloudTPUv4 chips. Therefore, we introduce a training-efficient framework with progressive pre-training.

Pre-training pipeline. Figure 2 outlines our pipeline. In Stage 1, we train the ViT from scratch using only high-quality videos and guidance from Unmasked Teacher. The masked video modeling fully mines knowledge from the videos, resulting in a model that excels at video-only tasks. In Stage 2, we equip the pre-trained ViT with a text encoder and cross-modal decoder, initialized with the well-prepared language model. And we conduct multi-modality training with large-scale vision-text pairs, enabling the model to handle complex video-language tasks. It's worth noting that currently, open-source language models are larger and more diverse than vision models, making it easy to scale up our foundation models. For example, the largest OPT [92] has 175B parameters, while ViT-G [90] only has 1.8B.

Pre-training objectives. For both stages, we utilize Unmasked Teacher to perform Unmasked Token Alignment (**UTA**). In Stage 2, we employ three other popular objectives: **(i)** Video-Text Contrastive (**VTC**) learning, which aims to align the pooled unmasked video and text embeddings. We use the symmetric contrastive loss [4] to maximize the mutual information. **(ii)** Video-Text Matching (**VTM**) enhances cross-modal fusion by aligning the unmasked video and text tokens. We adopt the binary cross-entropy loss with hard negative mining [39, 36]. **(iii)** Masked Language Modeling (**MLM**) uses the cross-modal decoder to predict masked words from the other text and unmasked video tokens. We follow the BERT [17] strategy but mask 50% of the text tokens.

4. Experiments

4.1. Implementation

Datasets. Unless otherwise stated, we use Kinetics-710 dataset [40] in Stage 1, which is a combination of Kinetics-400, 600 and 700 [33, 9, 10] and excludes any repeated or leaked videos. In Stage 2, we utilize image-text data for co-training [63, 36, 64], where images are treated as single-frame videos. We use three corpora as in [14]: **(i) 5M** Corpus comprises WebVid-2M [4] video-text pairs and CC3M [59] image-text pairs. **(ii) 17M** Corpus includes four other image-text datasets: COCO [44], Visual Genome [35], SBU Captions [51], and CC12M [11]. **(iii) 25M** Corpus uses a larger version of WebVid containing 10M video-text pairs.

Settings. In this paper, we consider two model configurations: ViT-B/16 [21] with BERT$_{base}$ [17] and ViT-L/16 with BERT$_{large}$. And CLIP-ViT-B/16 [55] and CLIP-ViT-L/14 are adopted as teachers for the base and large models, respectively. Since CLIP-ViT-L/14 uses a smaller patch size, we adopt a smaller input resolution (*i.e.*, 196) to align the token number. For Stage-1 pre-training, we follow most of the hyperparameter settings in VideoMAE [62]. However, we sparsely sample [67] 8 frames and use a masking ratio of 80%. By default, we train both models on 32 A100 with a batch size of 2048 for 200 epochs. The training on Kinetics-710 takes about **60** and **90** hours for ViT-B/16 and ViT-L/16, respectively. In Stage 2, we follow [36] to sample 4 frames and train for 10 epochs. Specifically, we mask 50% image and 80% video tokens. Both models are trained on 32 A100 with a batch size of 4096. The pre-training on 25M Corpus takes about **24** and **40** hours respectively for the base and large models. For more implementation details about training, please refer to the supplemental materials.

[U]	[M]	MAE	Memory (G)	SSV2	K400	MSR
✗	✗	✓	44.0	67.1	78.8	55.6
✓	✗	✓	52.5	**70.2**	83.9	64.5
✓	✓	✗	43.6	70.0	84.6	65.2
✓	✗	✗	**16.0**	**70.2**	**84.9**	**66.8**

Table 1: **Target design.** We benchmark ViT-B/16 in 32 A100 with a batch size of 2048. "[U]", "[M]" and "MAE" refers to unmasked token alignment, masked token recovering and pixel reconstruction [62] respectively. The pixel reconstruction conflict with our unmasked token alignment, and hinder the following multimodal learning.

Mask	Sampling	T-Down	SSV2	K400
Tube	Sparse	✗	**70.2**	84.3
Random	Sparse	✗	**70.2**	84.6
Semantic	Sparse	✗	**70.2**	**84.9**
Semantic	Dense	✗	69.8	84.0
Semantic	Sparse	✓	69.5	84.6

Table 2: **Mask type, sampling method and temporal downsampling.** Semantic masking [30] works best.

4.2. Ablation Study

We ablate the properties of UMT in both stages on both scene-related [33, 77] and temporal-related tasks [27, 36]. For single-modality learning, we pre-train ViT-B/16 for 200 epochs on SthSth V2 [27] or K400 [33] dataset. For multi-modality learning, we use K710 pre-trained models and further pre-train it for 10 epochs on 5M Corpus. Except for Table 1, where we use K400 pre-training.

Target. Table 1 presents a comparison of training targets. Compared with pixel reconstruction [62], our unmasked token alignment significantly improves the accuracy with only 36% memory cost. However, combining the two targets results in poor results on K400 and MSRVTT, indicating a conflict between low-level reconstruction and high-level alignment. Moreover, recovering the masked tokens has a detrimental effect, possibly due to the high masking ratio making high-level recovery too challenging. The results demonstrate our method is effective to learn temporal-sensitive and multimodal-friendly representation.

