EVEREST: Efficient Masked Video Autoencoder by Removing Redundant Spatiotemporal Tokens

Sunil Hwang [*1] **Jaehong Yoon** [*2] **Youngwan Lee** [*3 4] **Sung Ju Hwang** [3 5]

Abstract

Masked Video Autoencoder (MVA) approaches have demonstrated their potential by significantly outperforming previous video representation learning methods. However, they waste an excessive amount of computations and memory in predicting uninformative tokens/frames due to random masking strategies. (e.g., over 16 nodes with 128 NVIDIA A100 GPUs (Feichtenhofer et al., 2022)). To resolve this issue, we exploit the unequal information density among the patches in videos and propose ***EVEREST***, a *surprisingly efficient* MVA approach for video representation learning that finds tokens containing rich motion features and discards uninformative ones during both pre-training and fine-tuning. We further present an information-intensive frame selection strategy that allows the model to focus on informative and causal frames with minimal redundancy. Our method significantly reduces the computation and memory requirements of MVA, enabling the pre-training and fine-tuning on **a single machine with 8 GPUs** while achieving comparable performance to computation- and memory-heavy baselines on multiple benchmarks and the uncurated Ego4D dataset. We hope that our work contributes to reducing the barrier to further research on video understanding.

1. Introduction

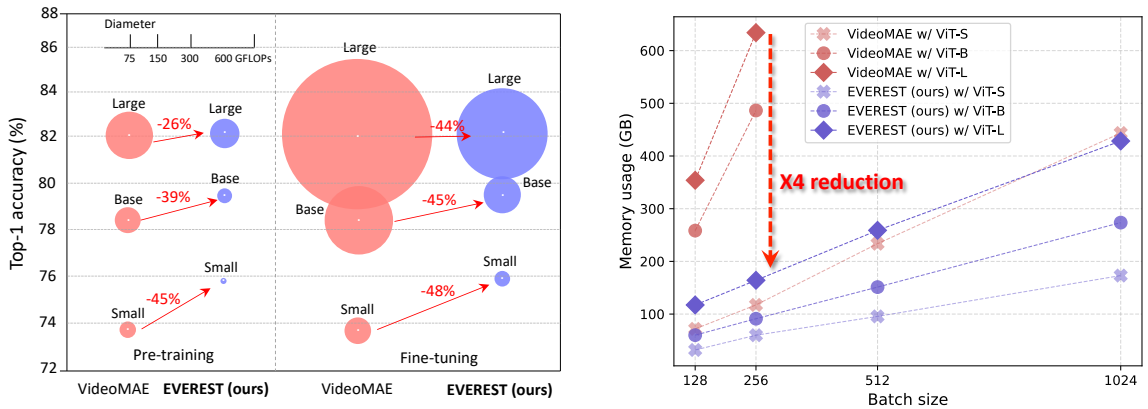
Massive video data floods the web and media daily with the rapid growth of portable devices equipped with cameras, such as AR glasses, smartphones, UAVs, and robots. However, direct utilization of user-generated video data to solve target task problems is nontrivial, as annotating them is time-consuming and expensive. One potential approach to tackle this problem is to learn generic representations from unlabeled video streams that can transfer to diverse downstream visual tasks. Video Representation Learning (VRL) (Fernando et al., 2017; Piergiovanni et al., 2020; Qian et al., 2021; Pan et al., 2021) methods allow learning spatial and temporal features from input video frames in a self-supervised manner without any human annotations. A caveat to such pre-training for video tasks is that, unlike image-based tasks with static information of objects in instances, video-based tasks involve temporal causality; that is, successive frames are strongly correlated in their semantics.

Recently, Masked Video Autoencoder (MVA) (Tong et al., 2022; Feichtenhofer et al., 2022), which learns to reconstruct randomly masked spatiotemporal regions in video clips, has shown impressive performance on various video-based problems. Yet, critical challenges remain in efficiently exploiting the spatiotemporal information in real-world videos: (1) Tokens (a pair of two temporally successive patches in the same space) in videos are not equally valuable to reconstruct, as the amount of their information depends not only on spatial importance but also on temporal redundancy. (2) Learning representations from videos is infeasible without a huge computing budget. MVA approaches (Feichtenhofer et al., 2022; Wang et al., 2023b;c) that reconstruct the *whole* video require excessively large amounts of computations, making it impractical without access to a substantial GPU cluster. For example, VideoMAE (Tong et al., 2022) takes about 27 hours to pre-train for 800 epochs with ViT-B using **64 NVIDIA V100 (32GB) GPUs**, and ST-MAE (Feichtenhofer et al., 2022) takes about 35.8 hours to pre-train for 800 epochs with ViT-L using **128 NVIDIA A100 (80GB) GPUs**.

To overcome such limited feasibility of video representation learning, we propose an ***E**fficient Masked **V**ideo AutoEncoder which removes **RE**dundant **S**patiotemporal **T**okens (**EVEREST**)*, that is a highly efficient training algorithm with token selection to alleviate the extremely high resource requirements for VRL. Unlike previous meth-

[*]Equal contribution [1]Korea Military Academy [2]UNC Chapel Hill [3]KAIST [4]ETRI [5]DeepAuto. Correspondence to: Sunil Hwang <sunil.hwang93@kma.ac.kr>, Sung Ju Hwang <sjhwang82@kaist.ac.kr>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) Computation (GFLOPs) Comparison **(b) Memory comparison**

Figure 1: **Efficiency of our EVEREST against VideoMAE on K400 dataset. (a) GFLOPs** for pre-training and fine-tuning. The bubble size is proportional to the GFLOPs of the model. **(b) Memory consumption** using one node equipped with 8× A100 (80GB). VideoMAE with ViT-B and -L fails to deploy the model due to ***out-of-memory*** if the batch sizes are 512 or larger. For a ViT-L backbone with a batch size of 256, our method achieves about 4× less memory consumption than VideoMAE. Please see Tabs. 1 and 3 for detailed results.

ods (Tong et al., 2022; Feichtenhofer et al., 2022; Wang et al., 2023a) reconstructing *all* patches, we aim to select a subset of *informative* visual tokens to learn based on the distance disparity across temporally adjacent tokens in the embedding space. Our *redundancy-robust token selection* approach successfully detects meaningful changes in the state/object of incoming videos in an online manner, discarding less meaningful tokens containing redundant information or meaningless backgrounds, without resorting to dense motion features from the incoming video, such as HOG or optical flows (Sun et al., 2023; Wang et al., 2023c).

This allows us to significantly reduce computational cost and memory usage while maintaining the quality of a representation model by backpropagating to only a few selected tokens that retain rich spatiotemporal information with minimal redundancy. As shown in Fig. 1, our EVEREST **saves computation costs by 26 ∼ 45%** in pre-training and **44 ∼ 48%** in fine-tuning across varying ViT scales while achieving competitive performance against strong VideoMAE baselines. We also achieve an impressive amount of memory reduction with all ViT backbones, which becomes more effective with larger models. For example, on a ViT-Large with a batch size of 256, our method achieves about **4× smaller memory consumption** than VideoMAE and **enables single-node training with larger batch sizes** using a large backbone, whereas existing VRL methods require immense memory occupancy and fail to train in the same setup as they go *out-of-memory*.

In addition, most current VRL (Arnab et al., 2021; Bertasius et al., 2021; Feichtenhofer et al., 2022; Tong et al., 2022) methods uniformly load frames at regular time intervals from each mini-batch clip. However, this strategy does not consider a temporal imbalance in information and noise that matters to real-world uncurated videos. Our EVEREST further performs *information-intensive frame selection*, which is carried out online through *probabilistic sampling* proportional to the ratio of redundancy-robust tokens in each frame and does not need any additional computational or parametric learning steps. Consequently, we can capture abundant spatiotemporal knowledge from highly sparse yet informative spatiotemporal regions in videos.

We extensively validate our proposed method on multiple benchmark datasets, including UCF101, HMDB51, K400, Something-Something V2, and Ego4D, and our EVEREST shows remarkable efficiency in terms of memory occupancy, computational cost, and training time compared to strong counterparts, achieving competitive performance.

Our contributions are as follows:

- We propose *redundancy-robust token selection*, an *efficient* VRL method that promptly selects the most informative tokens based on the states' change and discards redundant ones in an online manner, avoiding wasteful training on uninformative regions of videos.
- We further propose *information-intensive frame selection*, a strategy to select informative video frames from incoming videos, which allows the model to efficiently learn robust and diverse temporal representations in real-world uncurated videos.
- Our EVEREST has great potential to lower the barrier for video-related research that requires enormous

computing power and cost, as it shows comparable performance to existing methods while significantly reducing the computations, memory, and training time.

2. Related Work

Masked video autoencoder Inspired by self-supervised learning with Masked Image Modeling (He et al., 2022; Xie et al., 2021; Kakogeorgiou et al., 2022), several recent works on video representation learning (Wang et al., 2022b; Sun et al., 2023; Wang et al., 2023a;c) suggest spatiotemporal masking strategies given video streams. To capture spatial representation and temporal dynamics for unsupervised video streams, ST-MAE (Feichtenhofer et al., 2022) and VideoMAE (Tong et al., 2022) extend a masked image autoencoder to mask partial regions in arriving video clips via random and space-only random sampling, respectively. They find that spatiotemporal inductive bias in video clips helps a decoder predict input pixels in masked regions, allowing a higher masking ratio ($\sim 90\%$) than MIM ($\sim 60\%$ (Xie et al., 2021) and $\sim 75\%$ (He et al., 2022)) on image self-supervised learning. BEVT (Wang et al., 2022b) proposes to train image- and video-level masked autoencoders jointly by sharing weights of the encoder, formulated with Video Swin (Liu et al., 2022). They resort to random sampling given spatiotemporal inductive bias, which can be a good approximator with stochasticity during data-driven training. Nevertheless, selecting random tokens to reconstruct for Masked video autoencoder is inefficient since embedded tokens in video frames are not equally important, especially since the informativeness of each token is affected by the adjacent frames.

