Mask Again: Masked Knowledge Distillation for Masked Video Modeling

Xiaojie Li
Harbin Institute of Technology, Shenzhen
Peng Cheng Laboratory
xiaojieli0903@gmail.com

Shaowei He
Harbin Institute of Technology, Shenzhen
shaowei.hsw@gmail.com

Jianlong Wu *
Harbin Institute of Technology, Shenzhen
wujianlong@hit.edu.cn

Yue Yu
Peng Cheng Laboratory
yuy@pcl.ac.cn

Liqiang Nie *
Harbin Institute of Technology, Shenzhen
nieliqiang@gmail.com

Min Zhang
Harbin Institute of Technology, Shenzhen
zhangmin2021@hit.edu.cn

ABSTRACT

Masked video modeling has shown remarkable performance in downstream tasks by predicting masked video tokens from visible ones. However, training models from scratch on large-scale unlabeled data remains computationally challenging and time-consuming. Moreover, the commonly used random-based sampling techniques may lead to the selection of redundant or low-information regions, hindering the model from learning discriminative representations within the limited training epochs. To achieve efficient pre-training, we propose MaskAgain, an efficient feature-based knowledge distillation framework for masked video pre-training that facilitates knowledge transfer from a pre-trained teacher model to a student model. In contrast to previous approaches that align all visible token features with the teacher model at output layers, MaskAgain adopts a selective approach by masking visible tokens again at both the hidden and output layers of the transformer block. Attention mechanisms are utilized for informative feature selection. Extensive experimental results show that MaskAgain achieves comparable or even better performance than existing methods on benchmark datasets with much fewer training epochs and much less memory, which demonstrates that MaskAgain allows for efficient pre-training of accurate video models, reducing computational resources and training time significantly. Code is released at https://github.com/xiaojieli0903/MaskAgain.

CCS CONCEPTS

• **Computing methodologies** → **Activity recognition and understanding**.

KEYWORDS

Video Representation Learning; Masked Visual Modeling; Knowledge Distillation

ACM Reference Format:
Xiaojie Li, Shaowei He, Jianlong Wu, Yue Yu, Liqiang Nie, and Min Zhang. 2023. Mask Again: Masked Knowledge Distillation for Masked Video Modeling. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3581783.3612129

1 INTRODUCTION

Vision Transformers (ViTs) [10, 34, 35] achieve remarkable performance in computer vision tasks, especially on large-scale datasets with self-supervised learning. Masked visual modeling trains transformers for robust visual representations by predicting masked content. MAE [21] shows promise by randomly masking patches of input images and reconstructing missing pixels with a lightweight decoder. Similar ideas have extended to spatiotemporal representation learning for videos, resulting in VideoMAE [56] and ST-MAE [13], which utilize high masking ratios for memory-efficient training. Efficient masked visual modeling paradigms based on knowledge distillation have emerged to address the challenges of efficiency and scalability in training large-scale models on extensive unlabeled data. DMAE [2] aligns intermediate features and optimizes pixel reconstruction loss on masked inputs. MaskAlign [65] removes the pixel reconstruction module and aligns visible tokens with the teacher's features. In the video domain, MVD [60] additionally leverages pre-trained MIM models for masked feature prediction, effectively enhancing video model performance.

Previous studies have commonly relied on random-based sampling techniques, such as *random*[13], *tube*[56], and *frame* [62] masking, which assume a uniform probability distribution for selecting visible patches. However, these random-based strategies may lead to the selection of redundant or low-information regions. As depicted in Figure 1 (a) and Figure 1 (b), through at high mask rates, the subject of the action occupies only a small portion of the visible patches, while background tokens occupy a larger proportion. To enable efficient pre-training, we present **MaskAgain**, an efficient masked knowledge distillation framework for masked video modeling. MaskAgain employs a selective approach that helps to focus on informative regions, enabling the student model to learn

*Corresponding authors.

MM '23, October 29– November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612129

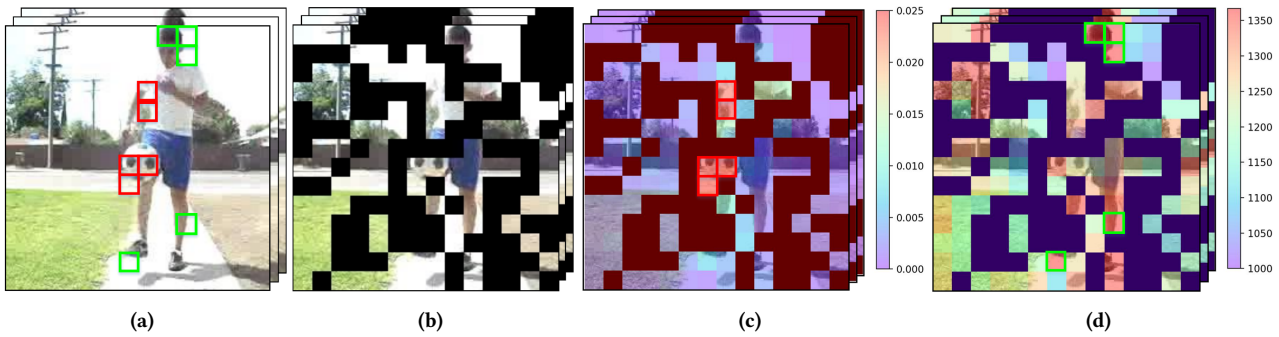


Figure 1: Attention Mechanisms in MaskAgain: (a) Original frames. (b) Masked frames using random-based masking. (c) Hidden level attention maps with red boxes denoting selected tokens with Top-5 attention values, showing a higher focus on tokens belonging to the same object, e.g., soccer. (d) Output level attention maps with green boxes denoting selected tokens with Top-5 attention values, highlighting tokens mostly concentrated on the action subject or distinctive elements, like the bent leg.

semantically rich knowledge from the teacher model. Moreover, MaskAgain conducts knowledge distillation at both the hidden and output layers of the transformer block, guided by attention mechanisms. This allows for learning discriminative feature representations efficiently which mitigates the reliance on long pre-training epochs.

At the hidden layer, MaskAgain utilizes attention maps from the transformer's multi-head attention structure to select informative token features at both temporally-global and temporally-local levels. These selected features are aggregated using a weighted summation function for generating a compact representation focused on crucial information for feature transfer. Two types of knowledge, temporally-global knowledge (**MAGlobal**) and temporally-local knowledge (**MALocal**), are defined to capture temporal changes throughout the entire video and transfer fine-grained frame-level knowledge. At the output layer, token-level activation-based attention maps are generated using visible tokens from the teacher network's encoder. Two knowledge transfer methods, feature hint alignment (**MAHint**) and relationship consistency transfer (**MARel**), are explored. MAHint aligns selected token features between the teacher and student, enabling the student to learn from the teacher's semantic features. MARel transfers valuable token-level relationship knowledge within a mini-batch, enhancing the student network's performance while reducing computational costs. The effectiveness of attention mechanisms is demonstrated through Figure 1 (c) and Figure 1 (d), showing the selective focus on informative tokens and the significance of token-level relationships in facilitating efficient knowledge transfer.

