VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training

Zhan Tong [1,2*] **Yibing Song** [2] **Jue Wang** [2] **Limin Wang** [1,3†]
[1]State Key Laboratory for Novel Software Technology, Nanjing University
[2]Tencent AI Lab [3]Shanghai AI Lab
tongzhan@smail.nju.edu.cn {yibingsong.cv, arphid}@gmail.com lmwang@nju.edu.cn

Abstract

Pre-training video transformers on extra large-scale datasets is generally required to achieve premier performance on relatively small datasets. In this paper, we show that video masked autoencoders (VideoMAE) are data-efficient learners for self-supervised video pre-training (SSVP). We are inspired by the recent ImageMAE [31] and propose customized video tube masking with an extremely high ratio. This simple design makes video reconstruction a more challenging and meaningful self-supervision task, thus encouraging extracting more effective video representations during the pre-training process. We obtain three important findings with VideoMAE: (1) An extremely high proportion of masking ratio (i.e., 90% to 95%) still yields favorable performance for VideoMAE. The temporally redundant video content enables higher masking ratio than that of images. (2) VideoMAE achieves impressive results on very small datasets (i.e., around 3k-4k videos) without using any extra data. This is partially ascribed to the challenging task of video reconstruction to enforce high-level structure learning. (3) VideoMAE shows that data quality is more important than data quantity for SSVP. Domain shift between pre-training and target datasets is an important factor. Notably, our VideoMAE with the vanilla ViT backbone can achieve 87.4% on Kinects-400, 75.4% on Something-Something V2, 91.3% on UCF101, and 62.6% on HMDB51, without using any extra data. Code is available at https://github.com/MCG-NJU/VideoMAE.

1 Introduction

Transformer [71] has brought significant progress in natural language processing [18, 7, 55]. The vision transformer [21] also improves a series of computer vision tasks including image classification [67, 89], object detection [8, 38], semantic segmentation [81], object tracking [14, 17], and video recognition [6, 3]. The multi-head self-attention upon linearly projected image/video tokens is capable of modeling global dependency among visual content either spatially or temporally. The inductive bias is effectively reduced via this flexible attention mechanism.

Training effective vision transformers (ViTs) typically necessitates large-scale supervised datasets. Initially, the pre-trained ViTs achieve favorable performance by using hundreds of millions of labeled images [21]. For video transformers [3, 6], they are usually derived from image-based transformers and heavily depend on the pre-trained models from large-scale image data (e.g., ImageNet [58]). Previous trials [3, 6] on training video transformers from scratch yield unsatisfied results (except for MViT [22] with a strong inductive bias). Therefore, the learned video transformers are naturally biased by image-based models, and it still remains a challenge that *how to effectively and efficiently train a vanilla vision transformer on the video dataset itself without using any pre-trained model*

*Work is done during internship at Tencent AI Lab. †Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

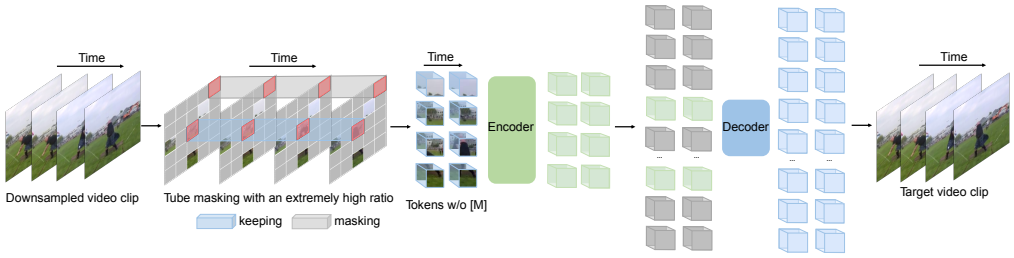


Figure 1: **VideoMAE** performs the task of masking random cubes and reconstructing the missing ones with an asymmetric encoder-decoder architecture. Due to high redundancy and temporal correlation in videos, we present the customized design of tube masking with an extremely high ratio (90% to 95%). This simple design enables us to create a more challenging and meaningful self-supervised task to make the learned representations capture more useful spatiotemporal structures.

or extra image data. Moreover, the existing video datasets are relatively small compared with image datasets, which further increases the difficulty of training video transformers from scratch. Meanwhile, self-supervised learning has shown remarkable performance by using large-scale image datasets [15, 9]. The learned representations have outperformed the ones via supervised learning when being transferred to downstream tasks. It is expected that this self-supervised learning paradigm can provide a promising solution to address the challenge of training video transformers.

Following the success of masked autoencoding in NLP [18] and images [31, 4], we present a new self-supervised video pre-training (SSVP) method, termed as *Video Masked Autoencoder* (VideoMAE). Our VideoMAE inherits the simple pipeline of masking random cubes and reconstructing the missing ones. However, the extra time dimension of videos makes them different from images in this masked modeling. First, video frames are often densely captured, and their semantics varies slowly in time [88]. This temporal redundancy would increase the risk of recovering missing pixels from the spatiotemporal neighborhood with little high-level understanding. Furthermore, video could be viewed as the temporal evolution of static appearance, and there exists a correspondence between frames. This temporal correlation could lead to information leakage (i.e., masked spatiotemporal content re-occurrence) during reconstruction unless a specific masking strategy is considered. In this sense, for each masked cube, it is easy to find a corresponding and unmasked copy in adjacent frames. This property would make the learned models identify some "shortcut" features that are hard to generalize to new scenarios.

To make video masked modeling more effective, in this paper, we present a customized design of tube masking with an extremely high ratio in our VideoMAE. First, due to temporal redundancy, we use an *extremely high* masking ratio to drop the cubes from the downsampled clips. This simple strategy not only effectively increases the pre-training performance but also greatly reduces the computational cost due to the asymmetric encoder-decoder architecture. Second, to consider temporal correlation, we devise a simple yet effective *tube masking* strategy, which turns out to be helpful in relieving the risk of information leakage for cubes with no or negligible motion during reconstruction. With this simple yet effective design in our VideoMAE, we are able to successfully train vanilla ViT backbones on the relatively small-scale video datasets such as Something-Something [26], UCF101 [61], and HMDB51 [35], which significantly outperform the previous state of the art under the setting without extra data. In summary, the main contribution of this paper is threefold:

- We present a simple but effective video masked autoencoder that unleashes the potential of vanilla vision transformer for video recognition. To the best of our knowledge, this is the first masked video pre-training framework of simply using plain ViT backbones. To relieve the information leakage issue in masked video modeling, we present the tube masking with an extremely high ratio, which brings the performance improvement to the VideoMAE.
- Aligned with the results in NLP and Images on masked modeling, our VideoMAE demonstrates that this simple masking and reconstruction strategy provides a good solution to self-supervised video pre-training. The models pre-trained with our VideoMAE significantly outperform those trained from scratch or pre-trained with contrastive learning methods.
- We obtain extra important findings on masked modeling that might be ignored in previous research in NLP and Images. (1) We demonstrate that VideoMAE is a data-efficient learner that could be successfully trained with only 3.5k videos. (2) Data quality is more important than quantity for SSVP when a domain shift exists between the source and target dataset.

2 Related Work

Video representation learning. Learning good video representations has been heavily investigated in the literature. The supervised learning methods [59, 76, 70, 10, 6] usually depend on the image backbones. The video encoder backbones are first pre-trained with image data in a supervised form. Then, these backbones are fine-tuned on the video dataset for classifying human actions. Meanwhile, some methods [68, 23, 22] directly train video backbones from videos in a supervised manner. Besides supervised learning, semi-supervised video representation learning has also been studied [60]. The representations of labeled training samples are utilized to generate supervision signals for unlabeled ones. Supervised or semi-supervised representation learning mainly uses a top-down training paradigm, which is not effective in exploring the inherent video data structure itself. Meanwhile, some multimodal contrastive learning methods [37, 43, 63] have been developed to learn video representation from noisy text supervision.

For self-supervised learning, the prior knowledge of temporal information has been widely exploited to design pretext tasks [79, 45, 83, 5] for SSVP. Recently, contrastive learning [29, 46, 30, 53, 25, 28] is popular to learn better visual representation. However these methods heavily rely on strong data augmentation and large batch size [24]. Predicting the video clip with autoencoders in pixel space has been explored for representation learning by using CNN or LSTM backbones [49, 62], or conducting video generation with autoregressive GPT [84]. Instead, our VideoMAE aims to use the simple masked autoencoder with recent ViT backbones to perform data-efficient SSVP.

Masked visual modeling. Masked visual modeling has been proposed to learn effective visual representations based on the simple pipeline of masking and reconstruction. These works mainly focus on the image domain. The early work [73] treated the masking and a noise type in denoised autoencoders [72] or inpainted missing regions with context [48] by using convolutions. iGPT [11] followed the success of GPT [7, 56] in NLP and operated a sequence of pixels for prediction. The original ViT [21] investigated the masked token prediction for self-supervised pre-training. More recently, the success of vision transformer has led to investigation of Transformer-based architectures for masked visual modeling [4, 20, 31, 80, 82, 90]. BEiT [4], BEVT [77] and VIMPAC [65] followed BERT [18] and proposed to learn visual representations from images and videos by predicting the discrete tokens [57]. MAE [31] introduced an asymmetric encoder-decoder architecture for masked image modeling. MaskFeat [80] proposed to reconstruct the HOG features of masked tokens to perform self-supervised pre-training in videos. VideoMAE is inspired by the ImageMAE and introduces specific design in implementation for SSVP. In particular, compared with previous masked video modeling [31, 77, 65], we present a simpler yet more effective video masked autoencoder by directly reconstructing the pixels. Our VideoMAE is the first masked video pre-training framework of simply using plain ViT backbones.

3 Proposed Method

In this section, we first revisit ImageMAE [31]. Then we analyze the characteristics of video data. Finally, we show how we explore MAE in the video data by presenting our VideoMAE.