Input selective training As benchmark training datasets often have massive scales and contain a lot of redundant or less meaningful instances, various works (Fayyaz et al., 2022; Wu et al., 2019; Yoon et al., 2022) have discussed sampling important instances from the entire dataset or focusing on localized information in each frame (Meng et al., 2022; Fayyaz et al., 2022; Kakogeorgiou et al., 2022; Yin et al., 2022) for efficient image recognition. However, they have no means to capture a temporal correlation across adjacent frames in video tasks. A few works recently suggest *supervised input selection* techniques for video tokens or frames. (Wang et al., 2022a) introduce a lightweight scorer network to select the most informative temporal and spatial token in incoming videos for supervised video classification tasks. (Park et al., 2022) suggest the greedy K-center search that iteratively finds the most distant patches in the geometrical space from video clips. (Gowda et al., 2021) train a single and global frame selector based on the ground truth for computing the importance score of single and paired frames. (Zhi et al., 2021) performs a frame selection for unsupervised videos, but they have to extract whole frames from the training video dataset in advance to compute the change of cumulative distribution in video frames and features, consuming substantial pre-processing time and storage for saving the extracted frames (it takes two days to extract SSv2 into frames and occupies 433 GB). Similarly, a few recently proposed MVA methods with adaptive token sampling require extracting dense motion information in advance or learnable parameters. MGM (Fan et al., 2023), MGMAE (Huang et al., 2023) and MME (Sun et al., 2023) generate motion-guided masking maps to reconstruct the informative tokens of the given videos, but they require motion vectors and optical flows, respectively.

However, extracting all video frames and computing their motion information (e.g., HoG or optical flows) in advance is unrealistic for online frame selection in videos. On the other hand, our EVEREST can perform rapidly without time- and memory-wasting steps before training.

3. Masked Video Autoencoders

3.1. Masked Video Autoencoder

Masked Image Modeling (MIM) (He et al., 2022; Xie et al., 2021) aims to learn representations of the input images by solving the regression problem in which a model predicts RGB pixel values in randomly masked patch regions of images. The recent success of MIM has led to breakthroughs in effectively capturing spatiotemporal information from incoming video streams, which we call masked video autoencoder (Tong et al., 2022; Feichtenhofer et al., 2022). Let $\boldsymbol{v} \in \mathbb{R}^{2\tau \times C \times H \times W}$ be a short clip consisting of 2τ successive frames from the entire video. A self-supervised model f be formulated into an encoder-decoder framework $f(\cdot) = D(E(\cdot))$ aims to reconstruct partially masked frames in $\boldsymbol{v}$, guided by spatial and temporal relationships between tokens in adjacent frames. The encoder takes tokenized embedding vectors from a pair of successive frames using a 3D convolution (Ma et al., 2022; Qing et al., 2023; Wang et al., 2023b). Let $\boldsymbol{k}_i$ be an i-th spatiotemporal embedding vector for a pair of two frames $\boldsymbol{v}[2i : 2i+1]$. Given $\boldsymbol{v}$, we reformulate the loss function as follows:

$$\begin{aligned} \ell(\boldsymbol{v}) = \sum_{i=0}^{\tau-1} \| D\left(E\left(\boldsymbol{m}_i \otimes \boldsymbol{k}_i; \mathcal{W}_E\right); \mathcal{W}_D\right) \\ -(\mathbf{1} - \boldsymbol{m}_i) \otimes \boldsymbol{v}[2i : 2i+1] \|_p, \end{aligned} \tag{1}$$

where $\boldsymbol{m} = G(J,\ \rho,\ \tau) \in \{0,1\}^{\tau \times J}$ and $G(\cdot)$ is a masking function depending on a specific policy, for example, random (or agnostic) (Tong et al., 2022) and space-only (or tube) (Feichtenhofer et al., 2022) masking techniques.

$J = \lfloor \frac{H}{s} \rfloor \cdot \lfloor \frac{W}{s} \rfloor$ denotes the number of patches per image and $\| \cdot \|_p$ denotes p-norm. $\mathcal{W}_E$ and $\mathcal{W}_D$ are a set of weights in the encoder and decoder, respectively. $\otimes$ is a dimensionality-preserving vector-matrix multiplication op-

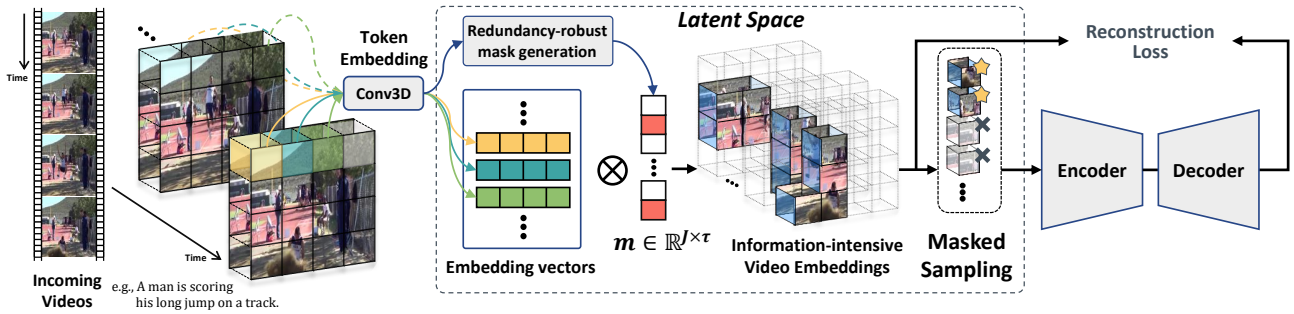


Figure 2: **Overview of EVEREST**. Our redundancy-robust mask generator selects tokens with a large disparity with the paired ones in the previous time dimension, indicating that they include rich motion features. Then, the model focuses on learning representation by reconstructing only sparsified videos containing abundant spatiotemporal information, which makes the VRL surprisingly efficient.

eration. A mask $\boldsymbol{m}_i \in \{0,1\}^J$ is drawn by the binary distribution B with a probability of ρ without replacement, that is $|\boldsymbol{m}_i| = [J \cdot \rho]$. After self-supervised pre-training to minimize Equation 1, the encoder transfers the learned representations to various downstream tasks. Unlike the samples from the image dataset, which are permutation-invariant as they are independent of each other, consecutive frames from the video stream inherently have a strong correlation and redundancy. Thus, masked video autoencoder can enjoy spatiotemporal inductive bias from other adjacent frames in the input clip, achieving good reconstruction quality even with lessened hints (i.e., a proportion of unmasked tokens). Indeed, MVA allows a much higher masking ratio $1 - \rho$ per video against MIM methods. This property is advantageous because masked modeling with a higher masking ratio significantly reduces computations when training the encoder-decoder framework.

3.2. Challenges in Masked Video Autoencoder

MVA methods (Tong et al., 2022; Feichtenhofer et al., 2022; Wang et al., 2023a) basically capture meaningful representations from pre-training videos via random masking strategies for input tokens, which are reasonable for curated and distributionally stable video datasets, yet, there is plenty of room for further development to make the model much more robust and computation-efficient. Here, we summarize two major **limitations** in the random sampling-based MVA:

(1) Patchified image tokens from a video clip are *not equally* important. At each iteration, MVA methods determine which tokens to mask according to specific random selection strategies (e.g., random, time-only, etc.). Yet, the relative amount of information in each token highly depends on the position of the informative objects and the correlation across patches within adjacent frames, which renders most of the tokens highly uninformative or redundant. These limitations lead to consuming massive training budgets in memory occupancy and suffering from slow convergence speed (e.g., training $3,200$ and $4,800$ epochs when using VideoMAE (Tong et al., 2022) on UCF101 and HMDB51, respectively). To address this information imbalance in visual tokens, several recent works have suggested sparsification/merging methods. EViT (Liang et al., 2022) is a supervised training method that fuses tokens by removing uninformative ones from the target task, therefore, inapplicable to video self-supervised learning. Token merging methods, including ToMe (Bolya et al., 2023), average multiple correlated/clustered tokens and let them have the same and marginalized values using average pooling. However, the masked video autoencoder encodes a few unmasked regions of video frames and aims to reconstruct raw frames. That is, token merging techniques are inappropriate for the decoding phase of masked autoencoder by design. Furthermore, the encoding phase with token merging may also fail to reconstruct raw frames at a pixel-level, due to their marginalized token features. This technical design is well-performed in the action recognition classification problem but deteriorates the performance of the VRL method.

(2) MVAs draw frames at *uniform time intervals* from video streams to train on. Real-world videos may include noisy and highly redundant frame sequences, often uninformative or even detrimental in representing temporal causal relationships and features of moving objects. However, the uniform sampling of video frames results in a waste of computational and memory resources in video understanding. To alleviate this problem, AdaFrame (Wu et al., 2019) trains the memory-augmented LSTM for gathering informative frames in given videos, leveraging supervised video labels to evaluate whether observing more frames is expected to produce more accurate predictions. SCSampler (Korbar et al., 2019) also requires the ground truth to jointly train the clip classifier and clip-level saliency model to obtain useful clips to the clip classifier. SMART (Gowda et al., 2021) trains

Single-frame Selector and Global Selector for the frame selection, which require the ground truth for computing the importance score of single and paired frames.