We demonstrate the efficacy of the MaskAgain framework on Kinetics-400, UCF101, and HMDB51 datasets, achieving impressive accuracies on various datasets comparable to or even surpassing models trained for longer epochs. Applying it to the ViT-B/16 model with just 400 epochs of distillation results in impressive Top-1 accuracies of 81.0% on Kinetics-400, 70.3 % on Something-Something-v2, 96.3% on UCF101, and 77.1% on HMDB51. These achievements are comparable to, or even surpass, the performance of the ViT-B/16 teacher model trained for 1600 epochs. MaskAgain enables efficient pre-training of highly accurate video models, significantly reducing computational resources and training time.

2 RELATED WORK

2.1 Masked Video Modeling

Video representation learning has seen significant progress in capturing spatial-temporal representations for video understanding tasks. Various learning approaches have been explored, including supervised methods [1, 11, 14, 32, 36, 58], semi-supervised approaches [50, 63] and self-supervised learning (SSL) methods [9, 13, 16, 19, 40, 56]. Among SSL methods, masked visual modeling has shown great promise due to its ability to learn discriminative video features without relying on high-quality labels. Various works have been proposed in this area, including MAE [21], BEiT [3], and MaskFeat [62]. MAE employs an end-to-end masked autoencoder to predict original pixels from masked image patches, reducing computational costs significantly. BEiT reconstructs visual tokens extracted by the tokenizer of DALL-E [47] instead of predicting pixels. MaskFeat directly predicts HOG features of masked contents for effective visual representation learning. In the video domain, VideoMAE [56] and ST-MAE [13] extend the MAE approach to videos and achieve impressive results using the tube or random masking strategies. These techniques leverage higher masking ratios, around 90%, compared to techniques used in the image domain, which typically employ around 60% masking [64] or 75% masking [21]. By effectively capturing spatial and temporal dynamics in unsupervised video streams, they have shown promising performance in video representation learning. Our proposed method builds upon the tube mask strategy utilized in VideoMAE, focusing on important feature learning through the selection of attention mechanisms and discovering that the attention mechanisms in different layers of the transformer block possess distinct properties, making them complementary to each other.

2.2 Knowledge Distillation

Knowledge distillation is a widely used model compression technique that transfers knowledge from a large teacher model to a smaller student model. It can be categorized into three main directions: response-based [22, 48, 67], feature-based [2, 27, 49, 52, 57, 60, 61], and relation-based methods [31, 33, 44, 55].

Feature-based distillation has received considerable attention, especially in masked visual modeling where labeled data is limited.

DMAE [2] demonstrates impressive results on ImageNet by incorporating feature distillation into the MAE structure. MaskAlign [65] introduces an asymmetric mask mechanism to align student and teacher features. G2SD [26] transfers task-agnostic and task-specific knowledge by aligning student predictions with hidden teacher features at visible and masked patches. In the video domain, MVD [60] leverages both image and video teachers to guide student feature reconstruction. Attention maps within transformer blocks have been identified as valuable intermediate features for knowledge distillation, and prior approaches like MobileBERT [52], TinyBERT [27], and MINILM [61] explore various methods to transfer attention maps.

Relation-based distillation complements feature distillation by preserving consistent relationships between instances in teacher and student models. CCKD [44] and LKD [31] transfer correlations among global or local views of the instances in the embedded space. Additionally, RKD [42] leverages distance-wise and angle-wise distillation losses that penalize structural differences in relations. In this work, we introduce a selective knowledge distillation approach with attention-guided masking at both hidden and output layers. Instead of directly transferring attention maps, we employ a weighted sum aggregation of important token features using masked attention maps. Additionally, we employ token-level relationship knowledge distillation, operating on informative tokens selected using activation maps at the output layer, which enhances the student's understanding of intra-sample and inter-sample relationships.

3 APPROACH

In this section, we first provide an overview of the existing masked video modeling framework. Then we introduce MAGlobal and MALocal, which selectively transfer knowledge using hidden level attention maps in transformer blocks to capture temporally-global and temporally-local knowledge, respectively. Next, we present MAHint and MARel, which utilize output level activation-based attention maps to select important tokens and facilitate feature alignment and relationship transfer. Finally, we present the overall loss function, and the overall pipeline is shown in Figure 2.

3.1 Preliminary

Given an input video clip $X \in \mathbb{R}^{T\times 3\times H\times W}$, with temporal length T and spatial resolution $H\times W$, we divide the clip into non-overlapping 3D patches of size $t \times P_h \times P_w$. This division results in $N = \frac{T}{t} \times \frac{H}{P_h} \times \frac{W}{P_w}$ patches. These patches are then linearly projected into one-dimensional token embeddings $\mathbf{z} \in \mathbb{R}^{B\times N\times d}$, where B is the batch size and d is the dimension of the token embedding. Following the tube masking strategy employed in VideoMAE [56], we apply tube masking to retain only a small portion of tokens, we mask a subset of tokens with a high masking ratio r (e.g., 90%) while retaining only a small portion of tokens for further computation within the transformer blocks of the ViT encoder.

Within each multi-head self-attention (MHA) head of the transformer block, the visible tokens $\mathbf{z}'_{vis} \in \mathbb{R}^{B\times N_{vis}\times d}$ are linearly projected into queries, keys, and values $Q, K, V \in \mathbb{R}^{B\times h\times N_{vis}\times d_k}$ using parameter matrices $W^Q, W^K, W^V \in \mathbb{R}^{d\times d_k}$, respectively. Here, h represents the number of heads, $N_{vis} = (1-r)\times N$ is the number of visible patches, and $d_k = \frac{d}{h}$ is the dimension of the query, key, and value features. The attention distributions $A \in \mathbb{R}^{B\times h\times N_{vis}\times N_{vis}}$, capturing high-level semantic relationships between tokens, are computed using the scaled dot-product of queries and keys, normalized by $\sqrt{d_k}$ and passed through the softmax function. The output features $\mathbf{z}_{vis} \in \mathbb{R}^{B\times N_{vis}\times d}$ are generated by aggregating the value features V through a weighted sum operation, allowing each token to incorporate information from other tokens.

3.2 Hidden Level Masked Knowledge Transfer

In this subsection, we utilize attention maps to guide selecting and aggregating informative token features, allowing us to capture both temporally-global and temporally-local knowledge.

3.2.1 *MAGlobal: Temporally-Global Knowledge Transfer.*

Figure 1 (c) visualizes the attention distribution of a visible token (e.g., soccer) using the teacher model's multi-head attention. The red boxes highlight the Top-5 tokens with the highest attention values, indicating a higher focus on tokens with token-category-specific features. However, the existing attention mechanism aggregates all visible token features using a weighted sum operation without distinguishing between token-class-specific and non-specific tokens, which may not be optimal for knowledge distillation. To address this limitation and encourage the student model to focus on token-class-specific important knowledge, we propose a temporally-global Top-k selection process, denoted as $\mathcal{F}_{global}$. This process involves calculating a corresponding attention map using the key features for each query, selecting the Top-k attention values in A, and then using the positions corresponding to these Top-k attention values to choose the corresponding value features in V. The mask generation function $\mathcal{F}_{global}$ is defined as:

$$\mathcal{F}_{global} : \mathbb{R}^{B\times h\times N_{vis}\times N_{vis}} \rightarrow \mathbb{R}^{B\times h\times N_{vis}\times N_{vis}}. \tag{1}$$