3.1 Revisiting Image Masked Autoencoders

ImageMAE [31] performs the masking and reconstruction task with an asymmetric encoder-decoder architecture. The input image $I \in \mathcal{R}^{3\times H\times W}$ is first divided into regular non-overlapping patches of size 16 $\times$16, and each patch is represented with *token embedding*. Then a subset of tokens are randomly masked with a high masking ratio (75%), and only the remaining ones are fed into the transformer *encoder* Φ_{enc}. Finally, a shallow *decoder* Φ_{dec} is placed on top of the visible tokens from the encoder and learnable mask tokens to reconstruct the image. The *loss function* is mean squared error (MSE) loss between the normalized masked tokens and reconstructed ones in the pixel space:

$$\mathcal{L} = \frac{1}{\Omega}\sum_{p\in\Omega} |I(p) - \hat{I}(p)|^2, \tag{1}$$

where p is the token index, Ω is the set of masked tokens, I is the input image, and $\hat{I}$ is the reconstructed one.

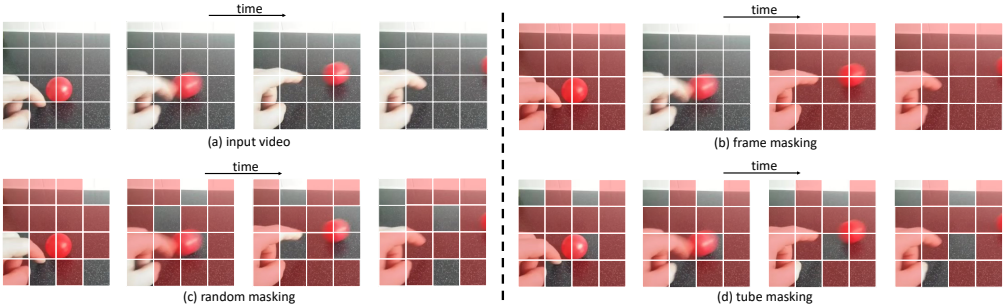


Figure 2: Slowness is a general prior in (a) video data [88]. This leads to two important characteristics in time: temporal redundancy and temporal correlation. Temporal redundancy makes it possible to recover pixels under an extremely high masking ratio. Temporal correlation leads to easily reconstruct the missing pixels by finding those corresponding patches in adjacent frames under plain (b) frame masking or (c) random masking. To avoid this simple task and encourage learning representative representation, we propose a (d) tube masking, where the masking map is the same for all frames.

3.2 Characteristics of Video Data

Compared with static images, video data contain temporal relations. We show the motivation of our VideoMAE by analyzing video characteristics.

Temporal redundancy. There are frequently captured frames in a video. The semantics vary slowly in the temporal dimension [88]. We observe that consecutive frames are highly redundant, as shown in Figure 2. This property leads to two critical issues in masked video autoencoding. First, it would be less efficient to keep the original temporal frame rate for pre-training. This would draw us to focus more on static or slow motions in our masked modeling. Second, temporal redundancy greatly dilutes motion representations. This would make the task of reconstructing missing pixels not difficult under the normal masking ratio (e.g., 50% to 75%). The encoder backbone is not effective in capturing motion representations.

Temporal correlation. Videos could be viewed as the temporal extension of static appearance, and therefore there exists an inherent correspondence between adjacent frames. This temporal correlation could increase the risk of information leakage in the masking and reconstruction pipeline. In this sense, as shown in Figure 2, we can reconstruct the masked patches by finding the spatiotemporal corresponding unmasked patches in the adjacent frames under plain random masking or frame masking. In this case, it might guide the VideoMAE to learn low-level temporal correspondence rather than high-level information such as spatiotemporal reasoning over the content. To alleviate this behavior, we need to propose a new masking strategy to make the reconstruction more challenging and encourage effective learning of spatiotemporal structure representations.

3.3 VideoMAE

To relieve the above issues in video masked modeling, we make the customized design in our VideoMAE, and the overall pipeline is shown in Figure 1. Our VideoMAE takes the *downsampled frames* as inputs and uses the *cube embedding* to obtain video tokens. Then, we propose a simple design of *tube masking with high ratio* to perform MAE pre-training with an asymmetric encoder-decoder architecture. Our backbone uses the vanilla ViT with *joint space-time attention*.

Temporal downsampling. According to the above analysis on temporal redundancy over consecutive frames, we propose to use the strided temporal sampling strategy to perform more efficient video pre-training. Formally, one video clip consisting of t consecutive frames is first randomly sampled from the original video V. We then use temporal sampling to compress the clip to T frames, each of which contains $H \times W \times 3$ pixels. In experiments, the stride τ is set to 4 and 2 on Kinetics and Something-Something, respectively.

Cube embedding. We adopt the joint space-time cube embedding [3, 22, 39] in our VideoMAE, where we treat each cube of size $2 \times 16 \times 16$ as one token embedding. Thus, the cube embedding layer obtains $\frac{T}{2} \times \frac{H}{16} \times \frac{W}{16}$ 3D tokens and maps each token to the channel dimension D. This design can decrease the spatial and temporal dimension of input, which helps to alleviate the spatiotemporal redundancy in videos.

Tube masking with extremely high ratios. First, temporal redundancy is a factor affecting VideoMAE design. We find that VideoMAE is in favor of extremely high masking ratios (e.g. 90% to 95%) compared with the ImageMAE. Video information density is much lower than images, and we expect a high ratio to increase the reconstruction difficulty. This high masking ratio is helpful to mitigate the information leakage during masked modeling and make masked video reconstruction a meaningful self-supervised pre-training task.

Second, temporal correlation is another factor in our VideoMAE design. We find even under the extremely high masking ratio, we can still improve the masking efficiency by proposing the temporal tube masking mechanism. Temporal tube masking enforces a mask to expand over the whole temporal axis, namely, different frames sharing the same masking map. Mathematically, the tube mask mechanism can be expressed as $\mathbb{I}[p_{x,y,\cdot} \in \Omega] \sim \mathrm{Bernoulli}(\rho_{\mathrm{mask}})$ and different time t shares the same value. With this mechanism, temporal neighbors of masked cubes are always masked. So for some cubes with no or small motion (e.g., finger cube in 4th row of Figure 2 (d)), we can not find the spatiotemporal corresponding content in all frames. In this way, it would encourage our VideoMAE to reason over high-level semantics to recover these totally missing cubes. This simple strategy can alleviate the information leakage for cubes with no or negligible motion, and turns out to be effective in practice for masked video pre-training.

Backbone: joint space-time attention. Due to the high proportion of masking ratio mentioned above, only a few tokens are left as the input for the encoder. To better capture high-level spatio-temporal information in the remaining tokens, we use the vanilla ViT backbone [21] and adopt the joint space-time attention [3, 39]. Thus, all pair tokens could interact with each other in the multi-head self-attention layer [71]. The specific architecture design for the encoder and decoder is shown in supplementary materials. The quadratic complexity of the joint space-time attention mechanism is a computational bottleneck, while our design of an extremely high masking ratio alleviates this issue by only putting the unmasked tokens (e.g., 10%) into the encoder during the pre-training phase.

4 Experiments

4.1 Datasets

We evaluate our VideoMAE on five common video datasets: Kinetics-400 [34], Something-Something V2 [26], UCF101 [61], HMDB51 [35], and AVA [27]. The Kinetics-400 contains around 240k training videos and 20k validation videos of 10s from 400 classes. The Something-Something V2 is another large-scale video dataset, having around 169k videos for training and 20k videos for validation. In contrast to Kinetics-400, this dataset contains 174 motion-centric action classes. These two large-scale video datasets focus on different visual cues for action recognition. UCF101 and HMDB51 are two relatively small video datasets, which contain around 9.5k/3.5k train/val videos and 3.5k/1.5k train/val videos, respectively. Compared with those large-scale video datasets, these two small datasets are more suitable for verifying the effectiveness of VideoMAE, as training large ViT models is more challenging on small datasets. Moreover, we also transfer the learned ViT models by VideoMAE to downstream action detection task. We work on AVA, a dataset for spatiotemporal localization of human actions with 211k training and 57k validation video segments. In experiments of downstream tasks, we fine-tune the pre-trained VideoMAE models on the training set and report the results on the validation set. The implementation details are described in Appendix § B.

4.2 Ablation Studies

In this subsection, we perform in-depth ablation studies on VideoMAE design with the default backbone of 16-frame ViT-B on Something-Something V2 (SSV2) and Kinetics-400 (K400). The specific architectures for the encoder and decoder are shown in Appendix § A. For fine-tuning, we perform TSN [76] uniform sampling on SSV2 and dense sampling [78, 23] on K400. All models share the same inference protocol, i.e., 2 clips $\times$ 3 crops on SSV2 and 5 clips $\times$ 3 crops on K400.