Mask type, sampling method, and temporal downsampling. Table 2 indicates that different masking strategies yield comparable results in SthSth V2. We contend that

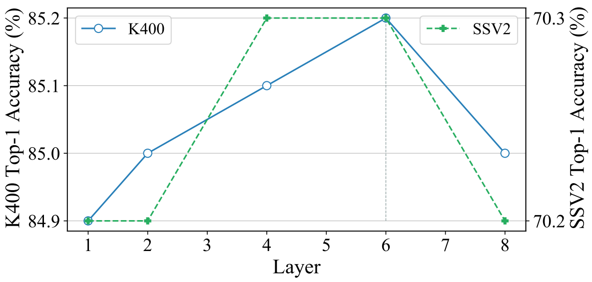


Figure 3: **Aligned layers.** Since the GPU memory and running speed are similar, we align the last 6 layers.

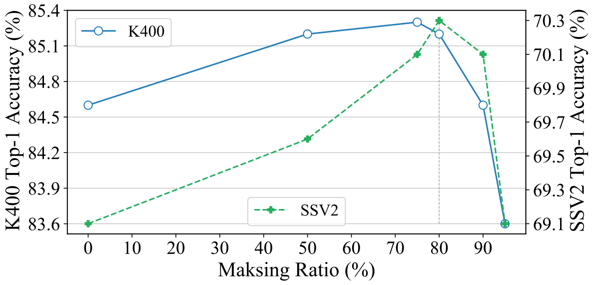


Figure 4: **Masking ratio.** We use the masking ratio of 0.8 for a better trade-off on both datasets.

recognizing the category of "something" is not necessary for SthSth V2, but it requires deducing the intricate motion between objects, thus random masking suffices. However, it is critical for K400 to identify the scene and objects, making semantic masking advantageous for knowledge distillation. Moreover, sparse sampling without temporal downsampling is more appropriate for our approach.

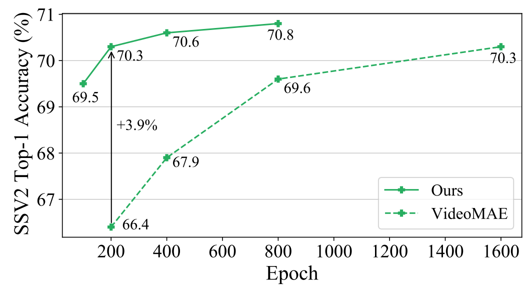
Aligned layers. We try to align more layers in Figure 3, and the losses are averaged across multiple layers. Since the GPU memory and running speed are similar, we simply align the last 6 layers for the best results.

Masking ratio. Figure 4 shows that proper high ratios work better. When using a ratio of 95%, the performances dramatically drop since it is too challenging for token alignment. Conversely, when removing masks, the task is too easy to learn the token relationships in space and time. By default, we adopt the ratio of 80% for better trade-offs.

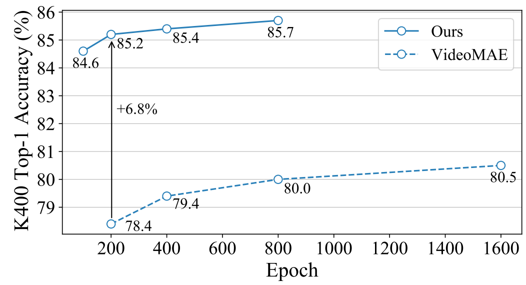
Training schedule. Figure 5 presents the results of different training schedules. On one hand, a longer training schedule consistently improves the performances on both benchmarks. On the other hand, compared to VideoMAE [62], our method shows a faster convergence speed. For example, when pre-training for 200 epochs, our models achieve 3.9% and 6.8% top-1 accuracy on SthSth V2 and Kinetics-400, respectively.

Why does UMT work? In Table 3, we investigate the crucial designs of our Unmasked Teacher. **(i) Spatiotemporal attention**: In the 2nd and 3rd parts, we compare the student with spatial attention and spatiotemporal attention during fine-tuning. Our results indicate that utilizing joint attention significantly enhances performance. Moreover, employing spatiotemporal attention during pre-training further improves performance (the 4th part), validating our assumption that joint attention encourages interaction among all unmasked tokens. **(ii) Masked modeling**: In the 4th part, we observe that masked modeling plays a crucial role. However, when using spatial attention during pre-training, masked modeling becomes detrimental. We argue that when processing each frame individually with a high mask ratio of 80%, the token alignment task becomes excessively challenging. **(iii) Teacher attention**: The 5th part shows that although CLIP-*ST* achieves better performance after fine-tuning, directly applying it as the teacher model leads to a performance drop. We contend that without post-training in the video domain, CLIP-*ST* may disrupt the representation learned in the image domain.

Outperforming the CLIP teacher. In the image do-



(a) Results on SthSth V2



(b) Results on Kinetics-400

Figure 5: **Training schedules.** A longer training schedule leads to more significant improvement.

Teacher	Mask	PT Student	FT Student	SSV2	K400
fine-tuning CLIP-*S*				57.4	82.0
fine-tuning CLIP-*ST*				68.0	82.5
CLIP-*S*	✗	ViT-*S*	ViT-*S*	54.5	82.4
CLIP-*S*	✓	ViT-*S*	ViT-*S*	54.0	82.2
CLIP-*S*	✗	ViT-*S*	ViT-*ST*	68.0	83.7
CLIP-*S*	✓	ViT-*S*	ViT-*ST*	67.2	83.4
CLIP-*S*	✗	ViT-*ST*	ViT-*ST*	69.1	84.6
CLIP-*S*	✓	ViT-*ST*	ViT-*ST*	**70.3**	**85.2**
CLIP-*ST*	✓	ViT-*ST*	ViT-*ST*	68.7	83.7

Table 3: **Why does UMT work?** "*S*" and "*ST*" refers to spatial and spatiotemporal attention respectively. Spatiotemporal attention and mask modeling are vital for UMT.