These challenges further exacerbate the problem when applying masked video autoencoder on *uncurated* first-person view real-world videos, *e.g.,* Ego4D (Grauman et al., 2022), which contains not only sparse motion information over space and time, but also suffers from a severe spatiotemporal redundancy. Therefore, determining which frames and spatiotemporal tokens to recover is crucial for practical and efficient video representation learning.

4. Efficient Masked Video Autoencoder by Removing Spatiotemporal Redundancy

4.1. Redundancy-robust Mask Generation

Valuable spatiotemporal information in video streams mainly comes from active visual changes rather than from static backgrounds. Thus, we aim to learn self-supervision on videos from only a few crucial regions containing minimal spatiotemporal redundancy. Let $\boldsymbol{k}_i$ be a token embedding of a pair of adjacent $2i^{th}$ and $(2i+1)^{th}$ frames from an input video clip $\boldsymbol{v} \in \mathbb{R}^{2\tau \times C \times H \times W}$, where $0 \leq i < \tau$. $\boldsymbol{k}_{i+1,j}$ indicates the jth token embedding of $\boldsymbol{k}_{i+1}$ and we measure the importance $I_{i+1,j}$ of $\boldsymbol{k}_{i+1,j}$ by computing the *distance* from the token embedding at the same region in the previous time step, $\boldsymbol{k}_{i,j}$:

$$I_{i+1,j} = S\left(\boldsymbol{k}_{i+1,j}, \boldsymbol{k}_{i,j}\right), \tag{2}$$

where $\boldsymbol{k}_i = \texttt{Conv3d}(\boldsymbol{v}[2i:2i+1]; \boldsymbol{w})$ and $S(\cdot,\cdot)$ indicates a distance function, such as Euclidian, negative Cosine Similarity, and negative Centered Kernel Alignment (Cortes et al., 2012). Throughout this paper, we use the ℓ_2 norm for S, which is simple yet empirically performs well, and we provide a discussion for the choice of the S in Appendix B. Tokens with a large disparity from tokens in a previous time period are considered important, indicating that they contain unique knowledge in the video and may have more important information than other tokens. Given $\boldsymbol{v}$, the model determines the token embedding vectors with the highest importance ratio of ρ_{pre} based on Eq. 2, which we call *Redundancy-robust (ReRo) Masking Generation*. We can drastically reduce the computational cost during training by **only propagating these selected embedding vectors** $\widetilde{\boldsymbol{k}}$ in a minibatch video clip at each iteration.

Next, we randomly sample a few token embeddings $\widetilde{\boldsymbol{k}}'$ from $\widetilde{\boldsymbol{k}}$ with the ratio of ρ_{post} to forward them to the encoder. And the corresponding decoder predicts RGB pixels on all embedding vectors $\widetilde{\boldsymbol{k}}_i \in \widetilde{\boldsymbol{k}}$ based on the encoder outputs. We set a high masking ratio in general, $\rho_{pre} \cdot \rho_{post} = 0.1$, to follow the settings of our MVA baselines. The objective

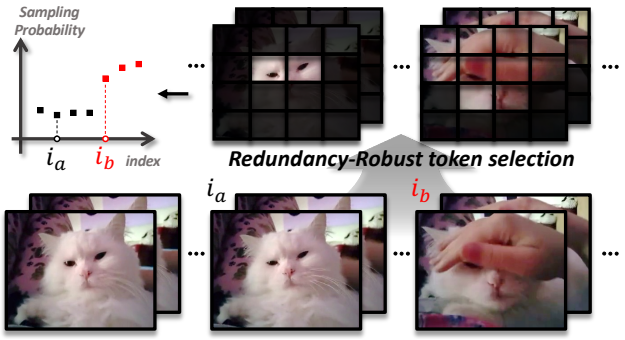


Figure 3: **Information-intensive frame selection.** We adaptively select frames based on the **ReRo** token frequency, which indicates significance compared to frames.

function of our EVEREST is formulated as follows:

$$arg\min_{\mathcal{W},\boldsymbol{w}} \sum_{n=1}^{N}\sum_{i=0}^{\tau-1} \left\| f_{\mathcal{W}}((\widetilde{\boldsymbol{k}}'_i)^n) - \boldsymbol{m}_i^n \otimes \boldsymbol{v}_n[2i:2i+1] \right\|_F \tag{3}$$

where $\boldsymbol{m}^n = G\left([J \cdot \rho_{pre}], \tau\right) \in \{0,1\}^{\tau \times J}$ and the size $|\boldsymbol{m}^n| = [J \cdot \tau \cdot \rho_{pre}]$. We omit the positional embedding term for notational simplicity. Whereas earlier MVA works used a fixed masking ratio per video cues for self-supervised training, our approach allows a dynamic masking rate for each frame by design, based on the occupancy of essential tokens. As shown in Fig. 7, each video frame contains a different amount of spatiotemporal information, and our dynamic masking strategy enables the model to focus on learning the valuable representation in a holistic view from incoming videos. We note that our EVEREST can be generalized well on both pre-training and fine-tuning phases with a negligible additional computational cost and training time (Please see Fig. 1). We find that removing unimportant tokens is crucial for both phases, obtaining more meaningful representations and achieving superior performances. (Please see Tab. 10). The overall process of our redundancy-robust token selection is illustrated in Fig. 2.

We emphasize that our ReRo mask generation is also significantly effective for motion-centric videos, e.g., Ego4D, since these *untrimmed real-world videos also contain a lot of temporally redundant visual information*. For example, in Fig. 8, a worker focuses on a grass mower to operate it well, and a person makes cookie dough, where visual scenes include much meaningless visual information, like empty space on the table.

4.2. On-the-fly Information-intensive Frame Selection

As discussed in Sec. 3.2, real-world video streams may contain many redundant frames and also often include non-useful intermediate clips, such as temporally glancing at uninspiring walls, grounds, or skies, that interfere with es-

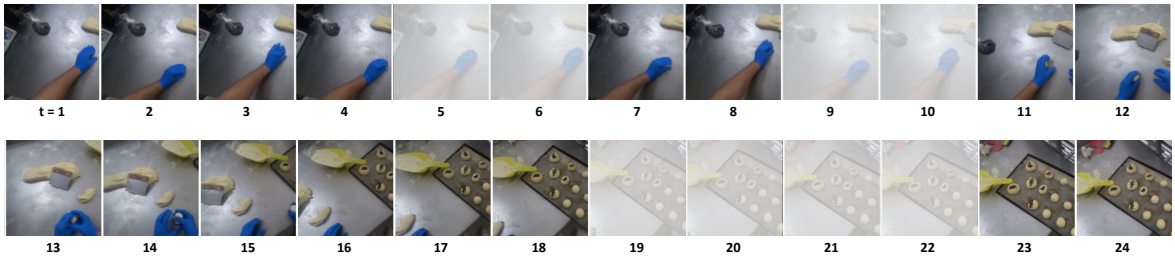


Figure 4: **Visualization of the proposed information-intensive frame selection on an uncurated dataset, Ego4D.** Unlike prior works that uniformly samples frames similar to each other, we adaptively sample the given video (24 frames) by probabilistic sampling the frames that have distinct spatiotemporal features (non-blurred frames).

timating a causality of the user's or cameraman's attention. However, most video-based learning methods follow a simple strategy to sample frames from uniform time intervals at each iteration. This approach is reasonable for well-curated video benchmark datasets (Kay et al., 2017; Soomro et al., 2012) containing only information-dense frames with fixed viewpoints, but often unsuitable for real-world videos, which are uncurated and more likely to contain redundant and overlapped frames.

To overcome the limitation, we propose to adaptively discard uninformative frames in the arrival video and build causal clips that represent the most crucial behaviors of the objects. We illustrate a simple overview of our *information-intensive frame selection* in Fig. 3 and visualize the sampled results in Fig. 4. We first select evenly spaced $[2\alpha \cdot \tau]$ frames as candidates, α times larger than the input clip size τ, where $\alpha > 1$. We count the number of chosen tokens c_i for the $2i^{th}$ and $(2i+1)^{th}$ frames based on the frequency of our *ReRo tokens*, described in Sec. 4.1. Then, the model iteratively trains on the input clip by drawing τ frames from $[\alpha \cdot \tau]$ candidates without replacement, with a probability of $\frac{c_i}{\sum_i c_i}$ that the paired $2i^{th}$ and $(2i+1)^{th}$ frames are drawn. The model trains video clips with a limited length at each iteration since longer clips require massive computations and memory. Therefore, we remark that information-intensive frame selection is crucial to better capture causality in the arrival video, as the model can observe longer video fragments while avoiding redundant frames.

5. Experiments

Experimental settings We validate our method on multiple video datasets: *UCF101* (Soomro et al., 2012), *HMDB51* (Kuehne et al., 2011), *Something-Something v2 (SSv2)* (Goyal et al., 2017), *Kinetics-400 (K400)* (Kay et al., 2017) and *Ego4D* (Grauman et al., 2022). We evaluate our information-intensive frame selection strategy during pre-training on *Object State Change Classification (OSCC)*task from Ego4D, containing raw and uncurated people's daily life videos. Given 8-second videos, OSCC classifies whether the object's state has changed by interacting with a camera wearer. Following the same training protocols as VideoMAE (Tong et al., 2022), we pre-train our EVEREST over benchmark datasets *without labels* and report the fine-tuning performance. For a fair comparison, we train both VideoMAE and our EVEREST using the same *one-node* equipped with 8 GPUs. For K400 and SSv2, although VideoMAE trained all models for 1600 and 2400 epochs, respectively, with multi-node GPUs (e.g., 64 V100 GPUs), we train both methods with several scaled ViT backbones for 200 epochs due to quick validation of the scalability. Except for K400, we use ViT-B as a backbone for the other benchmarks. Please see Appendix A for further implementation details.