Decoder design. The lightweight decoder is one key component of our VideoMAE. We conduct experiments with the different depths in Table 1a. Unlike in ImageMAE, a deep decoder here is important for better performance, while a shallow decoder could reduce the GPU memory consumption. We take 4 blocks for the decoder by default. The decoder width is set to half channel of the encoder (e.g., 384-d for ViT-B), following the design in the image domain.

blocks	SSV2	K400	GPU mem.
1	68.5	79.0	7.9G
2	69.2	79.2	10.2G
4	**69.6**	**80.0**	14.7G
8	69.3	79.7	23.7G

(a) **Decoder depth**. 4 blocks of decoder achieve the best trade-off. "GPU mem." is GPU memory during pre-training, benchmarked in one GPU with a batch size of 16.

case	ratio	SSV2	K400
tube	75	68.0	79.8
tube	90	**69.6**	**80.0**
random	90	68.3	79.5
frame	87.5*	61.5	76.5

(b) **Mask sampling**. We compare different masking strategies. Our proposed tube masking with an extremely high ratio works the best. $*$"87.5" means masking 14/16 frames.

input	target	SSV2	K400
$T\times\tau$	*center*	63.0	79.3
$T\times\frac{\tau}{2}$	$T\times\frac{\tau}{2}$	68.9	79.8
$T\times\tau$	$T\times\tau$	**69.6**	80.0
$T\times\tau$	$2T\times\frac{\tau}{2}$	69.2	**80.1**

(c) **Reconstruction target**. $T\times\tau$ denotes "frames $\times$stride". *center* denotes the center frame of the input clip. T is set to 16 as default. τ is set to 2 and 4 on SSV2 and K400, respectively.

case	SSV2	K400
from scratch	32.6	68.8
ImageNet-21k sup.	61.8	78.9
IN-21k+K400 sup.	65.2	-
VideoMAE	**69.6**	**80.0**

(d) **Pre-training strategy**. Our VideoMAE works the best without using any extra data. "sup." is supervised training.

dataset	method	SSV2	K400
IN-1K	ImageMAE	64.8	78.7
K400	VideoMAE	68.5	**80.0**
SSV2	VideoMAE	**69.6**	79.6

(e) **Pre-training dataset**. Our VideoMAE works the best when directly pre-training the models on the source datasets.

case	SSV2	K400
L1 loss	69.1	79.7
MSE loss	**69.6**	**80.0**
Smooth L1 loss	68.9	79.6

(f) **Loss function**. MSE loss works the best for the masking and reconstruction task in VideoMAE.

Table 1: Ablation experiments on **Something-Something V2** and **Kinetics-400**. Our backbone is 16-frame vanilla ViT-B and all models are pre-trained with *mask ratio* ρ=90% for 800 epochs, and fine-tuned for evaluation. We perform TSN [76] uniform sampling on SSV2 and dense sampling [78, 23] on K400. All models share the same inference protocol, i.e., 2 clips $\times$ 3 crops on SSV2 and 5 clips $\times$ 3 crops on K400. The default choice for our model is colored in gray.

Masking strategy. We compare different masking strategies in Table 1b. When increasing the masking ratio from 75% to 90% for tube masking, the performance on SSV2 boosts from 68.0% to 69.6%. Then, with an extremely high ratio, we find tube masking also achieves better performance than plain random masking and frame masking. We attribute these interesting observations to the redundancy and temporal correlation in videos. The conclusion on K400 is in accord with one on SSV2. One may note that the performance gap on K400 is lower than one on SSV2. We argue that the Kinetics videos are mostly stationary and scene-related. The effect of temporal modeling is not obvious. Overall, we argue that our default designs enforce the networks to capture more useful spatiotemporal structures and therefore make VideoMAE a more challenging task, which a good self-supervised learner hunger for.

Reconstruction target. First, if we only employ the center frame as the target, the results would decrease greatly as shown in Table 1c. The sampling stride is also sensitive. The result of small sampling stride $\frac{\tau}{2}$ is lower than default sampling stride τ (68.9% vs. 69.6% on SSV2). We also try to reconstruct $2T$ frames from the downsampled T frames, but it obtains slightly worse results on SSV2. For simplicity, we use the input downsampled clip as our default reconstruction target.

Pre-training strategy. We compare different pre-training strategies in Table 1d. Similar to previous trials [3, 6], training video transformers from scratch yields unsatisfied results on video datasets. When pre-trained on the large-scale ImageNet-21K dataset, the video transformer obtains better accuracy from 32.6% to 61.8% on SSV2 and 68.8% to 78.9% on K400. Using the models pre-trained on both ImageNet-21K and Kinetics further increases accuracy to 65.2% on SSV2. Our VideoMAE can effectively train a video transformer on the video dataset itself without using any extra data and achieve the best performance (69.6% on SSV2 and 80.0% on K400).

Pre-training dataset. First, we pre-train the ViT-B on ImageNet-1K for 1600 epochs, following the recipes in [31]. Then we inflate the 2D patch embedding layer to our cube embedding layer following [10] and fine-tune the model on the target video datasets. The results surpass the model trained *from scratch* as shown in Table 1e. We also compare the ImageMAE pre-trained model with VideoMAE models pre-trained on video datasets. We see that our VideoMAE models can achieve better performance than ImageMAE. However, when we try to transfer the pre-trained VideoMAE models to the other video datasets (e.g. from Kinetics to Something-Something), the results are slightly worse than their counterpart, which is directly pre-trained on its own target video datasets. We argue that domain shift between pre-training and target datasets could be an important issue.

dataset	training data	*from scratch*	MoCo v3	VideoMAE
K400	240k	68.8	74.2	**80.0**
Sth-Sth V2	169k	32.6	54.2	**69.6**
UCF101	9.5k	51.4	81.7	**91.3**
HMDB51	3.5k	18.0	39.2	**62.6**

Table 2: Comparisons with the results of previous **self-supvised pre-training methods** on different datasets. We take 16-frame ViT-B as the default backbone. Notably, here MoCo v3 and VideoMAE all only use the *unlabelled* data in the training set of each dataset for pre-training and are all fine-tuned for evaluation.

method	epoch	ft. acc.	lin. acc.	hours	speedup
MoCo v3	300	54.2	33.7	61.7	-
VideoMAE	800	**69.6**	38.9	19.5	**3.2**$\times$

Table 3: Comparisons with the **efficiency and effectiveness** on Something-Something V2. We report the fine-tuning (ft) and linear probing (lin) accuracy (%). The **wall-clock time** of pre-training is benchmarked in 64 Tesla V100 GPUs with PyTorch.

method	K400 $\rightarrow$ SSV2	K400 $\rightarrow$ UCF	K400 $\rightarrow$ HMDB
MoCo v3	62.4	93.2	67.9
VideoMAE	**68.5**	**96.1**	**73.3**

Table 4: Comparisons with the **feature transferability** on smaller datasets. We take 16-frame ViT-B as the default backbone. Notably, here MoCo v3 and VideoMAE are all pre-trained on Kinetics-400 with *unlabelled* data in the training set. Then the pre-trained model is fine-tuned on target datasets for evaluation.

Loss function. Table 1f contains an ablation study of loss function. We find that the MSE loss could achieve a higher result compared with the L1 loss and smooth L1 loss. Therefore, we employ the MSE loss by default.

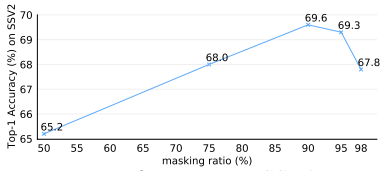
4.3 Main Results and Analysis

VideoMAE: data-efficient learner. The self-supervised video pre-training (SSVP) has been extensively studied in previous works, but they mainly use the CNN-based backbones. Few works have investigated transformer-based backbone in SSVP. Therefore, to demonstrate the effectiveness of VideoMAE for transformer-based SSVP, we compare two methods implemented by ourselves: (1) training from scratch and (2) pre-training with contrastive learning (MoCo v3 [15]). For training from scratch, we carefully tune these hyper-parameters to successfully pre-train ViT-Base from the training set of the dataset. For pre-training with MoCo v3, we strictly follow the training practice in its image counterpart and carefully avoid the collapse issue.

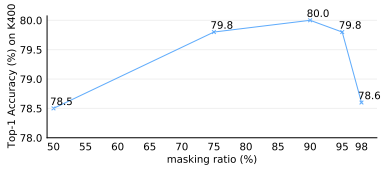
The recognition accuracy is reported in Table 2. We see that our VideoMAE significantly outperforms other two training settings. For instance, on the largest dataset of Kinetics-400, our VideoMAE outperforms training from scratch by around 10% and MoCo v3 pre-training by around 5%. This superior performance demonstrates that masked autoencoder provides an effective pre-training mechanism for video transformers. We also see that the performance gap between our VideoMAE and the other two methods becomes larger as the training set becomes smaller. Notably, even with only 3.5k training clips on HMDB51, our VideoMAE pre-training can still obtain a satisfying accuracy (around 61%). This new result demonstrates that VideoMAE is a more data-efficient learner for SSVP. This property is particularly important for scenarios with limited data available and different with contrastive learning methods.

We compare the efficiency of VideoMAE pre-training and MoCo v3 pre-training in Table 3. The task of masked autoencoding with a high ratio is more challenging and thereby requires more training epochs (800 vs. 300). Thanks to the asymmetric encoder-decoder in our VideoMAE and extremely high masking ratio, our pre-training time is much shorter than MoCo v3 (19.5 vs. 61.7 hours).

High masking ratio. In VideoMAE, one core design is the extremely high masking ratio. We perform an investigation of this design on the Kinetics-400 and Something-Something V2 datasets. The results are shown in Figure 3. We see that the best masking ratio is extremely high, and even 95% can achieve good performance for both datasets. This result is difference from BERT [18] in NLP and MAE [31] in images. We analyze the temporal redundancy and correlation in videos makes it possible for our VideoMAE to learn plausible outputs with such a high masking ratio.



(a) Performance on SSV2



(b) Performance on Kinetics-400

Figure 3: The effect of **masking ratio** on (a) Something-Something V2 and (b) Kinetics-400. We take 16-frame vanilla ViT-B as default. The results show that an extremely high masking ratio (90%) achieves the best efficiency and effectiveness trade-off on both video datasets.

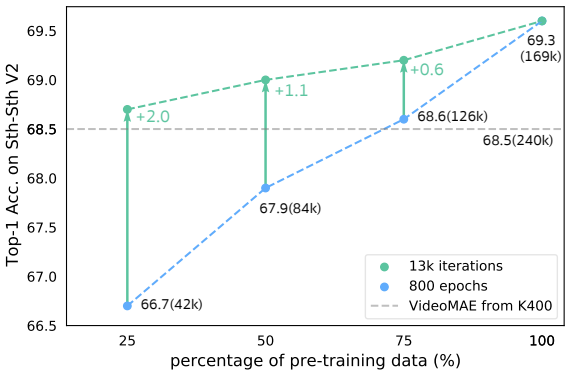


Figure 4: **Data efficiency** of VideoMAE representations. Our default backbone is 16-frame vanilla ViT-B. • denotes that all models are trained for the **same** 132k iterations, and • denotes that all models are trained for the **same** 800 epochs. Note that it takes 132k iterations to pre-train the model for 800 epochs on the full training set of Something-Something V2.