Teacher	Student	SSV2	K400
fine-tuning DINO-*ST*		65.0	80.8
DINO-*ST*	ViT-*ST*	67.9 (+2.9)	81.7 (+0.9)
fine-tuning CLIP-*ST*		68.0	82.5
CLIP-*ST*	ViT-*ST*	70.3 (+2.3)	85.2 (+2.7)
fine-tuning BeiTv2-*ST*		68.1	82.0
BeiTv2-*ST*	ViT-*ST*	70.3 (+2.2)	84.4 (+2.4)

Table 4: **Different teacher.** The student models significantly outperform the corresponding teacher models.

Image	Video	Text	Memory (G)	MSR	SSV2
50	60	50	35.6	66.9	80.6
50	80	50	18.6	**67.0**	**80.8**
75	80	50	18.6	66.5	80.6
75	90	50	13.1	65.9	79.5
90	95	50	12.1	65.7	79.2
50	80	25	18.6	66.5	80.1
50	80	75	18.6	66.6	78.2

Table 5: **Different masking ratios for multi-modality pre-training.** We benchmark ViT-B/16 in 16 A100 with a batch size of 2048. We report the average text-to-video retrieval R@1,5,10 accuracy of MSRVTT and SSV2-label. Masking 50% image and 80% video tokens works best.

main, the prior research [19] has shown that, CLIP itself with fine-tuning surpasses existing CLIP-targeted MIM methods [72, 73, 30, 53]. However, Table 3 indicates that in the video domain, the student model (the 4th part) clearly outperforms the teacher, *i.e.*, CLIP-*ST* with our elaborate fine-tuning. We attribute the success to masked video modeling with spatiotemporal attention, which encourages the model to capture long-term dependencies among objects.

Different teachers. In Table 4, we adopt different models [8, 55, 52] as the unmasked teachers. As expected, the student models clearly outperform the corresponding teacher models, which have undergone elaborate fine-tuning. It's important to note that both student and teacher models share the same architecture, further emphasizing the effectiveness of our approach.

Multi-modality masking ratios. In Table 5, we first alter the masking ratios of the image and video data. Since we co-train image-text and video-text data with the same batch size, the GPU memory primarily depends on the video masking ratio. As expected, processing images requires a lower masking ratio of 50%. Although higher masking ratios reduce memory consumption, the corresponding performances are lower. Additionally, masking too few (25%) or too many (75%) text tokens leads to inferior results.

Multi-modality pre-training objectives. For cross-modal retrieval, utilizing either VTC or VTM for visual-text pairs is necessary. In Table 6, all loss weights are set to 1. The 1st part reveals that VTM performs better than VTC. Besides, the 2nd part shows that combining VTC or MLM with VTM leads to a minor improvement, while integrating all three objectives significantly enhances the performance.

VTC	VTM	MLM	UTA	Memory (G)	MSR	SSV2
✓	✗	✗	✓	17.3	60.3	74.3
✗	✓	✗	✓	18.1	65.6	79.1
✓	✓	✗	✓	18.2	65.9	79.2
✗	✓	✓	✓	18.5	65.7	79.2
✓	✓	✓	✓	18.6	**67.0**	**80.8**
✓	✓	✓	✗	56.6	66.6	79.8

Table 6: **Objectives for multi-modality pre-training.** All the objectives are helpful to the downstream tasks.

Settings	MSR	SSV2
Baseline	**67.0**	**80.8**
one-stage pre-training	57.9	64.5
only random mask w/o UMT	66.9	80.1
+ visual projection alignment	66.4	80.5
+ visual & text projection alignment	66.0	80.0
+ extra pre-training w/o mask	66.5	80.5

Table 7: **Other designs for multi-modality pre-training.**

Lastly, without our unmasked teacher alignment, the memory usage triples, while the performances drop.

Other designs. Table 7 showcases alternative designs for our multi-modality pre-training. Firstly, we attempt to directly perform Stage 2 with a randomly initialized video encoder. For a fair comparison, we incorporate Kinetics-710 and conduct the same number of data iterations. However, the results demonstrate that the one-stage pre-training is challenging to converge, leading to poor performance. Secondly, we randomly mask the video without an unmasked teacher for supervision, which slightly reduces the overall performance. Additionally, we consider aligning the visual and text projection with the CLIP teacher, since the teacher model also adopts contrastive learning. However, introducing extra alignment tasks turns out to be redundant and even harmful. Finally, we conduct extra pre-training without masks after masked pre-training. Though it improves zero-shot performance (+1.5% higher average recall accuracy), the fine-tuned results are not as good as expected.

4.3. Single-modality tasks

We evaluate our method on two conventional video-only tasks: recognizing and localizing actions on six large-scale benchmarks, including the *Kinetics* family (*i.e.*, Kinetics-400, 600 and 700 [33, 9, 10]), *Moments in Time* V1 [49] and *Something-Something* V2 [27] for action recognition, and *AVA* V2.2 [28] for spatiotemporal localization.