EVEREST is memory and computationally efficient while achieving competitive performance against VideoMAE and K-centered variants. We extensively compare our proposed method against strong VRL baselines. Tab. 1 shows the results of the supervised and self-supervised methods on K400. Our redundancy-robust token selection, EVEREST, shows comparable performance to VideoMAE (Tong et al., 2022), while achieving *significant* computation and memory reduction during both pre-training and fine-tuning. Specifically, for ViT-S, EVEREST can reduce computational costs by 45.7% and 48.9%, respectively. It's also worth noting that EVEREST using ViT-L is more than $4\times$ more memory efficient than VideoMAE (e.g., 164.1 GB vs. 634.1 GB). Meanwhile, Tab. 2 shows that EVEREST achieves superior fine-tuning performance against recent VRL methods across UCF101 and HMDB. Specifically, EVEREST outperforms the best baseline, VideoMAE, by $2.1\%p \uparrow$ on UCF101 and $3.2\%p \uparrow$ on HMDB51. Also, we visualize the convergence plot of EVEREST on UCF101 in Fig. 5 Left. By pre-training at only 800 epochs, our EVEREST surpasses the fine-tuning accuracy of VideoMAE trained at 3200 epochs. That is, EVEREST reaches on-par performance with VideoMAE by using only **$\sim$14% of total computational costs**. We also against variants of a strong motion-based token selection method, K-centered patch sampling, with the modified vision transformer for video learning, including XViT(Bulat et al.,

Table 1: **Performance Comparison on K400.** PT and FT mean pre-training and fine-tuning, respectively. We use an input size of 16×224^2. Memory usage is measured with 256 batch size when pre-training. † is trained from random initialization. VideoMAE and EVEREST are pre-trained for 200 epochs, and we refer (Wightman, 2019) for computing ‡.

Method	Backbone	PT-Data	PT-GFLOPs	FT-GFLOPs	Memory usage (GB)	Top-1 Acc
MViT (Fan et al., 2021)†	MViT-S	✗	-	32.9	-	76.0
MViT (Fan et al., 2021)†	MViT-B	✗	-	70.5	-	78.4
ViViT FE (Arnab et al., 2021)	ViT-L	IN-21K	119.0‡	3980.0	N/A	81.7
K-centered (Park et al., 2022)	XViT	IN-1K	67.4‡	425.0	N/A	73.1
K-centered (Park et al., 2022)	Mformer	IN-1K	67.4‡	369.5	N/A	74.9
K-centered (Park et al., 2022)	TSformer	IN-1K	67.4‡	590.0	N/A	78.0
VideoMAE (Tong et al., 2022)	ViT-S	K400	11.6	57.0	117.4	73.5
VideoMAE (Tong et al., 2022)	ViT-B	K400	35.5	180.5	486.4	78.4
VideoMAE (Tong et al., 2022)	ViT-L	K400	83.1	597.2	634.1	82.0
EVEREST (Ours)	ViT-S	K400	**6.3** (↓ 45.7%)	**29.1** (↓ 48.9%)	**59.9** (↓ 49.0%)	**75.9**
EVEREST (Ours)	ViT-B	K400	**21.5** (↓ 39.4%)	**98.1** (↓ 45.7%)	**91.2** (↓ 81.3%)	**79.2**
EVEREST (Ours)	ViT-L	K400	**60.8** (↓ 26.8%)	**330.0** (↓ 44.7%)	**164.1**(↓ 74.1%)	**82.1**

Table 2: **Comparison with strong baselines on UCF101, HMDB51, and SSv2 datasets** without using the pre-training step on a large-scale dataset. Several results are drawn from Diba et al. (2021); Tong et al. (2022). *SR50* indicates *SlowOnly-R50*.

Method	Backbone	Extra data	T1 Acc (%) UCF101	HMDB51	SSv2
VCOP (Xu et al., 2019)	R(2+1)D	UCF101	72.4	30.9	N/A
CoCLR (Han et al., 2020)	S3D-G	UCF101	81.4	52.1	N/A
Vi^2 CLR(Diba et al., 2021)	S3D	UCF101	82.8	52.9	N/A
CoCLR (Han et al., 2020)	S3D-G	K400	87.9	54.6	N/A
Vi^2 CLR(Diba et al., 2021)	S3D	K400	89.1	55.7	N/A
RSPNet (Chen et al., 2021)	S3D-G	K400	93.7	64.7	55.0
ρSwAV$_{\rho=2}$ (Feich. et al., 2021)	SR50	K400	87.3	N/A	51.7
ρMoCo$_{\rho=2}$ (Feich. et al., 2021)	SR50	K400	91.0	N/A	54.4
ρBYOL$_{\rho=2}$ (Feich. et al., 2021)	SR50	K400	92.7	N/A	55.8
VideoMAE (Tong et al., 2022)	ViT-B	✗	91.3	62.6	64.3
EVEREST (Ours)	ViT-B	✗	**93.4**	**68.1**	**64.6**

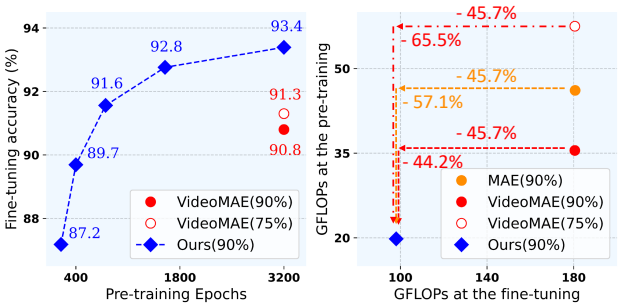


Figure 5: **Performance & GFLOPs Comparison** on UCF101 dataset. **(Left)** EVEREST outperforms VideoMAE for both masking ratios (75% and 90%) even at significantly fewer training epochs. **(Right)** Our EVEREST reduces GFLOPs during pre-training and fine-tuning compared to VideoMAE and ST-MAE.

2021), Mformer (Patrick et al., 2021), and TSformer (Bertasius et al., 2021). Our proposed method also surpasses these motion-based video understanding methods by significant margins in terms of Top-1 accuracy and GFLOPs.

EVEREST is highly beneficial for model deployment. To validate the efficiency of our EVEREST in terms of memory usage, we measure the memory consumption over multiple batch sizes and architecture scales during the pre-training phase on K400. We use one node equipped with 8×A100 (80GB) GPUs and compare EVEREST with VideoMAE. As shown in Tab. 3, VideoMAE shows ***out-of-memory*** when using a batch size of 512 and 1,024. Regarding that VideoMAE used the batch size of 1,024 in the original paper (Tong et al., 2022), VideoMAE inevitably has to lower the batch size to run training in a one-node environment, resulting in increased training time. However, the proposed redundancy-robust token selection approach allows to drastically reduce the memory usage, which means that it can be trained using only a one-node environment. We further compare pre-training and fine-tuning budgets with those of the state-of-the-art MVAs such as MME (Sun et al., 2023) and MVD (Wang et al., 2023c). In Tab. 4, when pre-training, EVEREST requires only 44%, 24%, and 55% of memory than VideoMAE, MME and MVD, respectively. Due to the generality of our EVEREST, we can apply EVEREST to the finetuning phase by prunning redundant tokens. As shown in Tab. 5, compared to other methods using full tokens, our EVEREST reducing computation and memory requirements while acheiving comparable accuracy.

EVEREST is also significantly rapid for pre-training compared to SoTA VMA baselines. To evaluate the wall-clock time efficiency, we compare pre-training time (PT-Time) with state-of-the-art MVAs by using the same batch size on K400. As shown in Tab. 4, EVEREST requires

Table 3: **Memory usage comparison** during the pre-training (one node, A100×8 GPUs) on K400. The memory gaps grow up when increasing batch size and model size. *OOM* indicates *out-of-memory*.

Method	Backbone	Effective Batch Size			
		128	256	512	1024
VideoMAE	ViT-S	71.9	117.4	233.6	443.6
	ViT-B	258.4	486.4	*OOM*	*OOM*
	ViT-L	353.9	634.1	*OOM*	*OOM*
EVEREST (Ours)	ViT-S	**31.9**	**59.9**	**95.6**	**173.1**
	ViT-B	**60.0**	**91.2**	**151.2**	**273.6**
	ViT-L	**117.3**	**164.1**	**258.9**	**428.4**

Table 4: **Pre-training time & Memory comparison with SoTA MVAs on K400.** We measure the pre-training time for an epoch under a single-node machine equipped with 8×A6000 (48GB) GPUs. We use ViT-B and a batch size of 128. Note that MME (Sun et al., 2023) should pre-compute the optical flow for the entire video data before pre-training. We exclude the time and memory of MME for optical flow computations in the table.

