Unmasked Teacher: Towards Training-Efficient Video Foundation Models

Supplementary Material

Stage	ViT-B	Output Size
Data	sparse sampling	$3\times8\times224\times224$
Patch Embedding	$1\times16\times16$, 768 stride $1\times16\times16$	$768\times8\times196$
Position Embedding	sine-cosine 768×1568	768×1568
Mask	semantic mask *mask ratio* $=\rho$	$768\times1568\cdot(1\text{-}\rho)$
Encoder	$\begin{bmatrix}\text{MHSA}(768)\\ \text{MLP}(3072)\end{bmatrix}\times12$	$768\times1568\cdot(1\text{-}\rho)$
Projection	$\begin{bmatrix}\text{LN}(768)\\ \text{MLP}(512)\end{bmatrix}\times K$	$K\times512\times1568\cdot(1\text{-}\rho)$

Table 17: **Architecture of video encoder.** We take ViT-B with 8-frame input as an example. "MHSA", "MLP" and "LN" refer to spatiotemporal multi-head self-attention, multi-layer perceptron and layer normalization. K means the layer number for unmasked token alignment. We mark the channel number, frame number, spatial size and token number by different colors.

config	SthSth V2	Kinetics
optimizer	AdamW [50]	
optimizer momentum	β_1, β_2=0.9, 0.95	
weight decay	0.05	
learning rate schedule	cosine decay [51]	
learning rate	1.2e-3	
batch size	2048	
warmup epochs [29]	40	
total epochs	default 200	
mask ratio	default 80%	
input frame	8	
drop path [34]	0	0.1 (B), 0.2 (L)
flip augmentation	*no*	*yes*
augmentation	MultiScaleCrop [0.66, 0.75, 0.875, 1]	

Table 18: **Stage-1 pre-training settings.**

config	5M & 17M & 25M
optimizer	AdamW [50]
optimizer momentum	β_1, β_2=0.9, 0.999
weight decay	0.02
learning rate schedule	cosine decay [51]
learning rate	1e-4
batch size	4096 (image), 4096 (video)
warmup epochs [29]	1
total epochs	10
mask ratio	50% (image), 80% (video), 50% (text)
input frame	4
drop path [34]	0.1 (B), 0.2 (L)
flip augmentation	*yes*
augmentation	MultiScaleCrop [0.5, 1]

Table 19: **Stage-2 pre-training settings.**

config	SthSth	Kinetics	MiT
optimizer	AdamW [50]		
optimizer momentum	β_1, β_2=0.9, 0.999		
weight decay	0.05		
learning rate schedule	cosine decay [51]		
learning rate	4e-4 (B), 8e-4 (L)	4e-4 (B), 2e-4 (L)	1e-4 (B/L)
batch size	512		
repeated augmentation	2	2	1
warmup epochs [29]	5	2	5
total epochs	30 (B), 17 (L)	35 (B), 20 (L)	40 (B), 20(L)
drop path [34]	0.1 (B), 0.2 (L)		
layer-wise lr decay [6]	0.75 (B), 0.85 (L)		
flip augmentation	*no*	*yes*	*yes*
label smoothing [67]	0.1		
cutmix [95]	1.0		
augmentation	RandAug(9, 0.5) [17]		

Table 20: **Action recognition fine-tuning settings.**

A. More implementation details

A.1. Model architecture and training details

In this section, we introduce the model architectures and training hyperparameters in our experiments.

Stage 1. In Stage 1, we train the video encoder from scratch, which is a vanilla ViT [23] without temporal downsampling. We use the same patch size for both ViT-B and ViT-L, *i.e.*, $1\times16\times16$ ($T\times H\times W$). To align with the unmasked teacher, we use a simple linear projection, including Layer Normalization [3] and one linear layer. The example architecture is shown in Table 17. For pre-training, we follow most of the hyperparameters in VideoMAE [69], as presented in Table 18. However, to prevent overfitting, we use drop path [34] in our approach.