Kinetics. Table 8 reports the SOTA methods with supervised and self-supervised learning on K400. On one hand, our UMT with intermediate fine-tuning outperforms the previous models that rely on web-scale pre-training, *e.g.*, the UMT-L achieves 0.4% higher top-1 accuracy than MTV-H [80] with only 1/10 of the FLOPs and 1/3 of the parameters. On the other hand, our UMT surpasses its counterparts with masked video modeling, *e.g.*, compared with

	Method	Backbone	Extra data	Input Size	GFLOPs	Param	Top-1	Top-5
supervised	SlowFast$_{101}$ [24]	R101+NL		80×224^2	$234\times3\times10$	60	79.8	93.9
	MViTv2-B [43]	MViTv2-B		32×224^2	$255\times1\times5$	37	81.2	95.1
	UniFormer-B[41]	UniFormer-B	IN-1K	32×224^2	$259\times3\times4$	50	83.0	95.4
	TimeSformer-L [7]	ViT-B	IN-21K	96×224^2	$2380\times3\times1$	121	80.7	94.7
	VideoSwin-L [45]	Swin-L	IN-21K	32×224^2	$604\times3\times4$	197	83.1	95.9
	Methods with web-scale data. FLD, ALIGN and CLIP consist of image-text pairs. WTS collects video-text pairs.							
	ViViT-H [2]	ViT-H	JFT-300M	32×320^2	$3981\times3\times4$	654	84.9	95.8
	CoVeR [91]	ViT-L	JFT-3B+SSV2+MiT+IN	16×448^2	$5860\times3\times1$	431	87.1	-
	CoCa [85]	ViT-g	JFT-3B+ALIGN-1.8B	16×576^2	N/A$\times3\times4$	1000+	88.9	-
	MTV-H [80]	ViT-H+B+S+T	IN-21K+WTS-60M	32×280^2	$6130\times3\times4$	1000+	89.9	98.3
	UniFormerV2-L [40]	ViT-L	CLIP-400M+K710†	64×336^2	$12550\times3\times2$	354	90.0	98.4
self-supervised	BEVT$_{800e}$ [68]	Swin-B	IN-1K	32×224^2	$282\times3\times4$	88	81.1	-
	MaskFeat$_{1600e}$ [71]	MViTv2-L		16×224^2	$377\times1\times10$	218	84.3	96.3
	ST-MAE-B$_{1600e}$ [23]	ViT-B		16×224^2	$180\times3\times7$	87	81.3	94.9
	ST-MAE-L$_{1600e}$ [23]	ViT-L	K600	16×224^2	$598\times3\times7$	304	84.9	96.2
	ST-MAE-L$_{1600e}$ [23]	ViT-L	K600†	16×224^2	$598\times3\times7$	304	86.5	97.2
	VideoMAE-B$_{1600e}$ [62]	ViT-B		16×224^2	$180\times3\times5$	87	81.5	95.1
	VideoMAE-L$_{1600e}$ [62]	ViT-L		16×224^2	$597\times3\times5$	305	85.2	96.8
	VideoMAE-L$_{1600e}$ [62]	ViT-L		16×320^2	$3958\times3\times5$	305	86.1	97.3
	MVD-H$_{800e}$ [69]	ViT-H	IN-1K	16×224^2	$1192\times3\times5$	633	87.2	97.4
	UMT-B$_{800e}$	ViT-B		8×224^2	$180\times3\times4$	87	85.7	97.0
	UMT-B$_{200e}$	ViT-B	K710	8×224^2	$180\times3\times4$	87	85.7	96.9
	UMT-B$_{200e}$	ViT-B	K710†	8×224^2	$180\times3\times4$	87	87.4	97.5
	UMT-L$_{400e}$	ViT-L		8×224^2	$596\times3\times4$	304	88.9	98.3
	UMT-L$_{200e}$	ViT-L	K710	8×224^2	$596\times3\times4$	304	89.1	98.2
	UMT-L$_{200e}$	ViT-L	K710†	8×224^2	$596\times3\times4$	304	90.3	**98.7**
	UMT-L$_{200e}$	ViT-L	K710†	16×224^2	$1434\times3\times4$	304	**90.6**	**98.7**

Table 8: **Comparison with the state-of-the-art methods on Kinetics-400.** For UMT, we use a masking ratio of 80%. The results using spatial resolution $>224^2$ are noted in blue. "†" marks the results with intermediate fine-tuning.

Method	Input Size	FLOPs (T)	Param (M)	K600		K700	
				Top-1	Top-5	Top-1	Top-5
SlowFast$_{101}$ [24]	80×224^2	7.0	60	81.8	95.1	71.0	89.6
MViTv2-L [43]	40×312^2	33.9	218	87.5	97.8	79.4	94.9
Methods with web-scale data.							
CoVeR [91]	16×448^2	17.6	431	87.9	-	79.8	-
CoCa [85]	16×576^2	N/A	1000+	89.4	-	82.7	-
UniFormerV2-L [40]	32×224^2	16.0	354	89.5	98.3	82.1	96.1
UniFormerV2-L [40]	64×336^2	75.3	354	90.1	98.5	82.7	96.2
MTV-H [80]	32×224^2	44.5	1000+	89.6	98.3	82.2	95.7
MTV-H [80]	32×320^2	73.6	1000+	90.3	98.5	83.4	96.2
UMT-B	8×224^2	2.2	87	87.8	97.8	78.5	94.3
UMT-L	8×224^2	7.2	304	90.4	98.7	83.2	96.5
UMT-L	16×224^2	17.2	304	**90.5**	**98.8**	**83.6**	**96.7**

Table 9: **Comparison with the state-of-the-art methods on Kinetics-600/700.** For UMT, we report the results with K710 pre-training and intermediate fine-tuning.

Method	Input Size	FLOPs (G)	Param (M)	MiT V1	
				Top-1	Top-5
ViViT-L [2]	32×224^2	3980×3	612	38.5	64.1
MTV-H [80]	32×224^2	3706×12	1000+	45.6	74.7
Methods with web-scale data.					
CoVeR [91]	16×448^2	5860×3	431	46.1	-
MTV-H [80]	32×280^2	6130×12	1000+	47.2	75.7
UniFormerV2-B [40]	8×224^2	148×12	115	42.7	71.5
UniFormerV2-L [40]	8×224^2	666×12	354	47.0	76.1
UniFormerV2-L [40]	8×336^2	1568×12	354	47.8	76.9
UMT-B	8×224^2	180×12	87	44.6	74.0
UMT-L	8×224^2	596×12	304	48.0	77.8
UMT-B	8×384^2	786×12	87	45.5	74.6
UMT-L	8×384^2	2440×12	304	**48.7**	**78.2**

Table 10: **Comparison with the state-of-the-art methods on Moments in Time V1.** Following the previous methods, we fine-tune the models pre-trained by Kinetics-400.