Method	PT-Time	Memory
VideoMAE (Tong et al., 2022)	18m 42s	150.3 GB
MME (Sun et al., 2023)	10m 15s	121.2 GB
MVD (Wang et al., 2023c)	51m 55s	274.9 GB
EVEREST (Ours)	**8m 18s**	**66.3** GB

Table 5: **EVEREST-Finetuning with other MVAs on K400.** We apply our EVEREST for finetuning the pre-trained models (ViT-B) by SoTA MVAs. We measure the memory usage with the same batch size of 128. While SoTA methods use full tokens during fine-tuning, EVEREST uses only redundancy-robust tokens.

PT-Method	FT-Method	GFLOPs	Memory	Top-1
VideoMAE	Full-token	180.5	362.5 GB	81.5
VideoMAE	**EVEREST**	98.1	178.4 GB	81.6
MME	Full-token	180.5	362.5 GB	81.8
MME	**EVEREST**	98.1	178.4 GB	82.0
MVD	Full-token	180.5	362.5 GB	83.4
MVD	**EVEREST**	98.1	178.4 GB	82.8

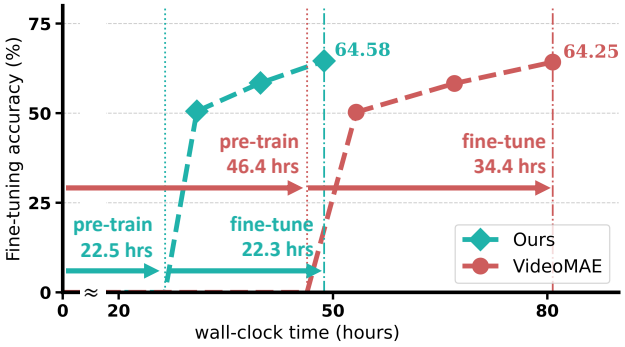


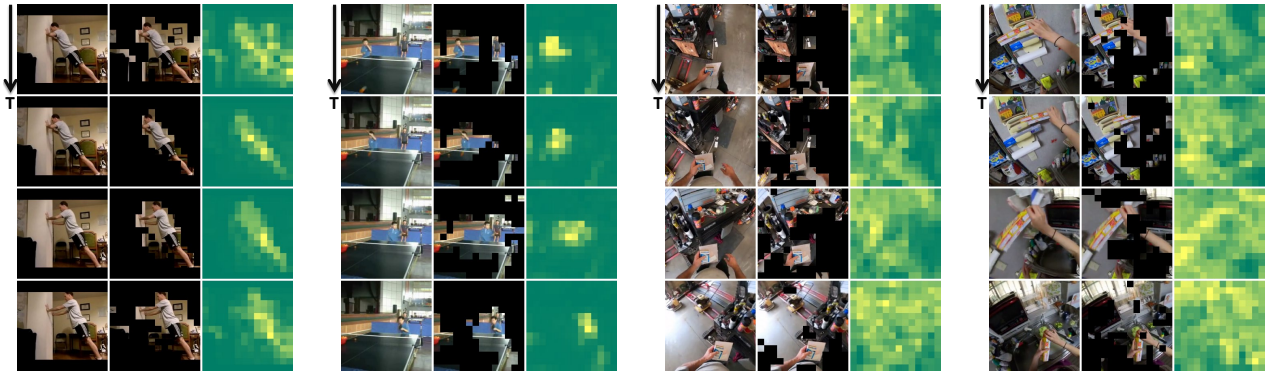
Figure 6: **Accuracy over training time** on SSv2. 4 NVIDIA RTX 3090 GPUs are used. We set ρ_{pre} to 0.3 and 0.8 for pre-train and fine-tuning, respectively.

only 45%, 16%, and 81% of pre-training time than VideoMAE, MME, and MVD, respectively. Note that we didn't reflect the heavy computation burdens of MME for extracting HOG and optical flow of all input video data before training. Similarly, the whole training time, including both pre-training and fine-tuning, can be drastically reduced, as shown in Fig. 6. In short, besides a gain in memory efficiency, we can maximally enjoy remarkably accelerated pre-training and fine-tuning in resource-constrained environments compared to existing MVA methods.

EVEREST has a potential for estimating motion information. Our method is also effective for egocentric videos, e.g., Ego4D, which capture dynamic views from camera motion. EVEREST captures informative objects even with the rapid change of view in first-person videos, as it masks out objects crucial to understanding the camera wearer's attention while not attending to the background. Specifically, we provide masked input examples with their importance heatmaps through our proposed redundancy-robust masking in Fig. 7. Our masking strategy enables the model to capture the most informative tokens containing static and dynamic moving of objects (e.g., *doing wall push up* (Fig. 7 (a)) and *playing ping pong* (Fig. 7 (b)). In addition, it captures informative objects even containing drastic viewpoint changes in videos, as it masks objects crucial for understanding the camera wearers' attention (Fig. 7 (c)), while not focusing on backgrounds such as walls and floors.

Furthermore, it is worth noting that egocentric videos contain temporally redundant visual information. Fig. 8 reveals that a worker focuses on a grass mower to operate it well, and a person makes cookie dough, where visual scenes include much meaningless visual information, such as empty space on the table. Our method automatically discards these noisy and redundant tokens in an online manner without requiring additional dense information, such as optical flows of incoming videos.

These abundant redundant scenes over a long time in real-world videos market VRL models to waste time and computations by learning on meaningless frames, which also can lead to poor local optima due to bias and catastrophic forgetting. To overcome the limitation, we adopt the redundancy robust sampling rate α to focus on frames with larger motions across frames (See Fig. 4). By constructing



(a) fixed and static **(b) fixed and dynamic** **(c) first-view egocentric videos**

Figure 7: **Examples of Redundancy-Robust (ReRo) Token Selection.** We visualize the raw video frames (left), Redundancy-Robust masking (middle), and obtained importance heatmaps (right) on UCF101 (**(a)** and **(b)**, ρ_{pre}=0.3) and Ego4D (**(c)**, ρ_{pre}=0.5). Our ReRo token selection successfully captures narrow/concentrated motion information **(a, b)** and distributed/multi-object motion cues **(c)**.

Table 6: **Comparison with public SoTA methods for OSCC on Ego4D val set.** We set our frame selection ratio $\alpha = 1.5$. Pre-trained data modalities from the OSCC dataset 'V' and 'T' refers to visual and text, respectively.

Method	Modality	Accuracy (%)
Egocentric VLP (Lin et al., 2022)	V+T	73.9
SViT (Escobar et al., 2022)	V	69.8
TarHeels (Islam & Bertasius, 2022)	V	70.8
EVEREST (Ours)	V	**76.2** (+5.4%p)

the given videos with fluent motion information, our model focuses more on learning the core part of the video. We set the default α to 1.5 and report the effect of α in Appendix C. We quantitatively compare our method with recent SoTA methods using the OSCC task on Ego4D in Tab. 6. We pre-train OSCC without labels and fine-tune it to classify whether the object's state has changed. With only visual information, we outperform the previous SoTA method, Egocentric VLP (Lin et al., 2022), which uses visual and text information, by 2.3%p ↑.

6. Conclusion

From the insight that not all video tokens are equally informative, we propose a simple yet efficient parameter-free token and frame selection method for video pre-training. We select redundancy-robust tokens based on their significance in the change of their states and train the model only with them, drastically reducing memory allocations and computations. In addition, we propose a frame selection technique to construct input video data by sampling frames with a probability proportional to the degree of their occupancy of selected tokens. This is useful for video representation learning with uncurated videos containing a number of redundant frames. We empirically demonstrate that our method is significantly more efficient in computations, memory, and training time than strong baselines. We believe our method could help democritizing (Seger et al., 2023) video-related research requiring immense computation budgets.

Impact Statement

Our work has a positive broader societal impact leading to environment-friendly AI research by drastically reducing energy and carbon costs in the training and inference phases. Current methods in the field of video understanding consume enormous monetary and environmental costs, and we believe that our work contributes to overcoming this limited feasibility of video-based research fields by enabling single-node training with larger batch sizes using a large backbone, whereas existing VRL methods require immense memory occupancy and suffer from deploying models in the same setup due to *out-of-memory*. We also aim to enhance the accessibility of those cutting-edge video models by employing practical resources on a single node and pursuing eco-friendly AI research.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.2019-0-00075 Artificial Intelligence Graduate School Program (KAIST)), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-00187238, Development of Large Korean Language Model Technology for Efficient Pre-

training), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NO. RS-2023-00256259), KAIST-NAVER Hypercreative AI Center, Google Research Grant, and Google Cloud Research Credits program with the award (KX3L-GHDM-4U6F-QLE0).

References

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., and Schmid, C. Vivit: A video vision transformer. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Bao, H., Dong, L., and Wei, F. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

Bertasius, G., Wang, H., and Torresani, L. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your vit but faster, 2023.

Bulat, A., Perez Rua, J. M., Sudhakaran, S., Martinez, B., and Tzimiropoulos, G. Space-time mixing attention for video transformer. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Chen, P., Huang, D., He, D., Long, X., Zeng, R., Wen, S., Tan, M., and Gan, C. Rspnet: Relative speed perception for unsupervised video representation learning. In *Proceedings of the AAAI National Conference on Artificial Intelligence (AAAI)*, 2021.

Cortes, C., Mohri, M., and Rostamizadeh, A. Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 2012.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Diba, A., Sharma, V., Safdari, R., Lotfi, D., Sarfraz, S., Stiefelhagen, R., and Van Gool, L. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Escobar, M., Daza, L., González, C., Pont-Tuset, J., and Arbeláez, P. Video swin transformers for egocentric video understanding@ ego4d challenges 2022. *arXiv preprint arXiv:2207.11329*, 2022.