It sets the Top-k values along the last dimension of the attention map to 1, while the rest $N_{vis} - k$ are set to 0. We then apply the mask generation function $\mathcal{F}_{global}$ to these attention maps to obtain the filtered attention maps M^s_{att-g} and M^t_{att-g}:

$$M^s_{att-g} = \mathcal{F}_{global}(A^s), \;\; M^t_{att-g} = \mathcal{F}_{global}(A^t), \tag{2}$$

where $A^s \in \mathbb{R}^{B\times h_s\times N_{vis}\times N_{vis}}$ and $A^t \in \mathbb{R}^{B\times h_t\times N_{vis}\times N_{vis}}$ are the attention maps generated by the student and teacher networks with h_s and h_t heads, respectively. The filtered attention maps retain only the Top-k values for each query, allowing the student model to focus on the most relevant token-class-specific knowledge associated with each query. The generated masks are applied to the original attention maps for temporally-global feature aggregation:

$$z^s_{att-g} = (M^s_{att-g} \odot A^s)V^s, \;\; z^t_{att-g} = (M^t_{att-g} \odot A^t)V^t, \tag{3}$$

where V^s and V^t are the value features from the student and teacher networks, respectively. To optimize knowledge transfer, a global attention loss function is defined as:

$$\mathcal{L}_{global} = \frac{1}{N_{global}} \sum_{i=1}^{N_{global}} \|\phi_s(z^s_{att-g})^{(i)} - \phi_t(z^t_{att-g})^{(i)}\|_2^2 \;, \tag{4}$$

where $\phi_s(\cdot)$ and $\phi_t(\cdot)$ are mapping functions to align the features from the student and teacher networks to a common dimension d_{proj}. And $N_{global} = B \times N_{vis} \times d_{proj}$.

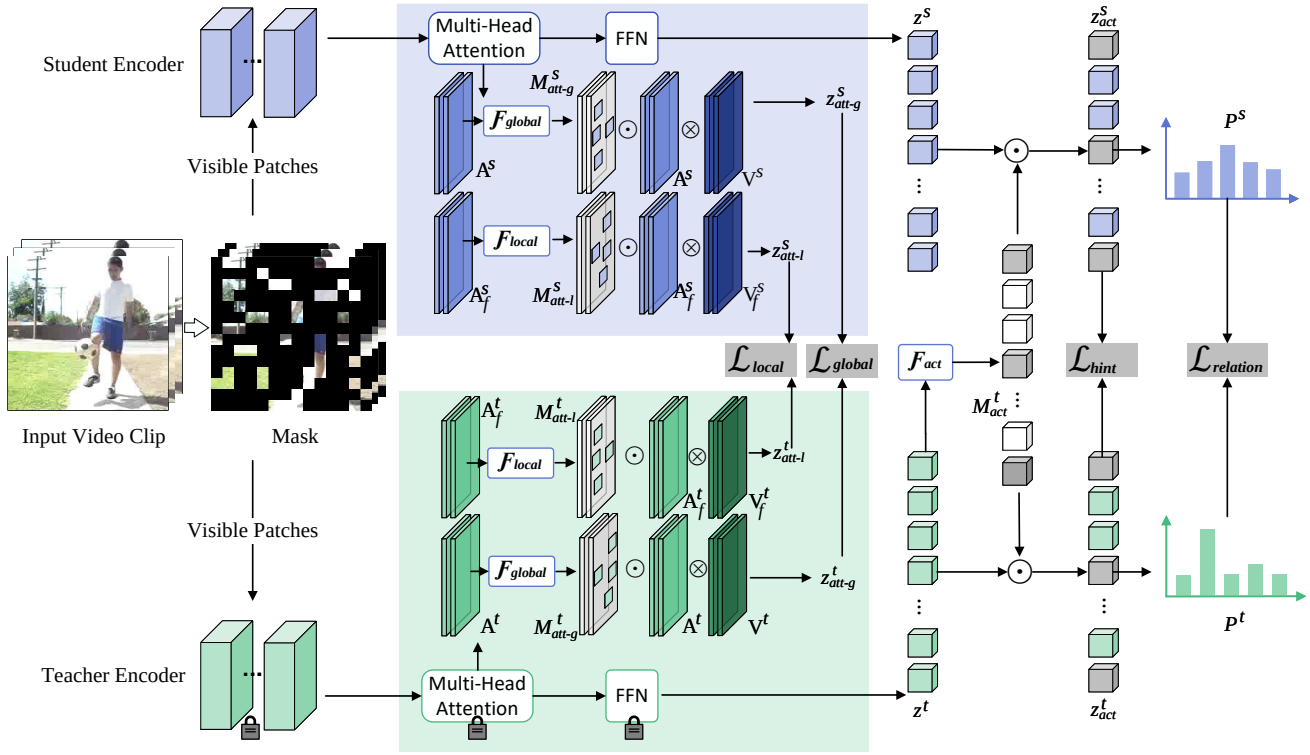


Figure 2: Overview of the MaskAgain model architecture: In the distillation stage, the teacher model remains fixed, and the student model is trained from scratch to match the teacher-encoded hidden level and output level knowledge, allowing for efficient pre-training without employing decoders in both models. Two attention-guided knowledge aggregation methods, $\mathcal{F}_{global}$ and $\mathcal{F}_{local}$, are utilized for temporally-global and temporally-local feature alignment with $\mathcal{L}_{global}$ and $\mathcal{L}_{local}$ at the hidden layer of the Transformer blocks. At the output layer of the Transformer block, an activation-based knowledge selection method, $\mathcal{F}_{act}$, is employed to propagate important token features, supervised by $\mathcal{L}_{hint}$, as well as their inter-sample and intra-sample relationships, supervised by $\mathcal{L}_{relation}$.

3.2.2 MALocal: Temporally-Local Knowledge Transfer.
To enable fine-grained temporally-local knowledge transfer and avoid global attention concentration on specific segments in the temporal dimension, we refine the temporally-global knowledge to the frame level by generating frame-level attention. This involves reshaping the attention maps to $A_f \in \mathbb{R}^{B\times h\times N_{vis}\times F\times N_{img}}$, where F is the temporal dimension of the input tokens, and $N_{img} = N_{vis}/F$ represents the number of visible patches per frame. We split the attention map for each query into F groups and perform knowledge selection on each frame using the attention map of that frame. The selection function $\mathcal{F}_{local}$ is defined as:

$$\mathcal{F}_{local} : \mathbb{R}^{B\times h\times N_{vis}\times F\times N_{img}} \rightarrow \mathbb{R}^{B\times h\times N_{vis}\times F\times N_{img}} . \quad (5)$$

It sets the Top-k attention values along the last dimension of each frame's attention map to 1, while the rest $N_{img} - k$ values are set to 0. This selection process allows us to retain only the most relevant information within each frame, which is then independently aggregated in the temporal dimension. The filtered attention maps M^s_{att-l} and M^t_{att-l} are obtained as follows:

$$M^s_{att-l} = \mathcal{F}_{local}(A^s_f),\ M^t_{att-l} = \mathcal{F}_{local}(A^t_f), \quad (6)$$

where $A^s_f \in \mathbb{R}^{B\times h_s\times N_{vis}\times F\times N_{img}}$ and $A^t_f \in \mathbb{R}^{B\times h_t\times N_{vis}\times F\times N_{img}}$ are the reshaped attention maps for the student and teacher networks, respectively. The filtered attention maps capture fine-grained knowledge within each frame.