VideoMAE [62] with 1600-epoch pre-training, the UMT-L with 400-epoch pre-training obtains 3.7% accuracy improvement. For K600 and K700, our UMT-L also obtains the SOTA performances (**90.5%** and **83.6%** see Table 9).

Moments in Time. As shown in Table 10, our UMT-L achieves **1.0%/1.7%** higher top-1/5 accuracy compared to the advanced UniFormerV2-L [40], while utilizing fewer FLOPs. Note that MiT is more challenging due to the large inter-class and intra-class variation, thus the results demonstrate the robustness and effectiveness of our method.

Something-Something. Distinct from previous benchmarks, this particular dataset requires complex and long-term modeling to accurately recognize temporal-related actions, such as "pretending to close something without actually closing it". Without any additional data, our UMT-L model outperforms the UniFormerV2-L [40] (74.4% *vs.* 73.0% in Table 11) which was specifically tailored for temporal modeling. Additionally, our approach achieves comparable performances to VideoMAE [62] with significantly fewer epochs. Intriguingly, VideoMAE performs worse when utilizing Kinetics for masked modeling, while our UMT performs even better. This demonstrates the versa-

Method	Extra Data	#F	FLOPs (G)	Param (M)	SSV2 Top-1	SSV2 Top-5
supervised						
SlowFast$_{101}$ [24]	K400	32	106×3	53	63.1	87.6
TDN$_{EN}$ [66]	IN-1K	87	198×3	88	69.6	92.2
TimeSformer-L [7]	IN-21K	96	2380×3	121	62.3	-
MViTv1-B [22]	K400	64	455×3	37	67.7	70.9
MViTv2-B [43]	K400	64	225×3	51	70.5	92.7
UniFormer-B [41]	IN-1K+K400	32	259×3	50	71.2	92.8
ViViT-L [2]	IN-21K+K400	32	3980×3	612	65.9	89.9
MTV-B [80]	IN-21K+K400	32	399×12	310	68.5	90.4
VideoSwin-B [45]	IN-21K+K400	32	321×3	88	69.6	92.7
CoVeR ↑448 [91]	JFT-3B+KMI	16	5860×3	431	69.9	-
UniFormerV2-B [40]	CLIP-400M	32	375×3	163	70.7	93.2
UniFormerV2-L [40]	CLIP-400M	32	1718×3	574	73.0	94.5
self-supervised						
BEVT$_{800e}$ [68]	IN-1K+K400	32	321×3	88	70.6	-
MaskFeat-L$_{1600e}$ ↑312 [71]	K400	16	2828×3	218	74.4	94.6
ST-MAE-L$_{1600e}$ [23]	K400*	16	598×3	304	72.1	93.9
ST-MAE-L$_{1600e}$ [23]	K600*	16	598×3	304	73.0	94.2
ST-MAE-L$_{1600e}$ [23]	K700*	16	598×3	304	73.6	94.4
VideoMAE-B$_{1600e}$ [62]	K400*	16	180×6	87	69.7	92.3
VideoMAE-B$_{2400e}$ [62]	-	16	180×6	87	70.8	92.4
VideoMAE-L$_{1600e}$ [62]	K400*	16	597×6	305	74.0	94.6
VideoMAE-L$_{2400e}$ [62]	-	16	597×6	305	74.3	94.6
UMT-B$_{800e}$	-	8	180×6	87	70.8	92.6
UMT-B$_{200e}$	K710*	8	180×6	87	70.8	92.4
UMT-L$_{400e}$	-	8	596×6	304	74.4	94.5
UMT-L$_{200e}$	K710*	8	596×6	304	**74.7**	**94.7**

Table 11: **Comparison with the state-of-the-art methods on Something-Something V2.** "#F" refers to the frame number. "KMI" refers to "K400+MiT+IN". "*" means the labels of extra data are not used for intermediate fine-tuning.

tility and adaptability of our method, which can be applied to diverse video domains with the same pre-training.

AVA. Table 12 presents the results of the action detection on AVA. Remarkably, our UMT achieves 2.0 mAP improvement over the advanced VideoMAE [62] with only K400 pre-training. Furthermore, our method achieves the impressive **39.8** mAP with K710 pre-training, showcasing its robust transferability for spatiotemporal understanding.

4.4. Multi-modality tasks

We further validate our model on two mainstream video-language tasks, including video-text retrieval (MSRVTT [77], DiDeMo [1], ActivityNet [34], LSMDC [57], MSVD [12] and Something-Something [36]) and video question-answering (ActivityNet-QA [87], MSRVTT-QA [75], MSRVTT-MC [86] and MSVD-QA [75]).

Zero-shot text-to-video retrieval. Table 13 indicates that the UMT-B outperforms the top-performing models [64, 36, 14] by **0.9%**, **5.0%**, and **4.6%** R@1 on MSRVTT, DiDeMo, and ActivityNet, respectively. In addition, our UMT-L has set new records across all datasets. We see scores of **42.6%**, **48.6%**, **42.8%**, **25.2%**, and **72.2%** on MSRVTT, DiDeMo, ActivityNet, LSMDC, and MSVD, in respective order. These notable results emphasize the exceptional robustness and effectiveness of our method.