Fan, D., Wang, J., Liao, S., Zhu, Y., Bhat, V., Santos-Villalobos, H., MV, R., and Li, X. Motion-guided masking for spatiotemporal representation learning. *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., and Feichtenhofer, C. Multiscale vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Fayyaz, M., Kouhpayegani, S. A., Jafari, F. R., Sommerlade, E., Joze, H. R. V., Pirsiavash, H., and Gall, J. Ats: Adaptive token sampling for efficient vision transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Feich., C., Fan, H., Xiong, B., Girshick, R., and He, K. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Feichtenhofer, C., Fan, H., Li, Y., and He, K. Masked autoencoders as spatiotemporal learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Fernando, B., Bilen, H., Gavves, E., and Gould, S. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Gowda, S. N., Rohrbach, M., and Sevilla-Lara, L. Smart frame selection for action recognition. In *Proceedings of the AAAI National Conference on Artificial Intelligence (AAAI)*, 2021.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Han, T., Xie, W., and Zisserman, A. Self-supervised co-training for video representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Huang, B., Zhao, Z., Zhang, G., Qiao, Y., and Wang, L. Mgmae: Motion guided masking for video masked autoencoding. *arXiv preprint arXiv:2308.10794*, 2023.

Islam, M. M. and Bertasius, G. Object state change classification in egocentric videos using the divided space-time attention mechanism. *arXiv preprint arXiv:2207.11814*, 2022.

Kakogeorgiou, I., Gidaris, S., Psomas, B., Avrithis, Y., Bursuc, A., Karantzalos, K., and Komodakis, N. What to hide from your students: Attention-guided masked image modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Korbar, B., Tran, D., and Torresani, L. Scsampler: Sampling salient clips from video for efficient action recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. Hmdb: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., and Xie, P. Not all patches are what you need: Expediting vision transformers via token reorganizations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Lin, K. Q., Wang, A. J., Soldan, M., Wray, M., Yan, R., Xu, E. Z., Gao, D., Tu, R., Zhao, W., Kong, W., et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., and Hu, H. Video swin transformer. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ma, Y., Yang, T., Shan, Y., and Li, X. Simvtp: Simple video text pre-training with masked autoencoders. *arXiv preprint arXiv:2212.03490*, 2022.

Meng, L., Li, H., Chen, B.-C., Lan, S., Wu, Z., Jiang, Y.-G., and Lim, S.-N. Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Pan, T., Song, Y., Yang, T., Jiang, W., and Liu, W. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Park, S. H., Tack, J., Heo, B., Ha, J.-W., and Shin, J. K-centered patch sampling for efficient video recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Patrick, M., Campbell, D., Asano, Y. M., Metze, I. M. F., Feichtenhofer, C., Vedaldi, A., and Henriques, J. F. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Piergiovanni, A., Angelova, A., and Ryoo, M. S. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Qian, R., Meng, T., Gong, B., Yang, M.-H., Wang, H., Belongie, S., and Cui, Y. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Qing, Z., Zhang, S., Huang, Z., Wang, X., Wang, Y., Lv, Y., Gao, C., and Sang, N. Mar: Masked autoencoders for efficient action recognition. *IEEE Transactions on Multimedia*, 2023.

Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., and Dafoe, A. Democratising ai: Multiple meanings, goals, and methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 715–722, 2023.

Soomro, K., Zamir, A. R., and Shah, M. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Sun, X., Chen, P., Chen, L., Li, C., Li, T. H., Tan, M., and Gan, C. Masked motion encoding for self-supervised video representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2235–2245, 2023.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Tong, Z., Song, Y., Wang, J., and Wang, L. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Wang, J., Yang, X., Li, H., Li, L., Wu, Z., and Jiang, Y.-G. Efficient video transformers with spatial-temporal token selection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022a.

Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., and Qiao, Y. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14560, 2023a.

Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.-G., Zhou, L., and Yuan, L. Bevt: Bert pre-training of video transformers. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., and Jiang, Y.-G. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023b.

Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Yuan, L., and Jiang, Y.-G. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6312–6322, 2023c.

Wightman, R. Pytorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Wu, Z., Xiong, C., Ma, C.-Y., Socher, R., and Davis, L. S. Adaframe: Adaptive frame selection for fast video recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. *arXiv preprint arXiv:2111.09886*, 2021.

Xu, D., Xiao, J., Zhao, Z., Shao, J., Xie, D., and Zhuang, Y. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Yin, H., Vahdat, A., Alvarez, J. M., Mallya, A., Kautz, J., and Molchanov, P. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yoon, J., Madaan, D., Yang, E., and Hwang, S. J. Online coreset selection for rehearsal-based continual learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=f9D-5WNG4Nv.

Zhi, Y., Tong, Z., Wang, L., and Wu, G. Mgsampler: An explainable sampling strategy for video action recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Appendix

Organization The supplementary file is organized as follows: Firstly, we explain the implementation details according to tasks in Appendix A and perform the distance function ablation study for redundancy-robust masking during pre-training in Appendix B. We provide additional experimental results for our EVEREST in Appendix C. We visualize more examples from EVEREST in Appendix D.

A. Implementation Details

Table 7: **Pre-training settings for K400, SSv2, UCF101, HMDB51 and OSCC.**

	K400	SSv2	UCF101	HDMB51	OSCC
optimizer	AdamW				
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$ (Chen et al., 2020)				
base learning rate	1.5e-4(S,L), 3e-4(B)	1.5e-4	1e-3	1e-3	1e-3
weight decay	0.05				
learning rate schedule	Cosine decay (Loshchilov & Hutter, 2016)				
flip augmentation	yes	no	yes	yes	yes
augmentation	MultiScaleCrop				
batch size	1024(S,B), 512(L)	256	192	192	256
warmup epochs	40	40	40	40	20
ρ_{pre}	0.3	0.3	0.3	0.3	0.5
sampling stride	4	2	4	2	4
total epochs	200	200	3200	4800	400

Table 8: **Fine-tuning settings for K400, SSv2, UCF101, HMDB51 and OSCC.**

	K400	SSv2	UCF101	HDMB51	OSCC
optimizer	AdamW				
optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$				
weight decay	0.05				
learning rate schedule	Cosine decay				
warmup epochs	5				
laer-wise lr decay	0.75 (Bao et al., 2021)				
flip augmentation	yes	no	yes	yes	yes
RandAug	$(9, 0.5)$ (Cubuk et al., 2020)				
label smoothing	0.1 (Szegedy et al., 2016)				
drop path	0.1				
base learning rate	5e-4(S), 1e-3(B), 2e-3(L)	5e-4	1e-3	1e-3	1e-4
batch size	384(S,B), 128(L)	48	128	128	32
ρ_{pre}	0.6	0.8	0.6	0.6	-
sampling stride	4	-	4	2	10
total epochs	75(S), 35(B,L)	50	100	50	30
multi-view	2×3	2×3	5×3	10×3	2×3

We validate our EVEREST on five video datasets: *K400* (Kay et al., 2017), *SSv2* (Goyal et al., 2017), *UCF101* (Soomro et al., 2012), *HMDB51* (Kuehne et al., 2011), and *Ego4D* (Grauman et al., 2022). We provide the hyperparameter setup for pre-training and fine-tuning in Tab. 7 and Tab. 8, respectively. During fine-tuning, we follow segment-based sampling (Wang et al., 2018) on SSv2. As we mentioned in Sec. 4.1 of the main paper, we adopt the same masking ratio with VideoMAE (Tong et al., 2022) and ST-MAE (Feichtenhofer et al., 2022) for masking input video, but we only reconstruct $(\rho_{pre} \times 100)\%$ of input video tokens during pre-training. And in fine-tuning, our model only takes $(\rho_{pre} \times 100)\%$ of input video tokens. We follow the linear learning rate scheduling of (Tong et al., 2022) and (Feichtenhofer et al., 2022), $lr = base_learning_rate \times batch_size/256$. Overall implementation of our method is built upon VideoMAE[1](Tong et al., 2022).

Video Action Recognition We extensively evaluated our method on multiple benchmark datasets for the video action recognition task, including K400, SSv2, UCF101, and HMDB51. For inference, we adapt common multi-view testing with T clips × three crops. That is, the model takes T temporal clips with three spatial crops to cover the overall length and space of the video. Then, we measure the average performance of all views.

Object State Change Classification (OSCC) The OSCC dataset is the subset of the Ego4d dataset, consisting of 41.1k/21.2k train/val 8-second videos. Note that, as the Ego4D dataset shows the characteristic of having motion cues that are distributed and containing multi objects in Fig. 7 (c) of the main paper, we adopt $\rho_{pre} = 0.5$. We train 400 and 30 epochs with a fixed sampling ratio of 4 and 10 in the pre-training and fine-tuning phases, respectively. We use $\alpha = 1.5$ only in pre-training.

B. Comparing Functions for Computing Token Embedding Distance

Table 9: Fine-tuning results on HMDB51 measured by varying the distance function in pre-training.

Distance Function	accuracy(%)
negative cosine	33.53
negative CKA	34.58
L1	42.22
L2	**42.81**

We analyze the effect of adopting different distance functions: L1 and L2 distance, negative cosine similarity, and negative CKA (Cortes et al., 2012). We train ViT-B on

[1] https://github.com/MCG-NJU/VideoMAE

Table 10: **The ratio of our ReRo Masking at pre-training (Left) and fine-tuning (Right)** on UCF101. We highlight the default redundancy-robust masking rate (ρ_{pre}) as red texts. We basically pre-train our model 800 epochs, but also report the results with 3,200 pre-training epochs following VideoMAE, denoted by †. * and ** denotes the results with 75% and 90% masking, respectively.