Finally, the filtered attention maps are applied to the reshaped attention maps for temporally-local feature aggregation:

$$z^s_{att-l} = (M^s_{att-l} \odot A^s_f)V^s_f,\ z^t_{att-l} = (M^t_{att-l} \odot A^t_f)V^t_f, \quad (7)$$

where $V^s_f \in \mathbb{R}^{B\times h_s\times N_{img}\times F\times d_k}$ and $V^t_f \in \mathbb{R}^{B\times h_t\times N_{img}\times F\times d_k}$ are obtained by reshaping the value features from the student and teacher networks, respectively. This operation allows for temporally-local feature aggregation, enabling the student model to effectively incorporate fine-grained information from each frame to enhance its knowledge transfer and improve the overall performance.

To optimize temporally-local knowledge transfer, we define a local attention loss function as follows:

$$\mathcal{L}_{local} = \frac{1}{N_{local}} \sum_{i=1}^{N_{local}} \|\phi'_s(z^s_{att-l})^{(i)} - \phi'_t(z^t_{att-l})^{(i)}\|_2^2 , \quad (8)$$

where $N_{local} = B \times N_{vis} \times F \times d'_{proj}$, where $\phi'_s(\cdot)$ and $\phi'_t(\cdot)$ are mapping functions that align the features from the student and teacher networks to a common dimension d'_{proj}. When the hidden layer dimensions are the same, the mappings are identical. If the dimensions differ, simple fully connected layers (FC) are used for mapping, ensuring effective knowledge transfer and alignment during training.

In our study, we combine both loss functions for knowledge distillation in the hidden layers:

$$\mathcal{L}_{hidden} = \alpha \mathcal{L}_{global} + \beta \mathcal{L}_{local}, \tag{9}$$

where α and β are weighting factors that balance the contributions of the global attention loss $\mathcal{L}_{global}$ and the local attention loss $\mathcal{L}_{local}$. By optimizing this combined loss, our approach effectively transfers both global and fine-grained local knowledge, leading to improved generalization of learned representations in the student model.

3.3 Output Level Masked Knowledge Transfer

Motivated by the importance of token-level information and the potential redundancy in output level features as shown in Figure 1 (d), we propose a method for knowledge distillation based on the output layer of the Transformer encoder, z_L, where we denote the number of blocks (i.e., Transformer encoders) as L. We aim to identify and transfer essential token-level knowledge while avoiding the convergence difficulties that students may face when learning redundant information.

3.3.1 MAHint: Feature Alignment of Important Tokens.
In this section, we generate an activation-based attention map for the final output layer $z_L \in \mathbb{R}^{B \times N_{vis} \times d}$, inspired by the attention map generation method used in convolutional neural networks [68]. The attention map, denoted as A_{act}, is generated by computing the sum of absolute values along the last dimension of the output features z_L:

$$A_{act} = \sum_{i=1}^{d} |z_L[:, :, i]|, \tag{10}$$

where $z_L[:, :, i]$ represents the i-th channel of the tensor z_L, and d is the number of channels in z_L. The resulting tensor A_{act} has dimensions $B \times N_{\text{vis}}$, where each element indicates the importance of the corresponding token based on the sum of its absolute activation values.

We define the selection function $\mathcal{F}_{act}$ to select the Top-k tokens using the activation-based attention map A^t_{act} generated by the teacher. The function $\mathcal{F}_{act}$ masks out all but the Top-k visible tokens of each sample (the rest $N_{vis} - k$ tokens are set to 0):

$$\mathcal{F}_{act} : \mathbb{R}^{B \times N_{vis}} \rightarrow \mathbb{R}^{B \times N_{vis}}. \tag{11}$$

We then apply $\mathcal{F}_{act}$ to the output features from the teacher network, resulting in the generated masks M^t_{act}:

$$M^t_{act} = \mathcal{F}_{act}(A^t_{act}). \tag{12}$$

Next, we use these masks to filter the output features from both the teacher and the student network. The masks are used as an indexing operation φ on the output features, retaining only the Top-k tokens for further knowledge transfer:

$$z^s_{act} = \varphi(z^s_L, M^t_{act}), \;\; z^t_{act} = \varphi(z^t_L, M^t_{act}), \tag{13}$$

where $z_{act} \in \mathbb{R}^{K \times d}$ and $K = B \times k$.

To optimize the knowledge transfer, we define an activation loss function $\mathcal{L}_{hint}$. This loss function measures the discrepancy between the filtered output features of the student and teacher networks:

$$\mathcal{L}_{hint} = \frac{1}{N_{act}} \sum_{i=1}^{N_{act}} \|\phi''_s(z^s_{act})^{(i)} - (z^t_{act})^{(i)}\|_2^2, \tag{14}$$

where $N_{act} = K \times d$, and $\phi''_s(\cdot)$ represents a mapping of the output layer features from the student model to the teacher model to a common dimension d. This process ensures that only the crucial knowledge is transferred, avoiding convergence difficulties that may arise from redundant information.

3.3.2 MARel: Relationship Transfer of Important Tokens.
We propose MARel to maintain correlation consistency between the teacher and student networks. It transfers the correlation between informative tokens within a mini-batch, providing valuable token-level relationship knowledge to improve the student's performance. The relationship measurement is based on cosine similarity $sim(\cdot)$, and self-comparisons are avoided by setting the relationship value to -1000 for each token with itself. For both the teacher and student networks, we compute probability matrices $\mathbf{P}^t \in \mathbb{R}^{K \times K}$ and $\mathbf{P}^s \in \mathbb{R}^{K \times K}$ using softmax over the cosine similarities between selected important token features respectively, where K represents the number of selected important output features. The temperature parameters τ_t and τ_s are used independently for the teacher and student, respectively:

$$\mathbf{P}^t_{i,j} = \frac{\exp(sim(\mathbf{z}^t_i, \mathbf{z}^t_j)/\tau_t)}{\sum_{m=1}^{K} \sum_{n=1}^{K} \exp(sim(\mathbf{z}^t_m, \mathbf{z}^t_n)/\tau_t)}. \tag{15}$$

$$\mathbf{P}^s_{i,j} = \frac{\exp(sim(\mathbf{z}^s_i, \mathbf{z}^s_j)/\tau_s)}{\sum_{m=1}^{K} \sum_{n=1}^{K} \exp(sim(\mathbf{z}^s_m, \mathbf{z}^s_n)/\tau_s)}. \tag{16}$$

To optimize knowledge transfer, we minimize the Kullback–Leibler divergence between the probability distributions of the teacher and student, defined as:

$$\mathcal{L}_{relation} = KL(\mathbf{P}^t || \mathbf{P}^s). \tag{17}$$

In the final stage of knowledge distillation at the output layers, we combine the relationship loss $\mathcal{L}_{relation}$ with the activation loss $\mathcal{L}_{hint}$. The combined output loss function is formulated as:

$$\mathcal{L}_{output} = \lambda \mathcal{L}_{relation} + \gamma \mathcal{L}_{hint}, \tag{18}$$

where λ and γ are weighting factors that control the influence of each loss term during optimization. This integrated loss function effectively leverages relationship knowledge and activation-based filtering to facilitate knowledge transfer, resulting in improved generalization of the learned representations in the student network.