Method	PT Data	Input Size	FLOPs (G)	Param (M)	AVA mAP
supervised					
SlowFast [24]	K400	32×224^2	138	53	23.8
SlowFast [24]	K600	64×224^2	296	59	27.5
MViTv1-B [22]	K400	64×224^2	455	36	27.3
MViTv1-B [22]	K600	32×224^2	236	53	28.7
MViTv2-B [43]	K400	32×224^2	225	51	28.1
MViTv2-B [43]	K700	32×224^2	225	51	31.3
MViTv2-L [43]	IN-21K+K700	40×312^2	2828	213	33.5
self-supervised					
MaskFeat-L [71]	K400	40×312^2	2828	218	36.3
MaskFeat-L [71]	K600	40×312^2	2828	218	37.8
ST-MAE-L [23]	K400	16×224^2	598	304	34.8
ST-MAE-L [23]	K700	16×224^2	598	304	37.3
VideoMAE-B [62]	K400	16×224^2	180	87	31.8
VideoMAE-L [62]	K400	16×224^2	597	305	37.0
VideoMAE-L [62]	K700	16×224^2	597	305	39.3
UMT-B	K400	8×224^2	180	87	32.7
UMT-B	K710	8×224^2	180	87	33.5
UMT-L	K400	8×224^2	596	304	39.0
UMT-L	K710	8×224^2	596	304	**39.8**

Table 12: **Comparison with the state-of-the-art methods on AVA v2.2.** All the self-supervised methods are with intermediate fine-tuning on the pre-training data.

Method	#Pairs	MSR	DDM	ANet	LSMDC	MSVD
Frozen [4]	5M	18.7	20.2	-	-	-
VIOLET [25]	138M	25.9	23.5	-	-	-
Singularity [36]	17M	34.0	37.1	30.6	-	-
OmniVL [64]	17M	34.6	33.3	-	-	-
VINDLU [14]	25M	32.0	36.9	30.9	-	-
CLIP4Clip [48]	400M	30.6	-	-	13.6	36.2
InternVideo [70]	646M	40.7	31.5	30.7	17.6	43.4
VideoCoCa [81]	4.8B	34.3	-	34.5	-	-
	5M	29.6	33.4	28.3	16.8	55.7
UMT-B	17M	35.5	41.9	33.8	18.1	58.8
	25M	35.2	41.2	35.5	19.1	60.3
	5M	33.3	34.0	31.9	20.0	68.1
UMT-L	17M	**42.6**	46.4	**42.8**	**25.2**	71.0
	25M	40.7	**48.6**	41.9	24.9	**72.2**

Table 13: Zero-shot text-to-video retrieval on MSRVTT ("MSR"), DiDeMo ("DDM"), AcitivityNet ("ANet"), LSMDC, and MSVD. We only report the R@1 accuracy. Models pre-trained with large-scale pairs are noted in gray.

Text-to-video retrieval. Table 14 lists the fine-tuned results, where our UMT-L significantly outperforms previous methods pre-trained with large-scale pairs [47, 79, 70]. Specifically, our UMT-L achieves **58.8%** (+3.6%), **70.4%** (+9.2%), **66.8%** (+4.6%), **43.0%** (+9.0%), and **80.3%** (+21.9%) on MSRVTT, DiDeMo, ActivityNet, LSMDC, and MSVD, respectively. Furthermore, Table 15 showcases its impressive performances on the temporally-heavy SSV2-label and SSV2-template datasets, *i.e.*, **73.3%** and **90.8%**, respectively. These results demonstrate its profound capability for temporal modeling.

Method	#Pairs	MSRVTT			DiDeMo			ActivityNet			LSMDC			MSVD		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
ClipBERT [37]	5.4M	22.0	46.8	59.9	20.4	48.0	60.8	21.3	49.0	63.5		-			-	
Frozen [4]	5M	31.0	59.5	70.5	34.6	65.0	74.7		-		15.0	30.8	39.8	33.7	64.7	76.3
VIOLET [25]	138M	34.5	63.0	73.4	32.6	62.8	74.7		-		16.1	36.6	41.2		-	
All-in-one [63]	138M	37.9	68.1	77.1	32.7	61.4	73.5	22.4	53.7	67.7		-			-	
LAVENDER [42]	30M	40.7	66.9	77.6	53.4	78.6	85.3		-		26.1	46.4	57.3	50.1	79.6	87.2
Singularity [36]	17M	42.7	69.5	78.1	53.1	79.9	88.1	48.9	77.0	86.3		-			-	
OmniVL [64]	17M	47.8	74.2	83.8	52.4	79.5	85.4		-			-			-	
HiTeA [84]	17M	46.8	71.2	81.0	56.5	81.7	89.7	49.7	77.1	86.7	28.7	50.3	59.0		-	
VINDLU [14]	25M	46.5	71.5	80.4	61.2	85.8	91.0	55.0	81.4	89.7		-			-	
CLIP4Clip [48]	400M	44.5	71.4	81.6	42.8	68.5	79.2	40.5	72.4	83.4	21.6	41.8	49.8	46.2	76.1	84.6
CLIP-ViP [79]	500M	54.2	77.2	84.8	50.5	78.4	87.1	53.4	81.4	90.0	29.4	50.6	59.0		-	
InternVideo [70]	646M	55.2	79.6	87.5	57.9	82.4	88.9	62.2	85.9	93.2	34.0	53.7	62.9	58.4	84.5	90.4
UMT-B	5M	46.3	72.7	82.0	54.8	83.0	89.0	52.1	80.5	89.6	30.3	51.8	61.4	67.0	92.7	96.7
	17M	50.6	75.4	83.5	60.8	85.1	91.0	56.1	82.5	91.2	32.3	54.5	61.9	70.8	93.7	96.6
	25M	51.0	76.5	84.2	61.6	86.8	91.5	58.3	83.9	91.5	32.7	54.7	63.4	71.9	94. 5	97.8
UMT-L	5M	53.3	76.6	83.9	59.7	84.9	90.8	58.1	85.5	92.9	37.7	60.6	67.3	76.9	96.7	98.7
	17M	56.5	80.1	**87.4**	66.6	89.9	**93.7**	66.6	88.6	94.7	41.4	63.8	72.3	78.8	97.3	98.8
	25M	**58.8**	**81.0**	87.1	**70.4**	**90.1**	93.5	**66.8**	**89.1**	**94.9**	**43.0**	**65.5**	**73.0**	**80.3**	**98.1**	**99.0**