We also visualize the reconstructed examples in Appendix § E. We see that even under an extremely high masking ratio, VideoMAE can produce satisfying reconstructed results. This implies VideoMAE is able to learn useful representations that capture the holistic spatiotemporal structure in videos.

Transfer learning: quality vs. quantity. To further investigate the generalization ability of VideoMAE in representation learning, we transfer the learned VideoMAE from Kinetics-400 to Something-Something V2, UCF101, and HMDB51. The results are shown in Table 4, and we compare them with MoCo v3 pre-training. The models pre-trained by VideoMAE are better than those pre-trained by MoCo v3, demonstrating that our VideoMAE learns more transferable representations.

Comparing Table 2 and Table 4, the transferred representation outperforms the original VideoMAE models trained from its own dataset on UCF101 and HMDB51. In contrast, the transferred representation is worse on Something-Something V2. To figure out whether this inconsistent result is caused by the large scale of Something-Something V2, we further perform a detailed investigation by decreasing the pre-training video numbers. In this study, we run two experiments: (1) pre-training with the same epochs and (2) pre-training with the same time budget. The result is shown in Figure 4. We see that more training iterations could contribute to better performance when we decrease the size of the pre-training set. Surprisingly, even with only 42k pre-training videos, we can still obtain better accuracy than the Kinetics pre-trained models with 240k videos (68.7% vs. 68.5%). This result implies that domain shift is another important factor, and data quality is more important than data quantity in SSVP when there exists a difference between pre-training and target datasets. It also demonstrates that VideoMAE is a data-efficient learner for SSVP.

Transfer learning: downstream action detection. We also transfer the learned VideoMAE on Kinetics-400 to downstream action detection dataset AVA. Following the standard setting [27], we evaluate on top 60 common classes with mean Average Precision (mAP) as the metric under IoU threshold of 0.5. The results are shown in the Table 5. After self-supervised pre-training on Kinetics-400, our VideoMAE with the vanilla ViT-B can achieve 26.7 mAP on AVA, which demonstrates the strong transferability of our VideoMAE. If the pre-trained ViT-B is additionally fine-tuned on Kinetics-400 with labels, the transfer learning performance can further increase about 5 mAP (from 26.7 to 31.8). More remarkably, when we scale up the pre-training configurations with larger video datasets (e.g. Kinetics-700) or more powerful backbones (e.g. ViT-Large and ViT-Huge), VideoMAE can finally obtain better performance. For example, our ViT-L VideoMAE pre-trained on Kinetics-700 achieves 39.3 mAP and ViT-H VideoMAE pre-trained on Kinetics-400 has 39.5 mAP. These results demonstrate that the self-supervised pre-trained models transfer well not only on action classification task but on more complex action detection task.

Method	Backbone	Pre-train Dataset	Extra Labels	$T \times \tau$	GFLOPs	Param	mAP
supervised [23]	SlowFast-R101	Kinetics-400	✓	8×8	138	53	23.8
CVRL [54]	SlowOnly-R50	Kinetics-400	✗	32×2	42	32	16.3
ρBYOL$_{\rho=3}$ [24]	SlowOnly-R50	Kinetics-400	✗	8×8	42	32	23.4
ρMoCo$_{\rho=3}$ [24]	SlowOnly-R50	Kinetics-400	✗	8×8	42	32	20.3
MaskFeat$_{\uparrow 312}$ [80]	MViT-L	Kinetics-400	✓	40×3	2828	218	37.5
MaskFeat$_{\uparrow 312}$ [80]	MViT-L	Kinetics-600	✓	40×3	2828	218	38.8
VideoMAE	ViT-S	Kinetics-400	✗	16×4	57	22	22.5
VideoMAE	ViT-S	Kinetics-400	✓	16×4	57	22	28.4
VideoMAE	ViT-B	Kinetics-400	✗	16×4	180	87	26.7
VideoMAE	ViT-B	Kinetics-400	✓	16×4	180	87	31.8
VideoMAE	ViT-L	Kinetics-400	✗	16×4	597	305	34.3
VideoMAE	ViT-L	Kinetics-400	✓	16×4	597	305	37.0
VideoMAE	ViT-H	Kinetics-400	✗	16×4	1192	633	**36.5**
VideoMAE	ViT-H	Kinetics-400	✓	16×4	1192	633	**39.5**
VideoMAE	ViT-L	Kinetics-700	✗	16×4	597	305	**36.1**
VideoMAE	ViT-L	Kinetics-700	✓	16×4	597	305	**39.3**

Table 5: **Comparison with the state-of-the-art methods on AVA v2.2.** All models are pre-trained and fine-tuned at image size 224^2. We report the mean Average Precision (mAP) on validation set. "Ex. labels ✗" means only *unlabelled* data is used during the pre-training phase and the pre-trained models are directly transferred to AVA. "Ex. labels ✓" means pre-trained models are additionally fine-tuned on the pre-training dataset with *labels* before transferred to AVA. $T \times \tau$ refers to frame number and corresponding sample rate.

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
TEINet$_{En}$ [40]	ResNet50$_{\times 2}$	ImageNet-1K	✓	8+16	99×10×3	50	66.5	N/A
TANet$_{En}$ [41]	ResNet50$_{\times 2}$		✓	8+16	99×2×3	51	66.0	90.1
TDN$_{En}$ [75]	ResNet101$_{\times 2}$		✓	8+16	198×1×3	88	69.6	92.2
SlowFast [23]	ResNet101	Kinetics-400	✓	8+32	106×1×3	53	63.1	87.6
MViTv1 [22]	MViTv1-B		✓	64	455×1×3	37	67.7	90.9
TimeSformer [6]	ViT-B	ImageNet-21K	✓	8	196×1×3	121	59.5	N/A
TimeSformer [6]	ViT-L		✓	64	5549×1×3	430	62.4	N/A
ViViT FE [3]	ViT-L	IN-21K+K400	✓	32	995×4×3	N/A	65.9	89.9
Motionformer [51]	ViT-B		✓	16	370×1×3	109	66.5	90.1
Motionformer [51]	ViT-L		✓	32	1185×1×3	382	68.1	91.2
Video Swin [39]	Swin-B		✓	32	321×1×3	88	69.6	92.7
VIMPAC [65]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	68.1	N/A
BEVT [77]	Swin-B	IN-1K+K400+DALLE	✗	32	321×1×3	88	70.6	N/A
MaskFeat$_{\uparrow 312}$ [80]	MViT-L	Kinetics-600	✓	40	2828×1×3	218	75.0	95.0
VideoMAE	ViT-B	Kinetics-400	✗	16	180×2×3	87	69.7	92.3
VideoMAE	ViT-L	Kinetics-400	✗	16	597×2×3	305	74.0	94.6
VideoMAE	ViT-S	*no external data*	✗	16	57×2×3	22	66.8	90.3
VideoMAE	ViT-B		✗	16	180×2×3	87	70.8	92.4
VideoMAE	ViT-L		✗	16	597×2×3	305	74.3	94.6
VideoMAE	ViT-L		✗	32	1436×1×3	305	**75.4**	**95.2**

Table 6: **Comparison with the state-of-the-art methods on Something-Something V2**. Our VideoMAE reconstructs normalized cube pixels and is pre-trained with a masking ratio of 90% for 2400 epochs. "Ex. labels ✗" means only *unlabelled* data is used during the pre-training phase. "N/A" indicates the numbers are not available for us.

4.4 Comparison with the state of the art

We compare with the previous state-of-the-art performance on the Kinetics-400 and Something-Something V2 datasets. The results are reported in Table 6 and Table 7. Our VideoMAE can easily scale up with more powerful backbones (e.g. ViT-Large and ViT-Huge) and more frames (e.g. 32). Our VideoMAE achieves the top-1 accuracy of 75.4% on Something-Something V2 and 87.4% on Kinetics-400 without using any extra data. We see that the existing state-of-the-art methods all depend on the external data for pre-training on the Something-Something V2 dataset. On the contrary, our VideoMAE without any external data significantly outperforms previous methods with the same input resolution by around 5%. Our ViT-H VideoMAE also achieves very competitive performance on the Kinetics-400 dataset without using any extra data, which is even better than ViViT-H with on

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
NL I3D [78]	ResNet101	ImageNet-1K	✓	128	359×10×3	62	77.3	93.3
TANet [41]	ResNet152		✓	16	242×4×3	59	79.3	94.1
TDN_{En} [75]	ResNet101		✓	8+16	198×10×3	88	79.4	94.4
TimeSformer [6]	ViT-L	ImageNet-21K	✓	96	8353×1×3	430	80.7	94.7
ViViT FE [3]	ViT-L		✓	128	3980×1×3	N/A	81.7	93.8
Motionformer [51]	ViT-L		✓	32	1185×10×3	382	80.2	94.8
Video Swin [39]	Swin-L		✓	32	604×4×3	197	83.1	95.9
ViViT FE [3]	ViT-L	JFT-300M	✓	128	3980×1×3	N/A	83.5	94.3
ViViT [3]	ViT-H	JFT-300M	✓	32	3981×4×3	N/A	84.9	95.8
VIMPAC [65]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	77.4	N/A
BEVT [77]	Swin-B	IN-1K+DALLE	✗	32	282×4×3	88	80.6	N/A
$MaskFeat_{\uparrow 352}$ [80]	MViT-L	Kinetics-600	✗	40	3790×4×3	218	87.0	97.4
ip-CSN [69]	ResNet152	*no external data*	✗	32	109×10×3	33	77.8	92.8
SlowFast [23]	R101+NL		✗	16+64	234×10×3	60	79.8	93.9
MViTv1 [22]	MViTv1-B		✗	32	170×5×1	37	80.2	94.4
MaskFeat [80]	MViT-L		✗	16	377×10×1	218	84.3	96.3
VideoMAE	ViT-S	*no external data*	✗	16	57×5×3	22	79.0	93.8
VideoMAE	ViT-B		✗	16	180×5×3	87	81.5	95.1
VideoMAE	ViT-L		✗	16	597×5×3	305	85.2	96.8
VideoMAE	ViT-H		✗	16	1192×5×3	633	**86.6**	**97.1**
$VideoMAE_{\uparrow 320}$	ViT-L	*no external data*	✗	32	3958×4×3	305	86.1	97.3
$VideoMAE_{\uparrow 320}$	ViT-H		✗	32	7397×4×3	633	**87.4**	**97.6**

Table 7: **Comparison with the state-of-the-art methods on Kinetics-400**. Our VideoMAE reconstructs normalized cube pixels. Here models are self-supervised pre-trained with a masking ratio of 90% for 1600 epochs on Kinetics-400. $VideoMAE_{\uparrow 320}$ is initialized from its 224^2 resolution counterpart and then fine-tuned for evaluation. "Ex. labels ✗" means only *unlabelled* data is used during the pre-training phase. "N/A" indicates the numbers are not available for us.