Stage 2. In Stage 2, we equip the pre-trained video encoder with a text encoder and cross-modal decoder. Following Singularity [40], for the base model, we use the first 9 layers and the last 3 layers of BERT$_{base}$ to initialize the text encoder and decoder, respectively. While for our large model, we respectively adopt the first 19 layers and the 5 layers of BERT$_{large}$. For pre-training, we set all the loss weights to 1. And more details are shown in Table 19.

Action Recognition. We adopt the Stage-1 pre-trained video encoder and add an extra classification layer for fine-tuning. Detailed hyperparameters for different datasets are shown in Table 20. In our experiments, we have tried to fine-tune the Stage-2 pre-trained video encoder, but the results on Kinetics are similar.

Action Detection. Following VideoMAE [69] and ST-MAE [25], we add ROIAlign with MaxPooling to generate the regions of interest. Since we the Kinetics pre-trained models adopt sparse sampling [73], we use a frame span of 300 for action detection, which is the default frame number

config	AVA v2.2
optimizer	AdamW [50]
optimizer momentum	β_1, β_2=0.9, 0.999
weight decay	0.05
learning rate schedule	cosine decay [51]
learning rate	1.25e-4
batch size	128
warmup epochs [29]	5
total epochs	30 (B), 25 (L)
drop path [34]	0.2 (B), 0.4 (L)
layer-wise lr decay [6]	0.75 (B), 0.85 (L)
flip augmentation	*yes*

Table 21: **Action detection fine-tuning settings.**

config	ActivityNet	MSRVTT	MSVD
optimizer		AdamW [50]	
optimizer momentum		β_1, β_2=0.9, 0.999	
weight decay		0.02	
learning rate schedule		cosine decay [51]	
learning rate	4e-5 (B/L)	2e-5 (B/L)	2e-5 (B)
batch size		256	
warmup epochs [29]		1	
total epochs	12 (B), 10 (L)	8 (B/L)	15 (B), 6 (L)
input frame		12	
drop path [34]	0.2 (B), 0.3 (L)	0.2 (B), 0.4 (L)	0.2 (B), 0.4 (L)
flip augmentation		*yes*	
augmentation		MultiScaleCrop [0.5, 1]	

Table 22: **Video question-answering fine-tuning settings.**

of Kinetics videos. More details are listed in Table 21.

Video-text retrieval. For fine-tuning, we adopt the same architecture as in Stage 2, but we only apply VTC and VTM losses. For all datasets, we sparsely sample 12 frames for both training and testing. More details are listed in Table 24. For a fair comparison, we follow Singularity [40] to apply flip augmentation for SSV2 retrieval, which may harm the performance of this temporal-related dataset.

Video question-answering. Following the previous works [40, 15, 43], we formulate this task as text generation instead of classification. We add an extra multi-modal decoder that takes the output of the cross-modal decoder as the keys/values. And it decodes the answer text with "[CLS]" as a start. We follow [40, 15] to adopt the same architecture as the cross-modal decoder, and initialize it using the pre-trained cross-modal decoder. As for multiple-choice question-answering, we follow [40, 43, 15] to convert it to a text-to-video retrieval task, where the question and candidate answers are concatenated. The detailed hyperparameters are shown in Table 22 and Table 23.

B. More results

B.1. Video-text retrieval

Table 25 and Table 26 show more zero-shot and fine-tuned retrieval results on MARVTT [83], DiDeMo [1], ActivityNet [38], LSMDC [63] and MSVD [13].

config	MSRVTT-MC
optimizer	AdamW [50]
optimizer momentum	β_1, β_2=0.9, 0.999
weight decay	0.02
learning rate schedule	cosine decay [51]
learning rate	8e-5 (B), 4e-5 (L)
batch size	256
warmup epochs [29]	0
total epochs	5
input frame	12
drop path [34]	0.2 (B), 0.3 (L)
flip augmentation	*yes*
augmentation	MultiScaleCrop [0.5, 1]