3.4 Overall Objective

The overall objective of the training process is to minimize the total loss, which is the sum of the individual loss functions. The total loss function $\mathcal{L}_{total}$ is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{hidden} + \mathcal{L}_{output}. \tag{19}$$

In this process, the student model learns to effectively match the teacher model's hidden level and output level knowledge, leading

to a more efficient and accurate student model. The MaskAgain algorithm is presented in Appendix B.

4 EXPERIMENTS

4.1 Implementation Details

We evaluate our approach on three video action recognition datasets: Kinetics-400 [28], UCF-101 [51], and HMDB-51 [29]. Kinetics-400 is a large-scale dataset with 240K video clips, covering 400 human action classes. UCF-101 and HMDB-51 are relatively smaller datasets, each containing approximately 9.5K/3.5K train/validation videos and 3.5K/1.5K train/validation videos, respectively, across 101 and 51 action classes. For self-supervised pre-training, we utilize the training set of either UCF-101 or Kinetics-400. During the evaluation, we follow the same split 1 as in [56] for UCF-101 and HMDB-51.

MaskAgain is applied to regular ViT models with different capacities, including ViT-S and ViT-B. On UCF101, the ViT-B teacher model is pre-trained on the UCF101 dataset for 3200 epochs. The ViT-B teacher model is pre-trained on K400 for 1600 epochs. In the distillation phase, the student model is pre-trained from scratch on the K400 dataset for 400 epochs. Subsequently, the pre-trained student model is fine-tuned on downstream video tasks. Both the pre-training and fine-tuning video clips have a length of 16 frames. The K400 experiments are conducted on 8 NVIDIA V100 GPUs, while the UCF101 experiments are conducted on 4 NVIDIA V100 GPUs. Moreover, we assign values of $d_{proj} = d'_{proj} = d = 768$ and utilize a Top-k approach with a threshold of Top-10. In the context of MaskAgain-B, we designate α as 3, β as 1, λ as 0.1, and γ as 0.2. For MaskAgain-S, we adjust β to 0.1, γ to 0.1 and we employ an MLP to project the student's output layer onto the teacher's feature space directly, using FC layers to map both the teacher and the student to the same feature space in the hidden layer, respectively. More comprehensive details can be found in the supplementary materials.

4.2 Comparison with state-of-the-art methods

We present a comprehensive comparison of MaskAgain with previous methods on three video recognition tasks. Table 1 shows that MaskAgain outperforms most self-supervised methods in terms of performance, while also being computationally efficient. Even when compared to video transformers pre-trained on ImageNet-21K, MaskAgain achieves competitive results. For instance, ViViT-L, with similar performance, incurs 4.4 times higher computational costs than MaskAgain. Additionally, MaskAgain-B achieves comparable performance to the teacher model, requiring only a quarter of the training epochs on the Kinetics-400 dataset.

We further evaluate MaskAgain's transfer learning capability on UCF101 and HMDB51 datasets. As shown in Table 2, MaskAgain-B outperforms previous works that rely on carefully designed contrastive learning and masked modeling methods. Compared to the teacher network, MaskAgain learns more transferable representations, achieving higher accuracies on both UCF101 (96.3% compared with 96.1%) and HMDB51 (77.1% compared with 73.3%). This result indicates that our approach learns more transferable representations without using pixel-level reconstruction.

Table 1: Comparison with existing works on Kinetics-400. MaskAgain-S denotes that the student is ViT-S and the teacher is ViT-B. The compute cost of a single view × the number of views (temporal clips with spatial crops) represents the inference cost (GFLOPs).

Method	Extra Data	Top-1	Top-5	GFLOPs	Param
SlowFast R101+NL [14]	-	79.8	93.9	234×30	60
X3D-XL [12]	-	79.1	93.9	48×30	11
MViTv1-B [11]	-	80.2	94.4	170×5	37
TSM-8 [32]	IN-1K	74.1	91.2	43×30	24
LGD [46]	IN-1K	79.4	94.4	N/A	N/A
VideoSwin-B [35]	IN-1K	80.6	94.6	282×12	88
Uniformer-S [30]	IN-1K	79.8	93.4	110×12	21
ViT-VTN [39]	IN-21K	78.6	93.7	4218×1	11
TimeSformer [4]	IN-21K	80.7	94.7	2380×3	121
Mformer-B [43]	IN-21K	79.7	94.2	370×30	109
Mformer-L [43]	IN-21K	80.2	94.8	1185×30	382
X-ViT [5]	IN-21K	80.2	94.7	N/A×3	N/A
ViViT-L FE [1]	IN-21K	81.7	93.8	3980×3	N/A
BEVT Swin-B [59]	IN-1K	80.6	N/A	282×12	88
OmniMAE ViT-B [17]	IN-1K	80.8	N/A	180×15	87
ST-MAE ViT-B [13]	-	81.3	94.9	180×21	87
VideoMAE ViT-S [56]	-	79.0	93.8	57×15	22
VideoMAE ViT-B [56]	-	81.5	95.1	180×15	87
MaskAgain-S	-	78.7	93.6	57×15	22
MaskAgain-B	-	81.0	94.6	180×15	87

Table 2: Comparison with previous methods on UCF101 and HMDB51.

Method	Extra Data	Param	UCF101	HMDB51
VideoMoCo R2+1D [41]	K400	15	78.7	49.2
MemDPC R2D3D [20]	K400	32	86.1	54.5
Vi^{2}CLR S3D [8]	K400	9	89.1	55.7
CORP Slow-R50 [24]	K400	32	93.5	68.0
CVRL Slow-R50 [45]	K400	32	92.9	67.9
CVRL Slow-R152 [45]	K600	328	94.4	70.6
ρBYOL Slow-R50 [15]	K400	32	94.2	72.1
VIMPAC ViT-L [54]	HowTo100M	307	92.7	65.9
VideoMAE ViT-B [56]	K400	87	96.1	73.3
MaskAgain-S	K400	22	92.9	72.0
MaskAgain-B	K400	87	**96.3**	**77.1**

The consistent performance across datasets, as demonstrated in Tables 1 and 2, highlights the robustness of MaskAgain without requiring any modifications. Additionally, experimental results on Something-Something V2 (SSV2) [18] show that MaskAgain exhibits stronger temporal modeling capabilities. For more experimental details regarding SSV2, please refer to Appendix A.3.

4.3 Ablation Experiments

We conduct ablation experiments to analyze the impact of different components in MaskAgain. The models are pre-trained on the UCF-101 dataset for 100 epochs, with ViT-B as the encoder. The batch sizes used for pre-training and fine-tuning are 24 and 16, respectively, for a duration of 50 epochs. For the experiments reported in Table 3 and 6, the models are trained for 400 epochs of pre-training on the UCF-101 dataset, followed by 100 epochs of fine-tuning.

Table 3: Effect of individual components in MaskAgain.

Hint	MAHint	MAGlobal	MALocal	MARel	UCF101
✓					88.1
	✓				88.5
	✓			✓	88.7
	✓		✓		89.2
	✓	✓			89.6
	✓		✓	✓	89.6
	✓	✓	✓		89.7
	✓	✓		✓	89.7
	✓	✓	✓	✓	**89.8**

Influence of Individual Components. We conduct ablation experiments to analyze the impact of different components in MaskAgain for video action recognition. The experiments are performed on the UCF-101 dataset using the ViT-B model as the encoder. We present the results in Table 3.