JFT-300M pre-training (86.6% v.s. 84.9%). When fine-tuned with larger spatial resolutions and input video frames, the performance of our ViT-H VideoMAE can further boost from 86.6% to 87.4%.

5 Conclusion

In this paper, we have presented a simple and data-efficient self-supervised learning method (VideoMAE) for video transformer pre-training. Our VideoMAE introduces two critical designs of extremely high masking ratio and tube masking strategy to make the video reconstruction task more challenging. This harder task would encourage VideoMAE to learn more representative features and relieve the information leakage issue. Empirical results demonstrate this simple algorithm works well for video datasets of different scales. In particular, we are able to learn effective VideoMAE only with thousands of video clips, which has significant practical value for scenarios with limited data available.

Future work VideoMAE could be further improved by using larger webly datasets, larger models (e.g., ViT-G) and larger spatial resolutions of input video (e.g., 384^2). VideoMAE only leverages the RGB video stream without using additional audio or text stream. We expect that audio and text from the video data can provide more information for self-supervised pre-training.

Broader impact Potential negative societal impacts of VideoMAE are mainly concerned with energy consumption. The pre-training phase may lead to a large amount of carbon emission. Though the pre-training is energy-consuming, we only need to pre-train the model once. Different downstream tasks can then share the same pre-trained model via additional fine-tuning. Our VideoMAE unleashes the great potential of vanilla vision transformer for video analysis, which could increase the risk of video understanding model or its outputs being used incorrectly, such as for unauthorized surveillance.

Acknowledgements and disclosure of funding Thanks to Ziteng Gao, Lei Chen and Chongjian Ge for their help. This work is supported by National Natural Science Foundation of China (No. 62076119, No. 61921006), the Fundamental Research Funds for the Central Universities (No. 020214380091), Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202125), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *Advances in Neural Information Processing Systems*, 2020.

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems*, 2020.

[3] 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.

[4] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*, 2022.

[5] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

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

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 2020.

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, 2020.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[10] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[11] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, 2020.

[12] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning*, 2020.

[13] Peihao Chen, Deng Huang, Dongliang He, Xiang Long, Runhao Zeng, Shilei Wen, Mingkui Tan, and Chuang Gan. Rspnet: Relative speed perception for unsupervised video representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[14] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *CVPR*, 2021.

[15] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[16] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[17] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*, 2019.

[19] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[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 preprint arXiv:2111.12710*, 2021.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[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.

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

[24] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[25] Chongjian Ge, Youwei Liang, Yibing Song, Jianbo Jiao, Jue Wang, and Ping Luo. Revitalizing cnn attentions via transformers in self-supervised visual representation learning. In *Advances in Neural Information Processing Systems*, 2021.

[26] 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/CVF International Conference on Computer Vision*, 2017.

[27] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[28] Sheng Guo, Zihua Xiong, Yujie Zhong, Limin Wang, Xiaobo Guo, Bing Han, and Weilin Huang. Cross-architecture self-supervised video representation learning. In *CVPR*, 2022.

[29] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European Conference on Computer Vision*, 2020.

[30] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems*, 2020.

[31] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[32] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[33] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[34] 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. *arXiv preprint arXiv:1705.06950*, 2017.

[35] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *IEEE/CVF International Conference on Computer Vision*, 2011.

[36] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequence. In *IEEE/CVF International Conference on Computer Vision*, 2017.

[37] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *CoRR*, abs/2001.05691, 2020.

[38] 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 *IEEE/CVF International Conference on Computer Vision*, 2021.

[39] 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.

[40] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. TEINet: Towards an efficient architecture for video recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[41] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[42] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[43] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[44] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[45] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, 2016.

[46] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2019.

[48] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[49] Viorica Patraucean, Ankur Handa, and Roberto Cipolla. Spatio-temporal video autoencoder with differentiable memory. *arXiv preprint arXiv:1511.06309*, 2015.

[50] Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[51] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joo F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems*, 2021.

[52] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[53] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[54] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge J. Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[55] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

[56] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019.

[57] 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.

[58] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

[59] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 2014.

[60] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[61] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[62] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using lstms. *International Conference on Machine Learning*, 2015.

[63] Jonathan C. Stroud, David A. Ross, Chen Sun, Jia Deng, Rahul Sukthankar, and Cordelia Schmid. Learning video representations from textual web supervision. *CoRR*, abs/2007.14937, 2020.

[64] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

[65] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. Vimpac: Video pre-training via masked token prediction and contrastive learning. *arXiv preprint arXiv:2106.11250*, 2021.

[66] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision*, 2020.

[67] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021.

[68] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015.

[69] Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video classification with channel-separated convolutional networks. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[70] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[72] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning*, 2008.

[73] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 2010.

[74] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European Conference on Computer Vision*, 2020.

[75] 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.

[76] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[77] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[78] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[79] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *IEEE/CVF International Conference on Computer Vision*, 2015.

[80] 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.

[81] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Advances in Neural Information Processing Systems*, 2021.

[82] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[83] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[84] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using VQ-VAE and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[85] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. *arXiv preprint arXiv:2006.15489*, 2020.

[86] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *IEEE/CVF International Conference on Computer Vision*, 2019.

[87] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[88] Zhang Zhang and Dacheng Tao. Slow feature analysis for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[89] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

[90] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image bert pre-training with online tokenizer. *International Conference on Learning Representations*, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Shown in Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Shown in Section 5.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Shown in Section 4.1 and Appendix § B.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Shown in Appendix § B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Because of the computation costs, we did not run the experiments multiple times.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please see Table 3 and Appendix § B.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] Shown in Appendix § F.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We used publicly available datasets whose licenses allow research usage.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] To the best of our knowledge, the data we used contains no personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Appendix

In this appendix, we provide more details of VideoMAE from the following aspects:

- The detailed architecture illustration is in § A.
- The implementation details are in § B.
- Experimental results are in § C where there are ablation studies on the Something-Something V2 and Kinectics-400 datasets, and a downstream evaluation task (i.e., action detection). More comparisons with state-of-the-art methods on the UCF101 and HMDB51 datasets are included as well.
- Results analysis is in § D.
- Visualization of reconstructed samples is in § E.
- License of the datasets is in § F.

A Architectures

We use an asymmetric encoder-decoder architecture for video self-supervised pre-training and discard the decoder during the fine-tuning phase. We take the 16-frame vanilla ViT-Base for example, and the specific architectural design for the encoder and decoder is shown in Table 8. We adopt the joint space-time attention [3, 39] to better capture the high-level spatio-temporal information in the remaining tokens.

Stage	Vision Transformer (Base)	Output Sizes
data	stride $4\times1\times1$ on *K400* stride $2\times1\times1$ on *SSV2*	$3\times16\times224\times224$
cube	$2\times16\times16$, 768 stride $2\times16\times16$	$768\times\mathbf{8}\times196$
mask	tube mask *mask ratio* = ρ	$768\times\mathbf{8}\times[196\times(1\text{-}\rho)]$
encoder	$\begin{bmatrix} MHA(768) \\ MLP(3072) \end{bmatrix}\times12$	$768\times\mathbf{8}\times[196\times(1\text{-}\rho)]$
projector	MLP(384) & *concat learnable tokens*	$384\times\mathbf{8}\times196$
decoder	$\begin{bmatrix} MHA(384) \\ MLP(1536) \end{bmatrix}\times4$	$384\times\mathbf{8}\times196$
projector	MLP(1536)	$1536\times\mathbf{8}\times196$
reshape	*from* 1536 *to* $3\times2\times16\times16$	$3\times16\times224\times224$

Table 8: **Architectures details of VideoMAE.** We take 16-frame vanilla ViT-Base for example. "MHA" here denotes the joint space-time self-attention. The output sizes are denoted by $\{C\times T\times S\}$ for channel, temporal and spatial sizes.

B Implementation Details

We conduct the experiments with 64 GPUs for both pre-training and fine-tuning on the Something-Something V2 and Kinetics-400 datasets. The experiments on the smaller UCF101 and HMDB51 datasets are trained with 8 GPUs. The experiments on the AVA dataset are conducted with 32 GPUs. We linearly scale the base learning rate w.r.t. the overall batch size, *lr* = *base learning rate* $\times$ *batch size / 256*. We adopt the PyTorch [47] and DeepSpeed[2] frameworks for faster training. We have made the code[3] and pre-trained models[4] public to facilitate future research in self-supervised video pre-training.