Table 23: **Multi-choice video question-answering fine-tuning settings.**

B.2. Dataset descriptions

We show the statistics of pre-training datasets in Table 27, and downstream datasets in Table 28.

config	MSRVTT	DiDeMo	ActivityNet	LSMDC	MSVD	SSV2-label	SSV2-template
optimizer	AdamW [50]						
optimizer momentum	β_1, β_2=0.9, 0.999						
weight decay	0.02						
learning rate schedule	cosine decay [51]						
learning rate	2e-5 (B/L)	2e-5 (B), 4e-5 (L)	4e-5 (B/L)	2e-5 (B/L)	2e-5 (B/L)	5e-5 (B/L)	1e-4 (B/L)
batch size	256						
warmup epochs [29]	1						
total epochs	10 (B), 7(L)	12 (B), 5 (L)	20 (B/L)	10 (B), 8 (L)	10 (B/L)	10 (B/L)	10 (B), 8 (L)
input frame	12						
max text length	32	64	150	96	64	25	25
drop path [34]	0.2 (B), 0.3 (L)	0.1 (B), 0.3 (L)	0.1 (B), 0.2 (L)	0.1 (B), 0.2 (L)	0.2 (B), 0.3 (L)	0.1 (B), 0.2 (L)	0.1 (B), 0.2 (L)
flip augmentation	*yes*						
augmentation	MultiScaleCrop [0.5, 1]						

Table 24: **Video-text retrieval fine-tuning settings.**

Method	**#Pairs**	**Type**	**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
UMT-B	5M	T2V	29.6	52.8	61.9	33.4	58.3	67.0	28.3	53.0	64.2	16.8	30.5	37.6	55.7	83.7	92.2
		V2T	26.2	46.7	54.9	32.0	58.7	68.2	25.9	50.2	61.7	12.9	27.4	33.6	60.6	85.7	92.2
	17M	T2V	35.5	59.3	68.6	41.9	66.7	75.0	33.8	59.1	70.4	18.1	33.1	40.0	58.8	86.1	91.5
		V2T	31.6	53.5	64.1	40.3	66.6	75.8	31.6	56.2	67.9	16.0	29.9	35.7	61.9	86.9	91.3
	25M	T2V	35.2	57.8	66.0	41.2	65.4	74.9	35.5	60.6	71.8	19.1	33.4	42.2	60.3	86.7	91.9
		V2T	30.3	50.7	61.4	40.8	67.7	76.7	32.8	57.6	69.2	15.7	30.6	37.4	64.0	86.3	90.4
UMT-L	5M	T2V	33.3	58.1	66.7	34.0	60.4	68.7	31.9	60.2	72.0	20.0	37.2	43.7	68.1	92.1	95.2
		V2T	30.2	51.3	61.6	36.2	60.0	68.6	30.0	59.1	71.3	16.1	32.0	39.2	68.1	92.5	96.3
	17M	T2V	**42.6**	**64.4**	**73.1**	46.4	70.0	78.8	**42.8**	**69.6**	79.8	**25.2**	**43.0**	50.5	71.0	93.3	96.4
		V2T	**38.6**	**59.8**	**69.6**	46.5	72.2	79.5	**40.7**	**67.6**	**78.6**	**23.2**	37.7	44.2	69.1	91.5	94.8
	25M	T2V	40.7	63.4	71.8	**48.6**	**72.9**	**79.0**	41.9	68.9	**80.3**	24.9	41.7	**51.8**	**72.2**	**94.2**	**96.9**
		V2T	37.1	58.7	68.9	**49.9**	**74.8**	**81.4**	39.4	66.8	78.3	21.9	**37.8**	**45.7**	**72.4**	**93.4**	**95.8**

Table 25: **Zero-shot** retrieval results on MSRVTT, DiDeMo, AcitivityNet, LSMDC, and MSVD.