Comparison with Hint: We compare MAHint with the baseline method Hint, which employs per-token alignment without selection on the output level of the model. MAHint, which incorporates important token selection before knowledge distillation, outperforms Hint, showcasing its effectiveness.

Effect of Individual Components: We evaluate the contributions of MAGlobal, MALocal, and MARel, the individual components in MaskAgain. Combining these components leads to performance improvements, demonstrating their complementary nature in enhancing knowledge distillation.

Efficiency and Scalability: MaskAgain exhibits efficiency and scalability advantages. Despite only 400 epochs of training, it achieves performance comparable to a teacher model pre-trained for 3200 epochs and fine-tuned under the same conditions (achieves 91.3% on UCF101 dataset reported in VideoMAE [56]), emphasizing its practicality and effectiveness.

In conclusion, our ablation experiments validate the efficiency and effectiveness of MaskAgain in video representation learning.

Comparison of Different Knowledge Distillation Methods in the Hidden Level. We compare different methods for knowledge distillation in the hidden level of the transformer block, focusing on their effectiveness in transferring attention maps and value features. The evaluated methods include: (1) Attention Transfer: Directly aligning attention maps. (2) Value Transfer: Directly aligning value features. (3) Weighted Value Transfer: Aggregating value features using attention maps without selection. (4) MMGlobal: Aggregating value features using masked attention maps with knowledge selection.

Table 4 provides a comprehensive analysis of these methods on the UCF101 dataset, considering the loss functions used, the layer index for knowledge transfer, and the temperature parameter (τ_{att}) applied to the attention distributions. From the results, it is evident that MMGlobal stands out as the most effective method, achieving a Top-1 accuracy of 62.2% on the UCF101 dataset, which surpasses the performance of other knowledge distillation methods. MMGlobal's ability to aggregate essential information through weighted value transfer based on masked attention maps showcases its efficacy in knowledge transfer within the transformer block.

Influence of Top-k Selection Strategy. We investigate the impact of the "Top-k" selection strategy on the Top-1 accuracy for the UCF101 dataset, considering different components of MaskAgain. From the results presented in Figure 3, we observe the following trends for different components: (1) *MAHint* with "Top-k" selection achieves higher accuracy compared to transferring token features without selection (58.3%) and MAHint with a single selected token. For example, MAHint-k with 10 selected tokens achieves a Top-1 accuracy of 64.2%. (2) *MAGlobal* with 10 selected tokens outperforms MAHint with 1 selected token, achieving a Top-1 accuracy of 61.3%. However, increasing the number of selected tokens to 80 results in a slight drop in accuracy to 60.0%. (3) *MALocal* with 10 selected tokens achieves the highest Top-1 accuracy of 63.8%. Increasing or decreasing the number of selected tokens from this value results in a decline in performance. These findings show that appropriate k values contribute to the performance improvement of each component.

Table 4: Comparison of different methods for knowledge distillation in the hidden level of the transformer block. The model is pre-trained with both pixel reconstruction loss and knowledge distillation loss as supervision, except for the baseline in the first row.

Method	Loss Function	Layer Index	τ_{att}	UCF101
VideoMAE	MSELoss	16	1	39.9
Attention Transfer	KL-div	12	4	57.6
Attention Transfer	KL-div	12	1	57.0
Attention Transfer	MSELoss	12	1	57.3
Attention Transfer	MSELoss	8	1	53.3
Value Transfer	MSELoss	12	1	59.3
Weighted Value Transfer	MSELoss	12	4	50.5
Weighted Value Transfer	MSELoss	12	1	58.0
Weighted Value Transfer	MSELoss	12	0.1	59.9
MAGlobal	MSELoss	12	1	**62.2**

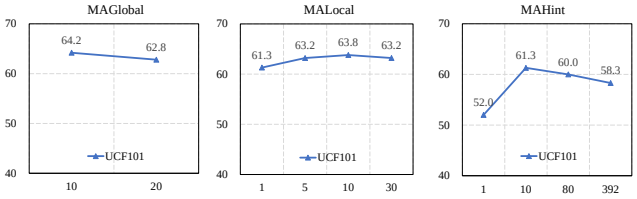


Figure 3: Influence of k: The horizontal axis of each subplot represents the k values, indicating the number of selected attention values on the attention maps for knowledge aggregation and transfer. The vertical axis represents the Top-1 accuracy on the UCF101 dataset.

Table 5: Influence of Position Selection Strategies.

Selection Strategy	UCF101
Selected by the teacher	63.7
Selected separately	**64.2**

Influence of Position Selection Strategies. We investigate different knowledge selection strategies in the MAGlobal loss of the middle-level MaskAgain. Two strategies are evaluated: "Selected by the teacher," where the teacher alone chooses positions, and the student model uses the selected positions to mask attention maps and aggregate value features; "Selected separately," where both the teacher and student independently select positions. Table 5 shows

Table 6: Influence of Temperature Coefficients.

Temperature		UCF101
τ_s	τ_t	
0.1	0.1	89.0
0.05	0.05	88.8
0.1	0.05	**89.8**

the impact of these strategies on UCF101 accuracy. Selecting positions separately leads to the highest accuracy of 64.2%, enabling dynamic knowledge selection. When positions are solely selected by the teacher, the accuracy slightly decreases to 63.7%.

Influence of Temperature Coefficient. We investigate the influence of temperature coefficients τ_s and τ_t on knowledge distillation. Table 6 shows that when τ_t is smaller than τ_s, we achieve an impressive accuracy of 89.8% on UCF101. This improvement may be attributed to the fact that a smaller τ_t produces a sharper target relationship distribution compared to the student's distribution. The sharpness emphasizes confident token relationships, leading to more effective knowledge transfer by focusing on important and relevant information while reducing the transfer of less relevant ones.

4.4 Visualization Analysis

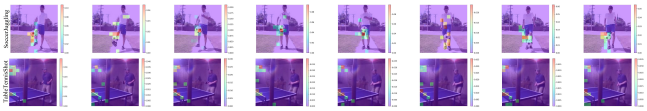
We visualize the attention maps of two action classes for three types of masked attentions defined in our method: temporally-local attention (M_{att-l}) used in MALocal, temporally-global attention (M_{att-g}) used in MAGlobal, and activation-based attention (M_{act}) generated by MAHint. As shown in Fig. 4, each attention method exhibits different patterns of token selection, emphasizing distinct visual cues for action recognition.

M_{att-l} focuses on frame-level token selection, capturing fine-grained temporal changes and visual patterns within each frame. This enables the student model to understand localized motion patterns and distinguish between different actions.

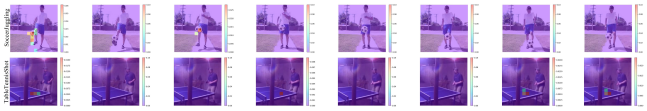
M_{att-g} performs global-level token selection, considering the entire video sequence to capture consistent context information across frames. It attends to tokens representing recurring visual patterns for a broader understanding of the video context.

M_{act} employs activation-based token selection, prioritizing tokens based on their significance in depicting the action subject or any discernible elements within the scene. It highlights tokens that carry significant semantic information associated with the visual cues of the action.