Something-Something V2. Our VideoMAE is pre-trained for 800 epochs on Something-Something V2 by default. During the fine-tuning phase, we perform the uniform sampling following TSN [76]. For evaluation, all models share the same inference protocol, i.e., 2 clips $\times$ 3 crops. The default

[2] https://github.com/microsoft/DeepSpeed
[3] https://github.com/MCG-NJU/VideoMAE
[4] https://github.com/MCG-NJU/VideoMAE/blob/main/MODEL_ZOO.md

config	Sth-Sth V2	Kinetics-400
optimizer	AdamW	
base learning rate	1.5e-4	
weight decay	0.05	
optimizer momentum	β_1, β_2=0.9, 0.95 [12]	
batch size	1024	
learning rate schedule	cosine decay [42]	
warmup epochs	40	
flip augmentation	*no*	*yes*
augmentation	MultiScaleCrop [76]	

Table 9: **Pre-training setting.**

config	Sth-Sth V2	Kinetics-400
optimizer	AdamW	
base learning rate	1e-3(S), 5e-4(B,L)	1e-3
weight decay	0.05	
optimizer momentum	β_1, β_2=0.9, 0.999	
batch size	512	512
learning rate schedule	cosine decay [42]	
warmup epochs	5	
training epochs	40 (S,B), 30 (L)	150 (S), 75 (B), 50 (L,H)
repeated augmentation	2	2
flip augmentation	*no*	*yes*
RandAug [16]	(9, 0.5)	(9, 0.5)
label smoothing [64]	0.1	0.1
mixup [87]	0.8	0.8
cutmix [86]	1.0	1.0
drop path	0.1 (S,B), 0.2 (L,H)	
dropout	0.5 (L)	0.5 (L,H)
layer-wise lr decay [4]	0.7 (S),0.75 (B,L)	0.7 (S),0.75 (B,L,H)

Table 10: **End-to-end fine-tuning setting in Table 6 and Table 7.**

config	Sth-Sth V2
optimizer	SGD
base learning rate	0.1
weight decay	0
optimizer momentum	0.9
batch size	1024
learning rate schedule	cosine decay
warmup epochs	10
training epochs	100
augmentation	MultiScaleCrop

Table 11: **Linear probing setting.**

settings of pre-training, fine-tuning, and linear probing are shown in Table 9, Table 12, and Table 11. For supervised training, we follow the recipe in [22] and train from scratch for 100 epochs. Note that we use *no flip augmentation* during both the pre-training and fine-tuning phase. We additionally adopt the repeated augmentation [32] during the fine-tuning phase in Table 6, which can further increase the Top-1 accuracy by 0.1% - 0.3%.

Kinetics-400. Our VideoMAE is pre-trained for 800 epochs on Kinetics-400 by default. During the fine-tuning phase, we perform the dense sampling following Slowfast [23]. For evaluation, all models share the same inference protocol, i.e., 5 clips $\times$ 3 crops. The default settings of pre-training and fine-tuning are shown in Table 9 and Table 12. For supervised training from scratch, we follow the recipe in [22] and train the model for 200 epochs. Note that we adopt the repeated augmentation [32] during the fine-tuning phase in Table 7, which can further increase the Top-1 accuracy by 0.8% - 1.0%.

config	AVA v2.2	
additional fine-tuning on Kinetics	✗	✓
optimizer	AdamW	
base learning rate	1e-3 (S), 2.5e-4 (B, L), 5e-4 (H)	5e-4
weight decay	0.05	
optimizer momentum	β_1, β_2=0.9, 0.999	
batch size	128	
learning rate schedule	cosine decay [42]	
warmup epochs	5	
training epochs	30 (S, B, L), 20 (H)	30 (S, B) 20 (L, H)
repeated augmentation	*no*	
flip augmentation	*yes*	
drop path	0.2	
layer-wise lr decay [4]	0.6 (S), 0.75 (B, L), 0.8 (H)	0.6 (S), 0.75 (B), 0.8 (L, H)

Table 12: **End-to-end fine-tuning setting in Table 5.**

UCF101. We follow a similar recipe on Kinetics for pre-training. Our VideoMAE is pre-trained with a masking ratio of 75% for 3200 epochs. The batch size and base learning rate are set to 192 and 3e-4, respectively. Here, 16 frames with a temporal stride of 4 are sampled. For fine-tuning, the model is trained with repeated augmentation [32] and a batch size of 128 for 100 epochs. The base learning rate, layer decay and drop path are set to 5e-4, 0.7 and 0.2, respectively. For evaluation, we adopt the inference protocol of 5 clips $\times$ 3 crops.

HMDB51. Our VideoMAE is pre-trained with a masking ratio of 75% for 4800 epochs. The batch size and base learning rate are set to 192 and 3e-4, respectively. Here, 16 frames with a temporal stride of 2 are sampled. For fine-tuning, the model is trained with repeated augmentation [32] and a batch size of 128 for 50 epochs. The base learning rate, layer decay and drop path are set to 1e-3, 0.7 and 0.2, respectively. For evaluation, we adopt the inference protocol of 10 clips $\times$ 3 crops.

AVA. We follow the action detection architecture in Slowfast [23] and use the detected person boxes from AIA [66]. The default settings of fine-tuning are shown in Table 12. For data augmentations, we resize the short side of the input frames to 256 pixels. We apply a random crop of the input frames to 224$\times$224 pixels and random flip during training. We use only ground-truth person boxes for training and the detected boxes with confidence $\geq$0.8 for inference.

C Additional Results

C.1 Training schedule

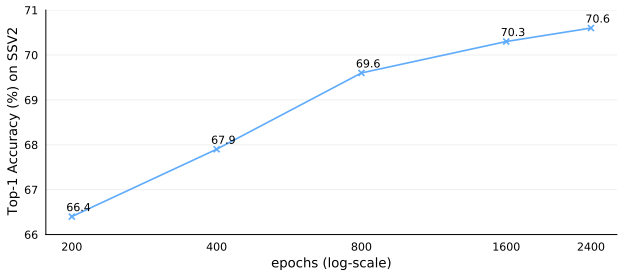
Figure 5 shows the influence of the longer pre-training schedule on the Something-Something V2 and Kinetics-400 datasets. We find that a longer pre-training schedule brings slight gains to both datasets. In the main paper, our VideoMAE is pre-trained for 800 epochs by default.

C.2 Comparison with the state-of-the-art methods

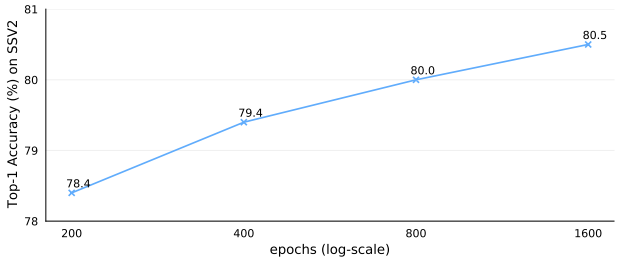
We present the detailed comparison with the state-of-the-art on UCF101 and HMDB51 in Table 13. Figure 6 additionally shows that our VideoMAE is a data-efficient learner that allows us to effectively train video transformers only from limited video data (e.g., 9.5k clips in UCF101, and 3.5k clips in HMDB51) without any ImageNet pre-training. VideoMAE significantly outperforms training from scratch, MoCo v3 pre-training [15], and the previous best performance from Vi^{2}CLR [19] without extra data on these small-scale video datasets. Compared with those large-scale video datasets, these two small datasets are more proper to verify the effectiveness of VideoMAE, as training large ViT models is more challenging on small datasets.

D Model result analysis

In this section, we add the analysis of model results. As shown in Figure 7 and Figure 8, our VideoMAE bring significant gain for most categories on SSV2, which implies that our VideoMAE



(a) Performance on Something-Something V2



(b) Performance on Kinetics-400

Figure 5: The effect of **training schedules** on (a) Something-Something V2 and (b) Kinetics-400. Here each point is a full training schedule. Our default ViT-B backbone is described in Table 8.

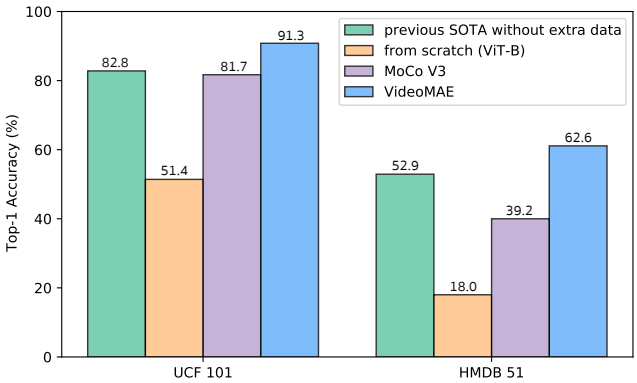


Figure 6: Comparison with VideoMAE, MoCo v3 [15], and Vi^{2}CLR [19] on UCF101 and HMDB51.

can capture more spatiotemporal structure representations than ImageMAE and ImageNet-21k supervised pre-trained model. On the other hand, we also notice that our VideoMAE performs slightly worse than other two models on some categories. To better understand how the model works, we select several examples from validation set. The examples are shown in Figure 9. For the example in the 1st row, we find our VideoMAE might not capture the motion information from very small object. We suspect that tokens containing the small motion might all be masked due to our extremely high masking ratio, so our VideoMAE could hardly reconstruct the masked small motion pattern. For the example in the 2nd row, we find our VideoMAE could capture the deformation of objects and movement from the squeeze of the hand, while this cannot be discriminated by image pre-training. We leave more detailed analysis of our VideoMAE for future work.