Method	**#Pairs**	**Type**	**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
UMT-B	5M	T2V	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
		V2T	44.4	72.8	80.7	52.9	80.2	85.8	50.0	79.8	88.2	29.8	52.2	60.5	67.0	92.5	96.3
	17M	T2V	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
		V2T	49.4	76.7	83.5	59.5	83.8	90.7	54.6	82.1	91.1	31.5	53.6	61.9	71.3	93.9	97.2
	25M	T2V	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
		V2T	49.0	77.0	84.7	59.5	84.9	90.5	56.0	83.5	91.7	32.7	53.5	63.2	74.0	94.6	97.3
UMT-L	5M	T2V	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
		V2T	51.4	76.3	82.8	59.5	84.5	90.7	55.4	84.4	92.9	36.2	58.9	65.7	73.6	96.3	98.1
	17M	T2V	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
		V2T	56.7	79.6	86.7	**66.4**	87.5	92.9	64.3	87.8	**94.8**	40.3	63.1	71.1	78.1	97.6	**98.7**
	25M	T2V	**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**
		V2T	**58.6**	**81.6**	**86.5**	65.7	**89.6**	**93.3**	**64.4**	**89.1**	**94.8**	**41.4**	**64.3**	**71.5**	**81.2**	**96.7**	**98.7**

Table 26: **Fine-tuned** retrieval results on MSRVTT, DiDeMo, AcitivityNet, LSMDC, and MSVD.

Dataset	#image/video	#text	Type
Kinetics-710 [44]	658K	0	Video
COCO [48]	113K	567K	image
Visual Genome [39]	100K	768K	image
SBU Captions [57]	860K	860K	image
CC3M [65]	2.88M	2.88M	image
CC12M [12]	11.00M	11.00M	image
WebVid-2M [5]	2.49M	2.49M	video
WebVid-10M [5]	10.73M	10.73M	video
5M corpus = CC3M+WebVid-2M	5.37M	5.37M	video+image
17M corpus = 5M+COCO+VG+SBU+CC12M	17.44M	18.57M	video+image
25M corpus = 17M+WebVid-10M−WebVid-2M	25.68M	26.81M	video+image

Table 27: **Statistics of pre-training datasets.**

Dataset	#video			#text			Avg Video
	Train	Val	Test	Train	Val	Test	Length (s)
Action Recognition							
Kinetics-400 [37]	240,436	19,787	-	-	-	-	10
Kinetics-600 [10]	366,006	27,935	-	-	-	-	10
Kinetics-700 [11]	529,573	33,861	-	-	-	-	10
Moments in Time V1 [55]	802,244	33,899	-	-	-	-	3
Something-Something V2 [30]	168,913	24,777	-	-	-	-	4
Action Detection							
AVA v2.2 [31]	235	64	131	-	-	-	900
Video-Text Retrieval							
MSRVTT [83]	7,010	-	1,000	140,200	-	1,000	15
DiDeMo [1]	8,496	1,094	1,036	8,496	1,094	1,036	29.3
ActivityNet Captions [38]	10,009	4,917	-	10,009	4,917	-	180
LSMDC [63]	101,055	-	1,000	101,055	-	1,000	4.7
MSVD [13]	1,200	100	670	1,200	100	670	15
SSV2-Template [40]	168,913	-	2,088	174	-	174	4
SSV2-Label [40]	168,913	-	2,088	109,968	-	1,989	4
Video Question-Answering							
ActivityNet-QA [93]	3,200	1,800	800	32,000	18,000	8,000	180
MSRVTT-QA [81]	6,513	497	2,990	158,581	12,278	72,821	15
MSRVTT-MC [92]	7,010	-	2,990	140,200	-	14,950	15
MSVD-QA [81]	1,161	245	504	29,883	6,415	13,157	15

Table 28: **Statistics of downstream datasets.**