Method	ρ_{pre}	top-1	GFLOPs
MAE	-	-	46.10
VideoMAE†	-	91.25*	57.50
	-	90.80**	35.48
Ours	0.5	90.91	23.43
	0.4	91.51	21.54
	0.3	**91.56**	19.81
	0.2	90.80	**18.22**
Ours†	0.3	**93.39**	19.81

Method	ρ_{pre}	top-1	GFLOPs
MAE	-	-	180.6
VideoMAE†	-	91.25	180.5
Ours	1.0	89.71	180.5
	0.8	90.17	137.5
	0.7	91.25	117.3
	0.6	**91.56**	98.1
	0.5	91.48	**79.8**
Ours†	0.6	**93.39**	98.1

HMDB51 for 200 and 50 epochs, respectively, only varying the distance function in pre-training as shown in Table 9. As shown, L1 and L2 outperform the alternatives. Therefore, we default to the L2 distance to measure the disparity between token embeddings in temporally adjacent frames.

C. Additional Experimental Results

Ablation study on redundancy-robust token selection One of our major contributions is the significantly enhanced computational efficiency during VRL, as we process a few latent vectors in the decoder to reconstruct only the motion-activated tokens in the given videos according to the redundancy-robust masking ratio ρ_{pre} (Please see Sec. 4.1). We set ρ_{pre} to 0.3 so that our EVEREST reconstructs the 30% of the essential spatiotemporal regions focusing on objects' movements and behaviors, from a sparsified video clip, which reduces 65.5% of GFLOPs at the pre-training phase (Please see Fig. 5 Right and Tab. 10 Left). Additionally, as reported in Tab. 10 Right, our approach can reduce the computational overhead at the fine-tuning phase, which is practically useful when transferring the learned representation learning model to downstream video tasks.

EVEREST fine-tuning using the pre-trained VideoMAE on K400. We also show the transferability of our redundancy-robust token selection by fine-tuning a pre-trained model with a different method, VideoMAE, on the K400 dataset. We borrow the pre-trained weights in the official repository[2]. Tab. 11 shows the results with different ViT backbones. Even using publicly available pre-trained models from VideoMAE, our EVEREST show competitive fine-tuning performance compared to the baseline while significantly reducing the computational cost. These results demonstrate that EVEREST can seamlessly adopt an otherwise pre-trained backbone.

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

Table 11: **Fine-tuning performance of EVEREST on K400 using the VideoMAE pre-trained model**. We adopt our redundancy-robust token selection method for fine-tuning K400. We use the pre-trained weights by VideoMAE for 1600 epochs and Vanilla indicates a standard fine-tuning of the video action recognition task leveraging all visual tokens.

Pre-training	Fine-tuning	Backbone	GFLOPs	Acc. (%)
VideoMAE	Vanilla	ViT-S	57.0	**79.0**
	+EVEREST	ViT-S	**29.1** (↓ 48.9%)	**78.8**
	Vanilla	ViT-B	180.5	**81.5**
	+EVEREST	ViT-B	**98.1** (↓ 45.7%)	**81.6**
	Vanilla	ViT-L	597.2	**85.2**
	+EVEREST	ViT-L	**330.0** (↓ 44.7%)	**84.8**

Table 12: Fine-tuning performance on UCF101 measured by a varying number of frames.

Method	# of frames		GFLOPs		Acc. (%)
	PT	FT	PT	FT	
VideoMAE	16	16	35.48	180.5	90.80
EVEREST	16	16	**19.81**	**98.1**	**93.39**
EVEREST	**24**	**24**	**30.65**	**159.4**	**94.42**

Increasing frame lengths from reduced memory of our EVEREST As shown in Fig. 1 in the main paper, our EVEREST drastically saves the computational cost in pre-training and fine-tuning. In Tab. 12, we analyze to compensate for the reduced memory occupancy by increasing the number of input frames from 16 to 24 in pre-training and fine-tuning to observe the results in case of having similar computational costs with VideoMAE. We achieve 94.42% accuracy on UCF101 while still having relatively lower GFLOPs than VideoMAE.

Table 13: Ablation study of on-the-fly information-intensive frame selection of EVEREST.

α	**Acc. (%)**
-	73.17
1.3	73.54
1.5	**73.85**
1.8	73.79
2.0	73.80

Table 14: Transferability comparison between VideoMAE and EVEREST on K400 and SSv2 to UCF101 and HMDB51 with ViT-S and ViT-B.

Method	backbone	Pre-train Dataset	Top1 Acc.(%) UCF101	HMDB51
VideoMAE	ViT-S	K400	84.2	54.2
EVEREST	ViT-S	K400	**89.2** (↑ 5.0%)	**61.4** (↑ 7.2%)
VideoMAE	ViT-B	SSv2	88.7	60.9
EVEREST	ViT-B	SSv2	**92.2** (↑ 3.5%)	**64.6** (↑ 3.7%)

The ratio of information-intensive frame selection We conduct an ablation study for the rate of information-intensive frame sampling α. We pre-train our model 100 epochs on the OSCC task by varying α from *not using (-)* to *2.0*. Interestingly, as shown in Tab. 13, the performance with α outperforms baselines (-). And we default $\alpha = 1.5$ for experiments, which shows the best performance than others.

Transferability of our EVEREST We also measure the fine-tuning performance on UCF101 and HMDB51 using larger pre-training datasets in Tab. 14. We perform fine-tuning experiments with 200 epochs of pre-trained models from K400 and SSv2 in Tab. 1 and Tab. 2. Impressively, our method gains a substantial performance enhancement over all experiments compared with VideoMAE.

Table 15: **The impact of the number of Conv3d layers for capturing Redundancy-Robust Token Selection.**

# of Conv3d layers	layer1 kernel size	layer2 kernel size	layer3 kernel size	# of params. (M)	Memory usage (GB)	GFLOPs	FT Acc. (%)
1	$2\times16\times16$	N/A	N/A	**94.2 / 86.3**	**9.2** / 17.6	**19.8 / 98.1**	**68.1**
2	$2\times4\times4$	$1\times4\times4$	N/A	102.5 / 94.6	11.9 / **17.5**	34.6 / 112.9	66.7
3	$2\times4\times4$	$1\times2\times2$	$1\times2\times2$	97.8 / 89.9	11.3 / **17.5**	38.3 / 116.6	66.8

The design of the embedding layer We perform an ablation study of our proposed EVEREST regarding the embedding layer design. In our default setting, we use a single Conv3d layer. And we adopt deeper neural networks for spatiotemporal embedding in this experiment. Note that we adjust the kernel size of each layer to maintain the output dimension. As shown in Tab. 15, stacking more layers for input embedding achieves lower performance than the default setting (i.e., # of Conv3D layer = 1) on the HMDB51 dataset. Even they require substantial additional training costs in terms of the number of parameters, GFLOPs, and memory usage in most cases of pre-training and fine-tuning.

Table 16: **The impact of informative-token selection.** It shows better performances both in pre-training and fine-tuning when the model prioritizes selection with the most far-distance (descending) embedded tokens rather than near-distance (ascending) tokens.

Method	ReRo Masking Pre-train	Fine-tune	Fine-tuning accuracy (%)
EVEREST	ascending	ascending	60.93
	ascending	descending	73.49
	descending	ascending	75.60
	descending	descending	**91.56**

The importance of selecting tokens containing rich motion information We select to learn temporally changing tokens based on the distance between token embeddings in adjacent frames, thereby, the tokens farther away from adjacent frames contain less redundant information. As shown in Tab. 16, when we reversely select near-distance tokens in pre-training and fine-tuning, it severely decreases the accuracy, and the performance is the worst if the model selects tokens via reverse strategies in both phases.

Table 17: Comparison with running-cell masking strategy.

Method	UCF101	HDMB51
Running-cell	91.0	61.4
EVEREST	**93.4**	**68.1**

Comparison with running-cell masking strategy We compare our method with running-cell masking (Qing et al., 2023), which is used in VideoMAE-V2 (Wang et al., 2023a). We pre-trained with these masking strategies and performed conventional full-finetuning. As shown in Tab. 17, our method clearly outperforms running-cell masking, demonstrating our effectiveness against other selection strategies given the same architecture. We further show clear limitations in running-cell masking that VideoMAE-V2 suffers from visual redundancy and cannot capture semantic importance (Please see Fig. 8).

Next, We emphasize that redundancy-robust masking is applicable not only in the pre-training phase, but also in performing video-based downstream tasks. We process only 60% of tokens ($\rho_{pre} = 0.6$) of the given video during fine-tuning. Surprisingly, our redundancy-robust masking for the downstream tasks gains increased performance than our variant without masking on the fine-tuning tasks ($\rho_{pre} = 1.0$) using only about 55% of GFLOPs, as shown in Tab. 10 Right. The results support our hypothesis that video data often contain redundant information and our selective video learning using the proposed masking strategy successfully captures

essential space-time regions in video inputs, allowing us to focus more on learning spatiotemporally meaningful features. The fine-tuning masking ratio is higher than the ratio at the pre-training stage, showing that the model uses more information to fully exploit the task-relevant cues from the given videos than pre-training, which aims to obtain general information.