E Visualization

We show several examples of reconstruction in Figure 10 and Figure 11. Videos are all randomly chosen from the validation set. We can see that even under an extremely high masking ratio, VideoMAE can produce satisfying reconstructed results. These examples imply that our VideoMAE

Method	Backbone	Extra data	Frames	Param	Modality	UCF101	HMDB51
OPN [36]	VGG	UCF101	N/A	N/A	V	59.6	23.8
VCOP [83]	R(2+1)D	UCF101	N/A	N/A	V	72.4	30.9
CoCLR [30]	S3D-G	UCF101	32	9M	V	81.4	52.1
Vi^{2}CLR [19]	S3D	UCF101	32	9M	V	82.8	52.9
VideoMAE	ViT-B	*no external data*	16	87M	V	**91.3**	**62.6**
SpeedNet [5]	S3D-G	Kinetics-400	64	9M	V	81.1	48.8
VTHCL [85]	SlowOnly-R50	Kinetics-400	8	32M	V	82.1	49.2
Pace [74]	R(2+1)D	Kinetics-400	16	15M	V	77.1	36.6
MemDPC [29]	R-2D3D	Kinetics-400	40	32M	V	86.1	54.5
CoCLR [30]	S3D-G	Kinetics-400	32	9M	V	87.9	54.6
RSPNet [13]	S3D-G	Kinetics-400	64	9M	V	93.7	64.7
VideoMoCo [46]	R(2+1)D	Kinetics-400	16	15M	V	78.7	49.2
Vi^{2}CLR [19]	S3D	Kinetics-400	32	9M	V	89.1	55.7
CVRL [54]	SlowOnly-R50	Kinetics-400	32	32M	V	92.9	67.9
CVRL [54]	SlowOnly-R50	Kinetics-600	32	32M	V	93.6	69.4
CVRL [54]	Slow-R152 (2$\times$)	Kinetics-600	32	328M	V	94.4	70.6
CORP$_f$ [33]	SlowOnly-R50	Kinetics-400	32	32M	V	93.5	68.0
ρSimCLR$_{\rho=2}$ [24]	SlowOnly-R50	Kinetics-400	8	32M	V	88.9	N/A
ρSwAV$_{\rho=2}$ [24]	SlowOnly-R50	Kinetics-400	8	32M	V	87.3	N/A
ρMoCo$_{\rho=2}$ [24]	SlowOnly-R50	Kinetics-400	8	32M	V	91.0	N/A
ρBYOL$_{\rho=2}$ [24]	SlowOnly-R50	Kinetics-400	8	32M	V	92.7	N/A
ρBYOL$_{\rho=4}$ [24]	SlowOnly-R50	Kinetics-400	8	32M	V	94.2	72.1
MIL-NCE [44]	S3D	HowTo100M	32	9M	V+T	91.3	61.0
MMV [1]	S3D-G	AS+HTM	32	9M	V+A+T	92.5	69.6
CPD [37]	ResNet50	IG300k	16	N/A	V+T	92.8	63.8
ELO [52]	R(2+1)D	Youtube8M-2	N/A	N/A	V+A	93.8	67.4
XDC [2]	R(2+1)D	Kinetics-400	32	15M	V+A	84.2	47.1
XDC [2]	R(2+1)D	IG65M	32	15M	V+A	94.2	67.1
GDT [50]	R(2+1)D	Kinetics-400	32	15M	V+A	89.3	60.0
GDT [50]	R(2+1)D	IG65M	32	15M	V+A	95.2	72.8
VideoMAE	ViT-B	Kinetics-400	16	87M	V	**96.1**	**73.3**

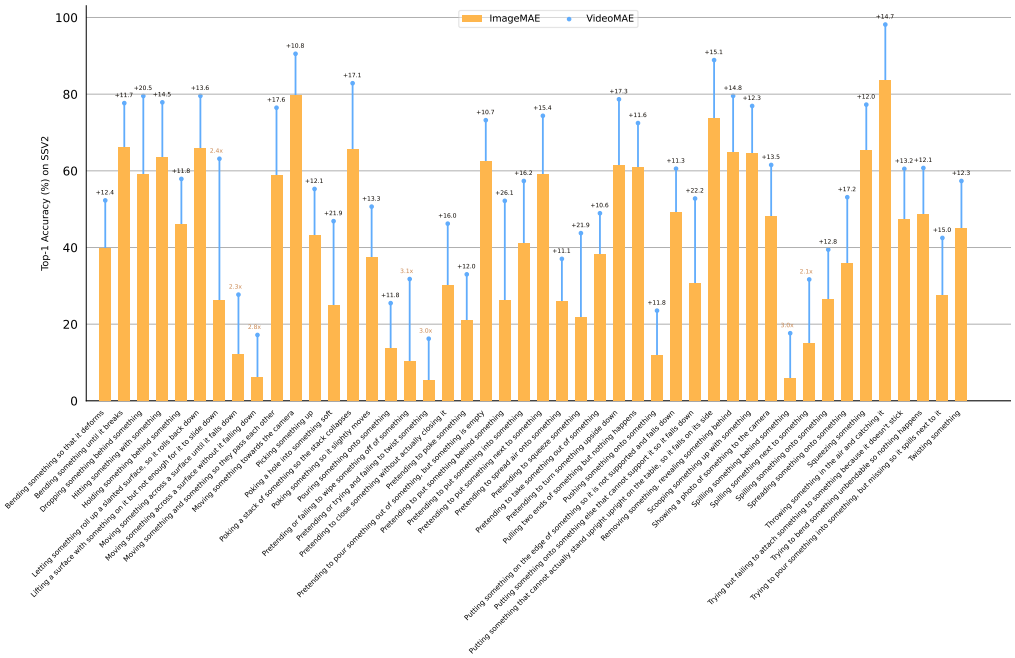
Table 13: **Comparison with the state-of-the-art methods on UCF101 and HMDB51.** Our VideoMAE reconstructs normalized cube pixels and is pre-trained with a masking ratio of 75% for 3200 epochs on UCF101 and 4800 epochs on HMDB51, respectively. We report fine-tuning accuracy for evaluation. 'V' refers to visual only, 'A' is audio, 'T' is text narration. "N/A" indicates the numbers are not available for us.

is able to learn more representative features that capture the holistic spatiotemporal structure in videos.

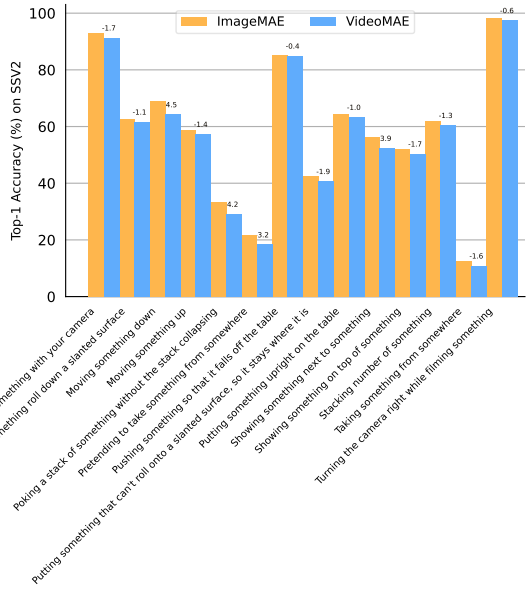
F License of Data

All the datasets we used are commonly used datasets for academic purpose. The license of the Something-Something V2[5] and UCF101[6] datasets is custom. The license of the Kinetics-400[7], HMDB51[8] and AVA[9] datasets is CC BY-NC 4.0[10].

[5]URL: https://developer.qualcomm.com/software/ai-datasets/something-something
[6]URL: https://www.crcv.ucf.edu/data/UCF101.php
[7]URL: https://www.deepmind.com/open-source/kinetics
[8]URL: https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database
[9]URL: https://research.google.com/ava/index.html
[10]URL: https://creativecommons.org/licenses/by/4.0

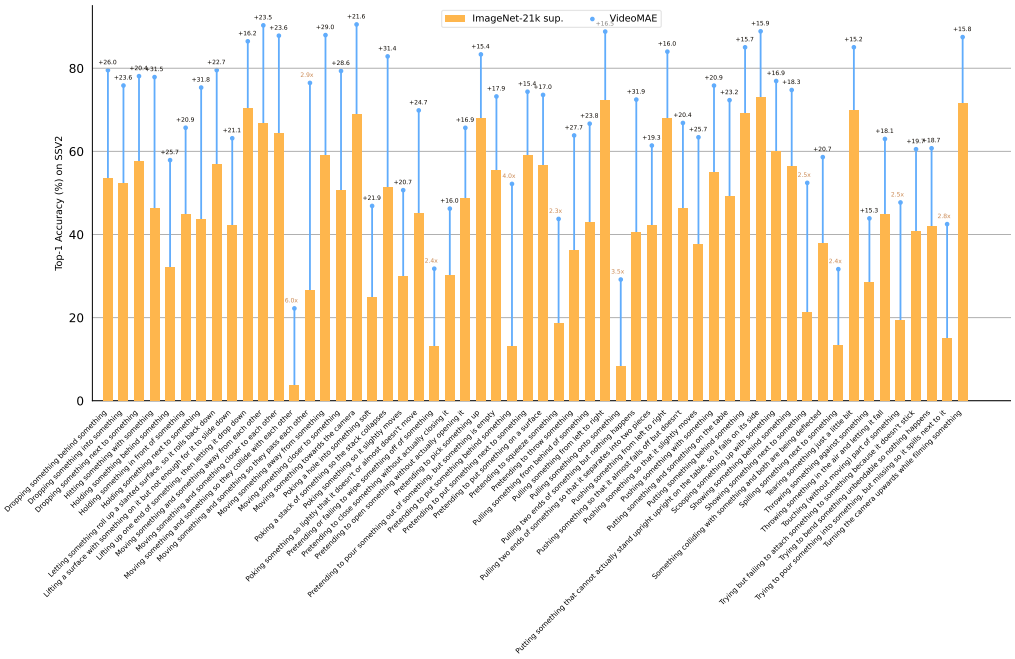


(a) Categories that VideoMAE outperforms ImageMAE. We only show those gain larger than 10%.