Table 14: Text-to-video retrieval on MSRVTT, DiDeMo, AcitivityNet, LSMDC, and MSVD. "#Pairs" denotes the number of pre-training pairs. Models pre-trained with large-scale pairs are noted in gray.

Method	#Pairs	SSV2-label			SSV2-template		
		R@1	R@5	R@10	R@1	R@5	R@10
CLIP4Clip [48]	400M	43.1	71.4	-	77.0	96.6	-
Singularity [36]	17M	47.4	75.9	-	77.6	96.0	-
VINDLU [14]	25M	53.1	81.8	-	83.3	**100**	-
HiTeA [84]	5M	55.2	81.4	89.1	85.6	**100**	**100**
UMT-B	5M	63.1	87.1	92.3	87.3	**100**	**100**
	17M	63.4	88.0	92.9	86.8	99.4	**100**
	25M	64.2	88.2	92.7	87.9	99.4	**100**
UMT-L	5M	70.5	92.3	95.5	90.2	99.4	**100**
	17M	73.1	**93.2**	96.4	**90.8**	**100**	**100**
	25M	**73.3**	92.7	**96.6**	**90.8**	99.4	**100**

Table 15: Text-to-video retrieval on the temporally-heavy SSV2-label [36] and SSV2-template datasets [36].

Video question-answering. As shown in Table 16, our UMT outperforms the methods specifically designed for QA such as JustAsk [82], and achieves comparable performance with state-of-the-art models that pre-trained with large-scale pairs [83, 70, 81], which demonstrates its powerful capability of complex multimodal reasoning.

5. Conclusion

In this paper, we propose using the image foundation model as the unmasked teacher for masked video modeling. Besides, we present a progressive pre-training framework for building environmentally friendly video foundation models, which handles both scene-related and temporal-related actions, as well as complex video-language understanding. We hope that our simple, scalable, and reproducible framework will facilitate further research on video foundation models for future AI systems.

Method	#Pairs	ANet	MSR-QA	MSR-MC	MSVD-QA
ClipBERT [37]	0.2M	-	37.4	88.2	-
ALPRO [38]	5M	-	42.1	-	45.9
JustAsk [82]	69M	38.9	41.5	-	47.5
VideoCLIP [76]	136M	-	-	92.1	-
All-in-one [63]	138M	-	44.3	92.0	47.9
MERLOT [89]	180M	41.4	43.1	90.9	-
VIOLET [25]	138M	-	43.9	91.9	47.9
Singularity [36]	17M	44.1	43.9	93.7	-
OmniVL [64]	17M	-	44.1	-	51.0
VINDLU [14]	25M	44.7	44.6	97.1	-
FrozenBiLM [83]	400M	43.2	47.0	-	54.8
InternVideo [70]	646M	-	47.1	-	55.5
VideoCoCa [81]	4.8B	-	46.0	-	56.9
UMT-B	5M	43.5	44.3	95.9	49.1
	17M	44.9	44.9	96.3	48.9
	25M	44.8	44.9	96.3	49.5
UMT-L	5M	45.1	45.5	96.8	51.3
	17M	47.3	46.4	**97.7**	53.4
	25M	**47.9**	**47.1**	97.3	**55.2**

Table 16: Video question-answering on ActivityNet-QA, MSRVTT-QA, MSRVTT-MC and MSVD-QA.

Acknowledgement

This work was supported in part by the National Key R&D Program of China (No. 2022ZD0160100, No. 2022ZD0160505, No. 2022ZD0160900), the National Natural Science Foundation of China (No. 62076119), the Joint Lab of CAS-HK, the National Natural Science Foundation of China under Grant (No. 62272450), the Shenzhen Research Program (RCJC20200714114557087), and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2020355).

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*, 2017. 2, 8

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2, 7, 8

[3] 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*, 2022. 3

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 4, 8, 9

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *International Conference on Learning Representations*, 2021. 1

[6] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *ArXiv*, abs/2106.08254, 2021. 3

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, 2021. 3, 7, 8

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021. 6

[9] João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *ArXiv*, abs/1808.01340, 2018. 2, 3, 4, 6

[10] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *ArXiv*, abs/1907.06987, 2019. 2, 3, 4, 6

[11] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 4

[12] David L. Chen and William B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics*, 2011. 2, 8

[13] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *ArXiv*, abs/2202.03026, 2022. 3

[14] Feng Cheng, Xizi Wang, Jie Lei, David J. Crandall, Mohit Bansal, and Gedas Bertasius. Vindlu: A recipe for effective video-and-language pretraining. *ArXiv*, abs/2212.05051, 2022. 4, 8, 9

[15] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 2

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2018. 4

[18] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019. 3

[19] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Shuyang Gu, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Clip itself is a strong fine-tuner: Achieving 85.7% and 88.0% top-1 accuracy with vit-b and vit-l on imagenet. *ArXiv*, abs/2212.06138, 2022. 3, 6