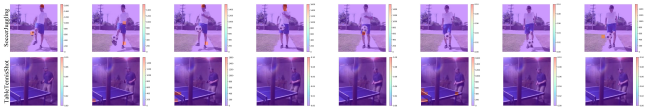
By leveraging these attention mechanisms for token selection, our MaskAgain framework effectively emphasizes informative tokens while filtering out less relevant scene information. This selective token attention enhances knowledge distillation efficiency and contributes to the improved performance of our masked video modeling approach. The visualization of masked attention maps offers insights into how each method guides the selection of important tokens, enriching the learned feature representations and leading to accurate action recognition.



(a) Temporally-Local masked attention maps (M_{att-l}).



(b) Temporally-Global masked attention maps (M_{att-l}).



(c) Activation-Based masked attention maps (M_{act}).

Figure 4: Visualization of masked attention maps from different components of MaskAgain. The attention mechanisms guide token selection to emphasize essential visual cues relevant to action recognition while filtering out less relevant scene information.

5 CONCLUSION

In conclusion, we have introduced MaskAgain, an efficient masked knowledge distillation framework tailored for masked video modeling. By addressing the challenges of memory consumption and effective knowledge transfer in transformer models with large-scale unlabeled data, MaskAgain offers a robust solution for pre-training video models. In the hidden layer, MaskAgain leverages attention maps to select crucial token features at both temporally-global and temporally-local levels, enabling more efficient and meaningful feature representations. At the output layer, token-level activation-based attention is utilized for knowledge transfer, with two methods, MAHint and MARel, explored to align features and maintain correlation consistency between teacher and student networks. The experimental results on Kinetics-400, UCF101, and HMDB51 datasets showcase the effectiveness of MaskAgain, achieving impressive accuracies comparable to or even surpassing models trained for longer epochs. This highlights the efficiency and efficacy of our approach in pre-training highly accurate video models while optimizing computational resources and training time. The integration of attention mechanisms in selective knowledge distillation paves the way for further advancements in computer vision tasks. Future research can explore leveraging attention-based selective distillation for other tasks and modalities, as well as investigating heterogeneous distillation between transformer and convolution models using activation value consistency.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (No. 62006140). Additionally, it is also supported in part by the Shandong Provincial Natural Science Foundation (No.: ZR2020QF106) and is in part based on research sponsored by Shenzhen Higher Education Institutions Stable Support Program (Key Project) under award number GXWD20220817123150002.

REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: a video vision transformer. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 6816–6826.

[2] Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. 2023. Masked autoencoders enable efficient knowledge distillers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 24256–24265.

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

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *Proceedings of the International Conference on Machine Learning*.

[5] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. 2021. Space-time mixing attention for video transformer. *Advances in neural information processing systems* 34 (2021), 19594–19607.

[6] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative pretraining from pixels. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1691–1703.

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops*. IEEE, 3008–3017.

[8] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. 2021. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1482–1492.

[9] Shuangrui Ding, Rui Qian, and Hongkai Xiong. 2022. Dual Contrastive Learning for Spatio-temporal Representation. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 5649–5658.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[11] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 6804–6815.

[12] Christoph Feichtenhofer. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 203–213.

[13] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. 2022. Masked Autoencoders As Spatiotemporal Learners. *Advances in Neural Information Processing Systems* (2022).

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 6201–6210.

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

[16] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross B. Girshick, and Kaiming He. 2021. A Large-Scale Study on Unsupervised Spatiotemporal Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3299–3309.

[17] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2022. OmniMAE: Single Model Masked Pretraining on Images and Videos. *arXiv* (2022).

[18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*. 5842–5850.

[19] Tengda Han, Weidi Xie, and Andrew Zisserman. 2019. Video Representation Learning by Dense Predictive Coding. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. IEEE, 1483–1492.

[20] Tengda Han, Weidi Xie, and Andrew Zisserman. 2020. Memory-augmented dense predictive coding for video representation learning. In *Proceedings of the European Conference on Computer Vision*. Springer, 312–329.

[21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2021), 15979–15988.

[22] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv* (2015).

[23] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv* (2012).

[24] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. 2021. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 7919–7929.

[25] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. 2016. Deep networks with stochastic depth. In *Proceedings of the European Conference on Computer Vision*. Springer, 646–661.

[26] Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, and Qixiang Ye. 2023. Generic-to-Specific Distillation of Masked Autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 15996–16005.

[27] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Conference on Empirical Methods in Natural Language Processing*. 4163–4174.

[28] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[29] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2556–2563.

[30] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. 2022. UniFormer: Unified Transformer for Efficient Spatial-Temporal Representation Learning. In *International Conference on Learning Representations*. OpenReview.net.

[31] Xiaojie Li, Jianlong Wu, Hongyu Fang, Yue Liao, Fei Wang, and Chen Qian. 2020. Local correlation consistency for knowledge distillation. In *Proceedings of the European Conference on Computer Vision*. Springer, 18–33.

[32] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 7082–7092.

[33] Yufan Liu, Jiajiong Cao, Bing Li, Chunfeng Yuan, Weiming Hu, Yangxi Li, and Yunqiang Duan. 2019. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7096–7104.

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE International Conference on Computer Vision*. 10012–10022.

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2022. Video swin transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3202–3211.

[36] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. 2021. Tam: Temporal adaptive module for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 13688–13698.

[37] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* (2016).

[38] Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. (2018).

[39] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. 2021. Video Transformer Network. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. IEEE.

[40] Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. 2022. Search-oriented micro-video captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*. 3234–3243.

[41] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. 2021. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 11205–11214.

[42] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3967–3976.

[43] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and Joao F. Henriques. 2021. Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers. In *Advances in Neural Information Processing Systems*. 12493–12506.

[44] Baoyun Peng, Xiao Jin, Jiaheng Liu, Shunfeng Zhou, Yichao Wu, Yu Liu, Dongsheng Li, and Zhaoning Zhang. 2019. Correlation Congruence for Knowledge Distillation. (2019), 5006–5015.

[45] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. 2021. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6964–6974.

[46] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. 2019. Learning Spatio-Temporal Representation With Local and Global Diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 12056–12065.

[47] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the International Conference on Machine Learning*.

[48] Jun Rao, Xv Meng, Liang Ding, Shuhan Qi, and Dacheng Tao. 2022. Parameter-efficient and student-friendly knowledge distillation. *arXiv* (2022).

[49] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representations*.

[50] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. 2021. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10389–10399.

[51] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* (2012).

[52] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2158–2170.

[53] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2818–2826.

[54] Hao Tan, Jie Lei, Thomas Wolf, and Mohit Bansal. 2021. VIMPAC: Video Pre-Training via Masked Token Prediction and Contrastive Learning. *arXiv* (2021).

[55] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*. OpenReview.net.

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

[57] Luting Wang, Xiaojie Li, Yue Liao, Zeren Jiang, Jianlong Wu, Fei Wang, Chen Qian, and Si Liu. 2022. Head: Hetero-assists distillation for heterogeneous object detectors. In *Proceedings of the European Conference on Computer Vision*. Springer, 314–331.

[58] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. 2021. TDN: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1895–1904.

[59] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. 2022. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 14713–14723.

[60] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. 2023. Masked Video Distillation: Rethinking Masked Feature Modeling for Self-supervised Video Representation Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[61] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems* 33 (2020), 5776–5788.

[62] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2022. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 14648–14658.