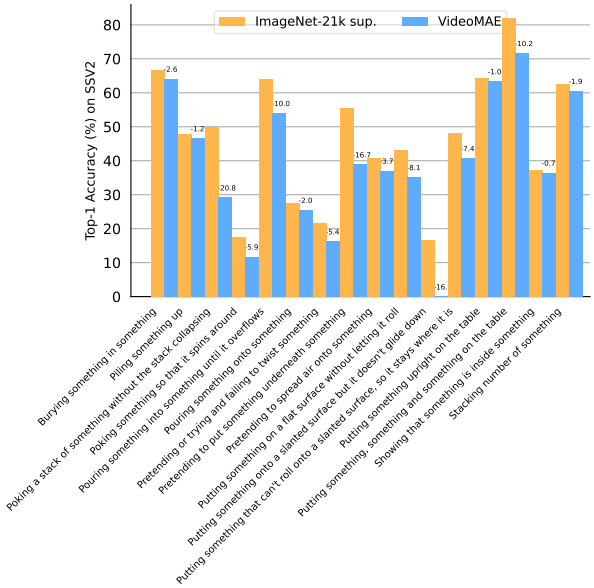


(b) Categories that ImageMAE outperforms VideoMAE.

Figure 7: ImageMAE (64.8%) vs. VideoMAE (69.6%) on Something-Something V2.



(a) Categories that VideoMAE outperforms ImageNet-21k supervised pre-trained model. We only show those gain larger than 15%.



(b) Categories that ImageNet-21k supervised pre-trained model outperforms VideoMAE.

Figure 8: ImageNet-21k supervised pre-trained model (61.8%) vs. VideoMAE (69.6%) on Something-Something V2.

Figure 9: Prediction examples of different models on Something-Something V2. For each example drawn from the validation dataset, the predictions with blue text indicating a correct prediction and red indicating an incorrect one. "GT" indicates the ground truth of the example.

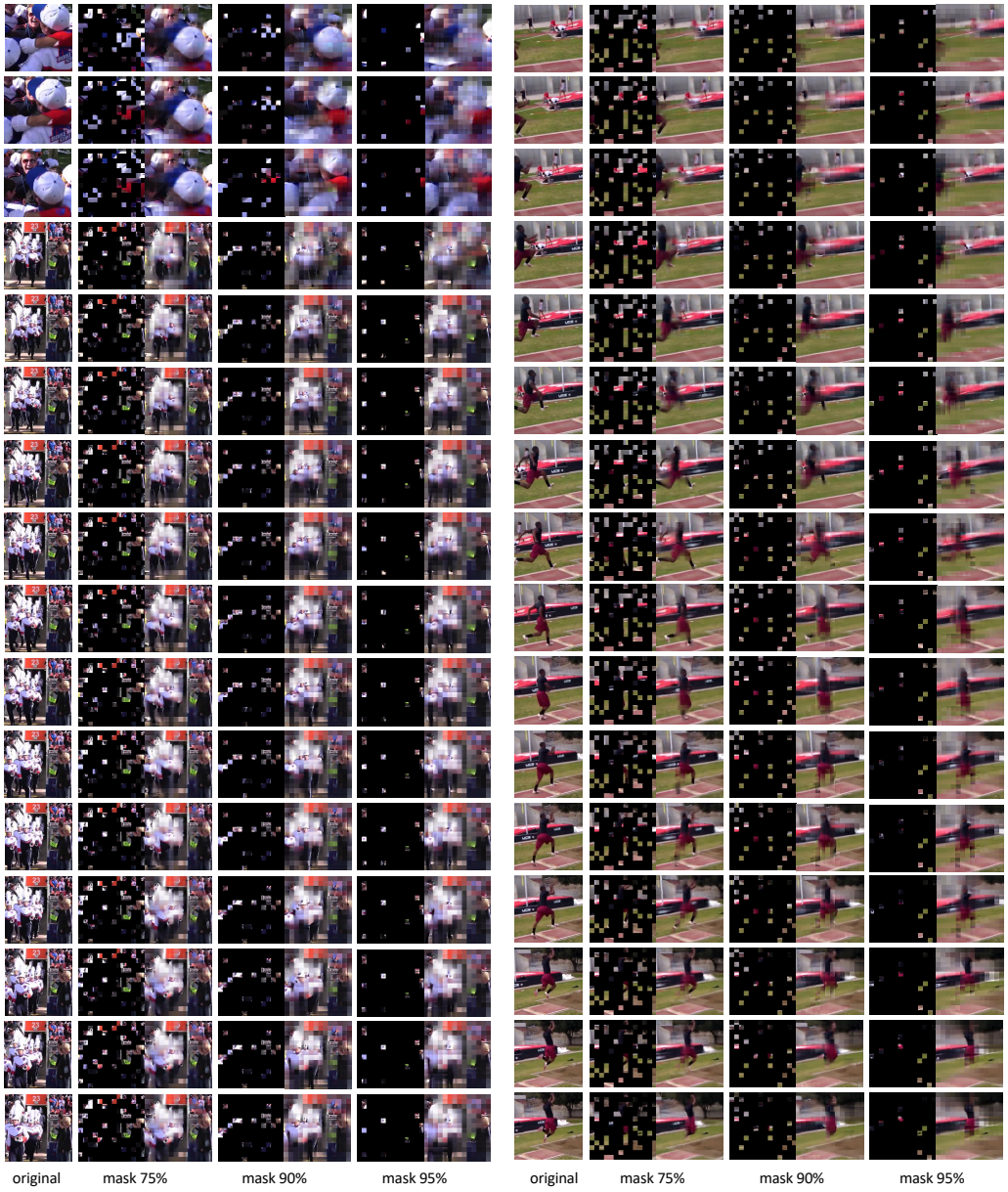


Figure 10: **Uncurated random videos** on Kinetics-400 *validation* set. We show the original video squence and reconstructions with different masking ratios. Reconstructions of videos are predicted by our VideoMAE pre-trained with a masking ratio of 90%.

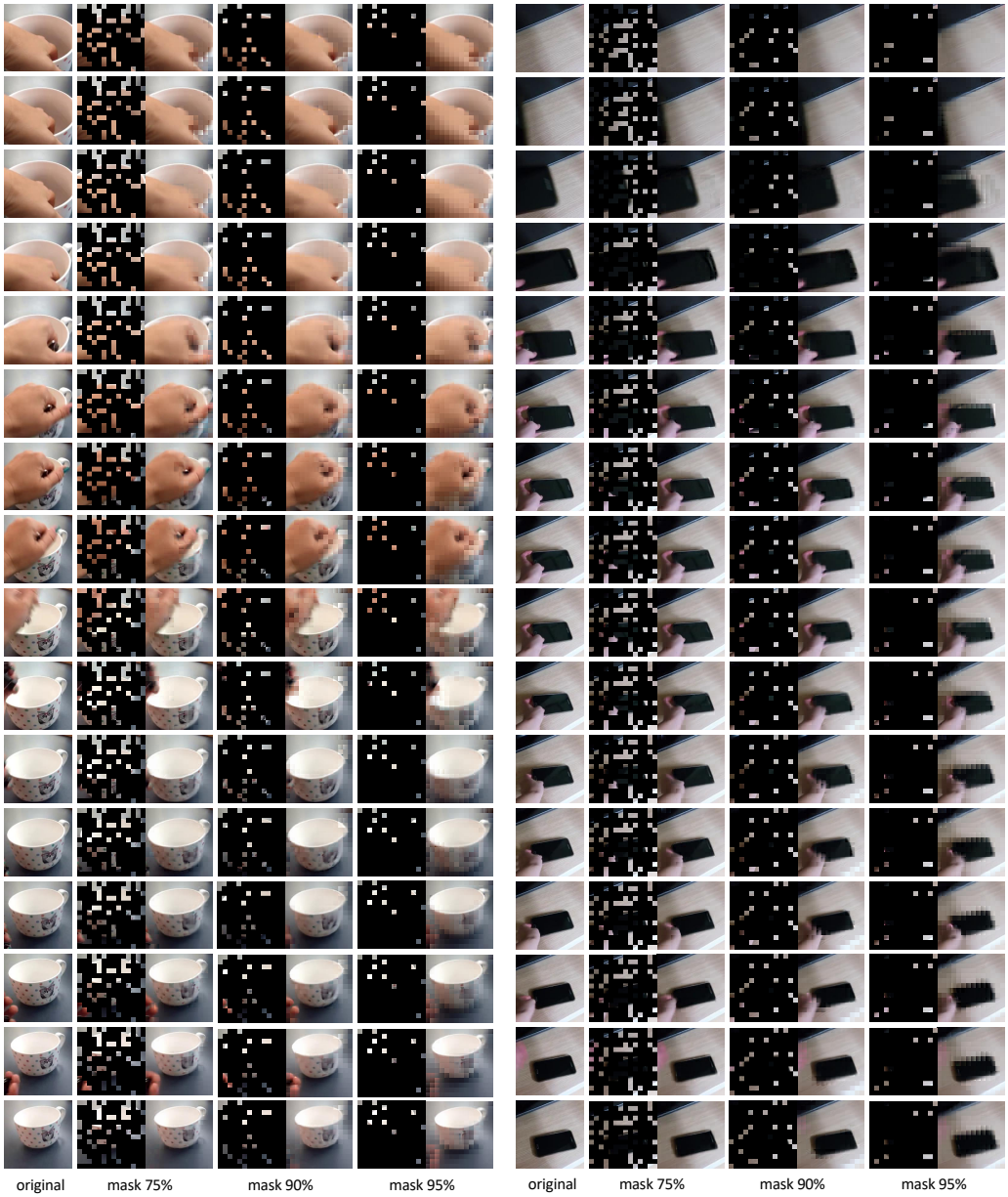


Figure 11: **Uncurated random videos** on Something-Something V2 *validation* set. We show the original video squence and reconstructions with different masking ratios. Reconstructions of videos are all predicted by our VideoMAE pre-trained with a masking ratio of 90%.