[20] Xiaoyi Dong, Jianmin Bao, Ting Zhang, Dongdong Chen, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, and Nenghai Yu. Peco: Perceptual codebook for bert pre-training of vision transformers. *ArXiv*, abs/2111.12710, 2021. 3

[21] 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 on Learning Representations*, 2021. 1, 3, 4

[22] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021. 8

[23] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 2022. 1, 3, 7, 8

[24] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE/CVF International Conference on Computer Vision*, 2019. 7, 8

[25] 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*, abs/2111.12681, 2021. 2, 3, 8, 9

[26] Peng Gao, Teli Ma, Hongsheng Li, Jifeng Dai, and Yu Jiao Qiao. Convmae: Masked convolution meets masked autoencoders. *ArXiv*, abs/2205.03892, 2022. 3

[27] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The “something something” video database for learning and evaluating visual common sense.

In *IEEE International Conference on Computer Vision*, 2017. 2, 3, 4, 6

[28] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A. Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2, 3, 6

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3

[30] Zejiang Hou, Fei Sun, Yen-Kuang Chen, Yuan Xie, and S. Y. Kung. Milan: Masked image pretraining on language assisted representation. *ArXiv*, abs/2208.06049, 2022. 3, 4, 6

[31] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *ArXiv*, abs/2207.13532, 2022. 3

[32] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 2017. 3

[33] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017. 2, 3, 4, 6

[34] 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*, 2017. 2, 8

[35] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 2017. 4

[36] Jie Lei, Tamara L Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *ArXiv*, abs/2206.03428, 2022. 3, 4, 8, 9

[37] 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*, 2021. 3, 9

[38] 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*, 2022. 9

[39] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 2021. 4

[40] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, and Y. Qiao. Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer. *ArXiv*, abs/2211.09552, 2022. 2, 3, 4, 7, 8

[41] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*, 2022. 7, 8

[42] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *ArXiv*, abs/2206.07160, 2022. 2, 3, 9

[43] Yanghao Li, Chaoxia Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. *ArXiv*, abs/2112.01526, 2021. 7, 8

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014. 4

[45] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 7, 8

[46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 2019. 2, 3

[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. *ArXiv*, abs/2104.08860, 2022. 8

[48] 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*, 2022. 8, 9

[49] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa M. Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2, 6

[50] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. *ArXiv*, abs/2208.02816, 2022. 3

[51] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 2011. 4

[52] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *ArXiv*, abs/2208.06366, 2022. 2, 3, 6

[53] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. A unified view of masked image modeling. *ArXiv*, abs/2210.10615, 2022. 3, 6

[54] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2015. 1

[55] 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*, 2021. 1, 2, 3, 4, 6

[56] 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*, 2021. 3

[57] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Joseph Pal, H. Larochelle, Aaron C. Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 2016. 2, 8

[58] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[59] 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*, 2018. 4

[60] Fangxun Shu, Biaolong Chen, Yue Liao, Ke Gao, Shuwen Xiao, Wenyu Sun, Xiaobo Li, Yousong Zhu, Jinqiao Wang, and Si Liu. Masked contrastive pre-training for efficient video-text retrieval. *ArXiv*, abs/2212.00986, 2022. 2

[61] Chenxin Tao, Xizhou Zhu, Gao Huang, Y. Qiao, Xiaogang Wang, and Jifeng Dai. Siamese image modeling for self-supervised vision representation learning. *ArXiv*, abs/2206.01204, 2022. 3

[62] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Neural Information Processing Systems*, 2022. 1, 3, 4, 5, 7, 8

[63] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *ArXiv*, abs/2203.07303, 2022. 3, 4, 9

[64] 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. *ArXiv*, abs/2209.07526, 2022. 2, 3, 4, 8, 9

[65] Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, and Tong Lu. Memory-and-anticipation transformer for online action understanding, 2023. 1

[66] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. TDN: Temporal difference networks for efficient action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 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*, 2016. 3, 4

[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. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022. 3, 7, 8

[69] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 7

[70] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *ArXiv*, abs/2212.03191, 2022. 1, 3, 8, 9

[71] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3, 7, 8

[72] Longhui Wei, Lingxi Xie, Wen gang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *European conference on computer vision*, 2022. 3, 6

[73] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *ArXiv*, abs/2205.14141, 2022. 6

[74] 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. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[75] 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 IEEE international conference on Multimedia*, 2017. 2, 8

[76] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *ArXiv*, abs/2109.14084, 2021. 3, 9

[77] 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*, 2016. 2, 4, 8

[78] Hongwei Xue, Peng Gao, Hongyang Li, Yu Jiao Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. Stare at what you see: Masked image modeling without reconstruction. *ArXiv*, abs/2211.08887, 2022. 3

[79] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-

trained image-text model to video-language representation alignment. *ArXiv*, abs/2209.06430, 2022. 1, 2, 8, 9

[80] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 6, 7, 8

[81] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Video-text modeling with zero-shot transfer from contrastive captioners. *ArXiv*, abs/2212.04979, 2022. 8, 9

[82] 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*, 2021. 9

[83] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *ArXiv*, abs/2206.08155, 2022. 9

[84] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. *ArXiv*, abs/2212.14546, 2022. 9

[85] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 1, 2, 3, 7

[86] 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*, 2018. 2, 8

[87] 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*, 2019. 2, 8

[88] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *ArXiv*, abs/2111.11432, 2021. 3

[89] 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*, 2021. 3, 9

[90] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4

[91] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M. Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *ArXiv*, abs/2112.07175, 2021. 2, 7, 8

[92] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068, 2022. 2, 4

[93] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 2019. 1

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