[63] Jianlong Wu, Wei Sun, Tian Gan, Ning Ding, Feijun Jiang, Jialie Shen, and Liqiang Nie. 2023. Neighbor-Guided Consistent and Contrastive Learning for Semi-Supervised Action Recognition. *IEEE Transactions on Image Processing* (2023).

[64] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 9643–9653.

[65] Hongwei Xue, Peng Gao, Hongyang Li, Yu Qiao, Hao Sun, Houqiang Li, and Jiebo Luo. 2022. Stare at What You See: Masked Image Modeling without Reconstruction. *arXiv* (2022).

[66] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In *In Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 6022–6031.

[67] Sukmin Yun, Jongjin Park, Kimin Lee, and Jinwoo Shin. 2020. Regularizing class-wise predictions via self-knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 13873–13882.

[68] Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv* (2016).

[69] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*. OpenReview.net.

Algorithm 1 Pseudocode of MaskAgain in PyTorch style.

```
# global_att: the attention map (A) of the last encoder
# global_value: the value (V)of the last encoder
# local_att: the reshaped global_att (A_f)
# local_value: the reshaped global_value (V_f)
# feature: the output of the L-th encoder (z_t)
# *_s: from the student model
# *_t: from the teacher model
    # Generate global_masks and select topk global_attentions
        respectively
    global_bool_topk_pos_t = topk(softmax(global_att_t))
    global_bool_topk_pos_s = topk(softmax(global_att_s))
    global_att_s = att_s * global_bool_topk_pos_s
    global_att_t = att_t * global_bool_topk_pos_t

    # Compute global loss
    L_global = mseloss((global_att_s @ global_value_s), (global_att_t
        @ global_value_t) )

    # Generate local_masks and select topk local_attentions
        respectively
    local_bool_topk_pos_t = topk(softmax(local_att_t))
    local_bool_topk_pos_s = topk(softmax(local_att_s))
    local_att_s = local_att_s * local_bool_topk_pos_s
    local_att_t = local_att_t * local_bool_topk_pos_t

    # Compute local loss
    L_local = mseloss((local_att_s @ local_value_s), (local_att_t @
        local_value_t))

    # Compute hidden loss
    L_hidden = α * L_global + β * L_local

    # Generate the activation-based attention
    activation_t = abs(feature_t).sum(-1)

    # Generate masks and select topk activation-based attentions
        respectively
    bool_topk_pos = topk(activation_t)
    sd = feature_s[bool_topk_pos]
    td = featuret_t[bool_topk_pos]

    # Compute hint loss
    L_hint= mseloss(sd, td)

    #Calculate the relation of the topk token
    relation_s = einsum(sd,sd)
    relation_t = einsum(td,td)

    # Calculate relation loss
    L_relation = kl_div(relation_s, relation_t)

    #Calculate output loss
    L_output = λ * L_relation + γ * L_hint

    #Calculate total loss
    L_total = L_hidden + L_output
```

A IMPLEMENTATION DETAILS

A.1 Pre-training setting

The pre-training experiments involve training the MaskAgain model on both the Kinetics-400 and UCF101 datasets. The pre-training settings for each dataset are provided in Table 7.

Table 7: Pre-training setting.

config	Kinetics-400	UCF101
optimizer	AdamW [38]	
base learning rate	1.5e-4	3e-4
weight decay	0.05	
optimizer momentum	β_1,β_2=0.9,0.95 [6]	
batch size	512	256
mask ratio	90%	75%
learning rate schedule	cosine decay [37]	
warmup epochs	40	
augmentation	MultiScaleCro	
drop path	0.1	0
short edge	256	240

A.2 Fine-tuning setting

The fine-tuning experiments are performed on four datasets: Kinetics-400, UCF101, HMDB51, and Something-Something V2. The fine-tuning settings for each dataset are provided in Table 8.

A.3 Results on Something-Something V2

Something-Something V2 is a challenging video dataset with diverse human-object interaction actions, making it valuable for evaluating temporal modeling capabilities. We fine-tuned the pre-trained MaskAgain student model from K400 on Something-Something V2. The performance comparison with previous works on Something-Something V2 is shown in Table 9. Despite having a smaller batch size and lower video resolution compared to SOTA methods, MaskAgain outperforms all supervised and self-supervised methods listed in the table.

Compared to VideoMAE (ViT-S), MaskAgain-S achieves a +0.3% Top-1 accuracy gain on Something-Something V2 with limited computational resources. Furthermore, MaskAgain-B outperforms the teacher model (pre-trained on K400 for 1600 epochs) by +0.6% using only 1/4 of the fine-tuning batch size. This indicates that MaskAgain has a stronger temporal modeling capability even with limited computational resources, making it effective for tasks that require better temporal information, such as Something-Something V2.

B THE PSEUDO-CODE OF MASKAGAIN

Table 8: Fine-tuning setting of MaskAgain.

config	Kinetics-400	UCF101	HMDB51	Something-Something V2
optimizer	AdamW			
base learning rate	1e-3	5e-4	1e-3	1e-3 (S), 5e-4 (B)
weight decay	0.05			
optimizer momentum	β_1, β_2=0.9, 0.999			
batch size	256	128	128	256 (S), 128 (B)
temporal stride	4			
warmup lr	1e-6	1e-8	1e-6	1e-8
min lr	1e-5			
inference protocol	5 clips × 3 crops	5 clips × 3 crops	5 clips × 3 crops	2 clips × 3 crops
learning rate schedule	cosine decay			
warmup epochs	5			
training epochs	150 (S), 75 (B)	100	50	40
short edge	256	240	320	240
repeated augmentation	2			
flip augmentation	*yes*	*yes*	*yes*	*no*
RandAug [7]	(9, 0.5)			
label smoothing [53]	0.1			
mixup [69]	0.8			
cutmix [66]	1.0			
drop path [25]	0.1	0.2	0.2	0.1
dropout [23]	0	0.5	0.5	0.5
layer-wise lr decay [3]	0.75	0.7	0.7	0.7 (S),0.75 (B)

Table 9: Comparison with previous works on Something-Something V2.

method	extra data	Top-1	GFLOPs	Param
supervised				
SlowFast R101 [14]	K400	63.1	106×3	53
TSM-RGB R50 [32]	IN-1K	63.3	62×6	24
TAM R50 [36]	IN-1K	66.0	99×6	51
TDN R101 [58]	IN-1K	69.6	198×3	88
MViTv1-B [11]	-	67.7	455×3	37
TimeSformer-HR [4]	IN-21K	62.5	1703×3	121
ViViT-L FE [1]	IN-21K+K400	65.9	995×12	N/A
Mformer-B [43]	IN-21K+K400	66.5	370×3	109
Mformer-L [43]	IN-21K+K400	68.1	1185×3	382
VideoSwin-B [35]	IN-21K+K400	69.6	321×3	88
self-supervised				
VIMPAC ViT-L [54]	HowTo100M	68.1	N/A×30	307
VideoMAE ViT-S [56]	K400	66.4	57×6	22
VideoMAE ViT-B [56]	K400	69.7	180×6	87
OmniMAE ViT-B [17]	IN-1K	69.5	180×6	87
OmniMAE ViT-B [17]	IN-1K+K400	69.0	180×6	87
MaskAgain-S	K400	66.7	57×6	22
MaskAgain-B	K400	70.3	180×6	87