D. Visualization of Redundancy-Robust Masking and Sampling

As shown in Fig. 7 of the main paper, our masking method successfully captures the most informative space of each frame in various conditions. In Fig. 8, we visualize masking strategies of other strong baselines, ST-MAE (Feichtenhofer et al., 2022) and VideoMAE (Tong et al., 2022). While these two methods sample masked tokens based on random probability without taking into account semantic motion information, our proposed method successfully captures temporally significant localized regions in the video. We further visualize our masking results in Fig. 9 for deeper understanding. The upper sample is similar in Fig. 7 (b) but changes the ρ_{pre} from 0.5 to 0.15. The masked result shows that our masking strategy could be better and save more computational costs if the view is fixed and the moving object is few. The bottom sample show that our masking strategy wrongly captured the blue line of each frame as informative tokens when ρ_{pre} is relatively large(=0.5). However, when ρ_{pre} is relatively small(=0.25, 0.15), it ignores the blue line of each frame and concentrates more on moving objects in the given video. In Fig. 10, We show selected frames via our adaptive frame sampling. Our method uses redundancy-robust token selection to draw frames to learn in an online manner based on the proportion that contains important tokens containing fluent motion information or significant change of temporal states. This strategy allows the model to take diverse frames from given videos while discarding redundant or low-information frames.

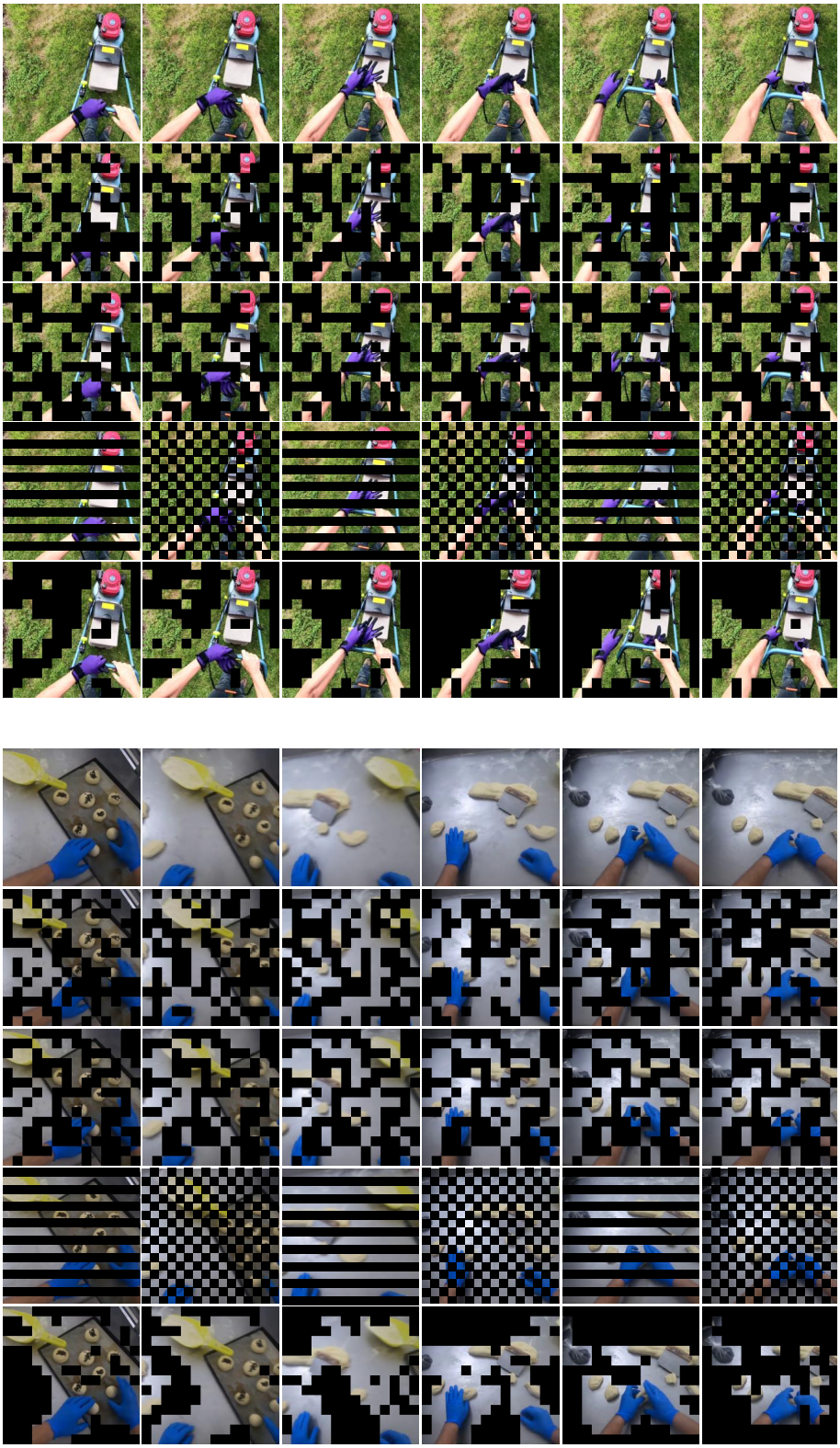


Figure 8: **Examples of masking strategies.** We visualize the masking strategy of ST-MAE (Feichtenhofer et al., 2022) (random), VideoMAE (Tong et al., 2022) (space-only), VideoMAE V2 (Wang et al., 2023a) (running cell (Qing et al., 2023)), and our EVEREST on Ego4D dataset. From the top, each row indicates the original frames, MAE masking, VideoMAE masking, VideoMAE V2 masking, and our proposed method, respectively. We set ρ_{pre} to 0.5.

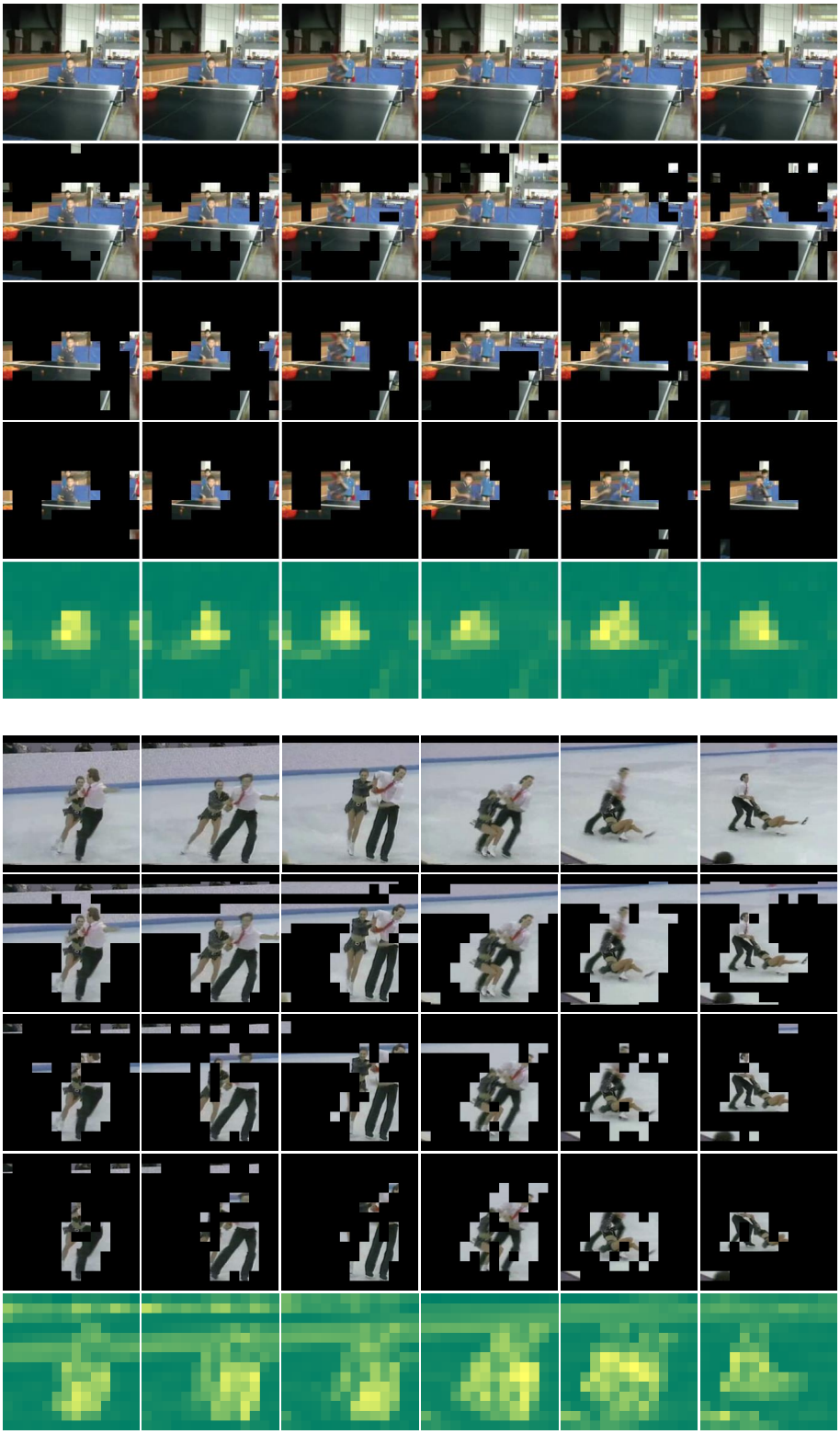


Figure 9: **More examples of Redundancy-robust token selection.** We show the original video frames in the first row, Redundancy-robust masking results in the second, third, and fourth row by varying the ρ_{pre} to 0.5, 0.25, and 0.15, respectively, and obtained importance heatmaps in the last row on UCF101 (Soomro et al., 2012) dataset.

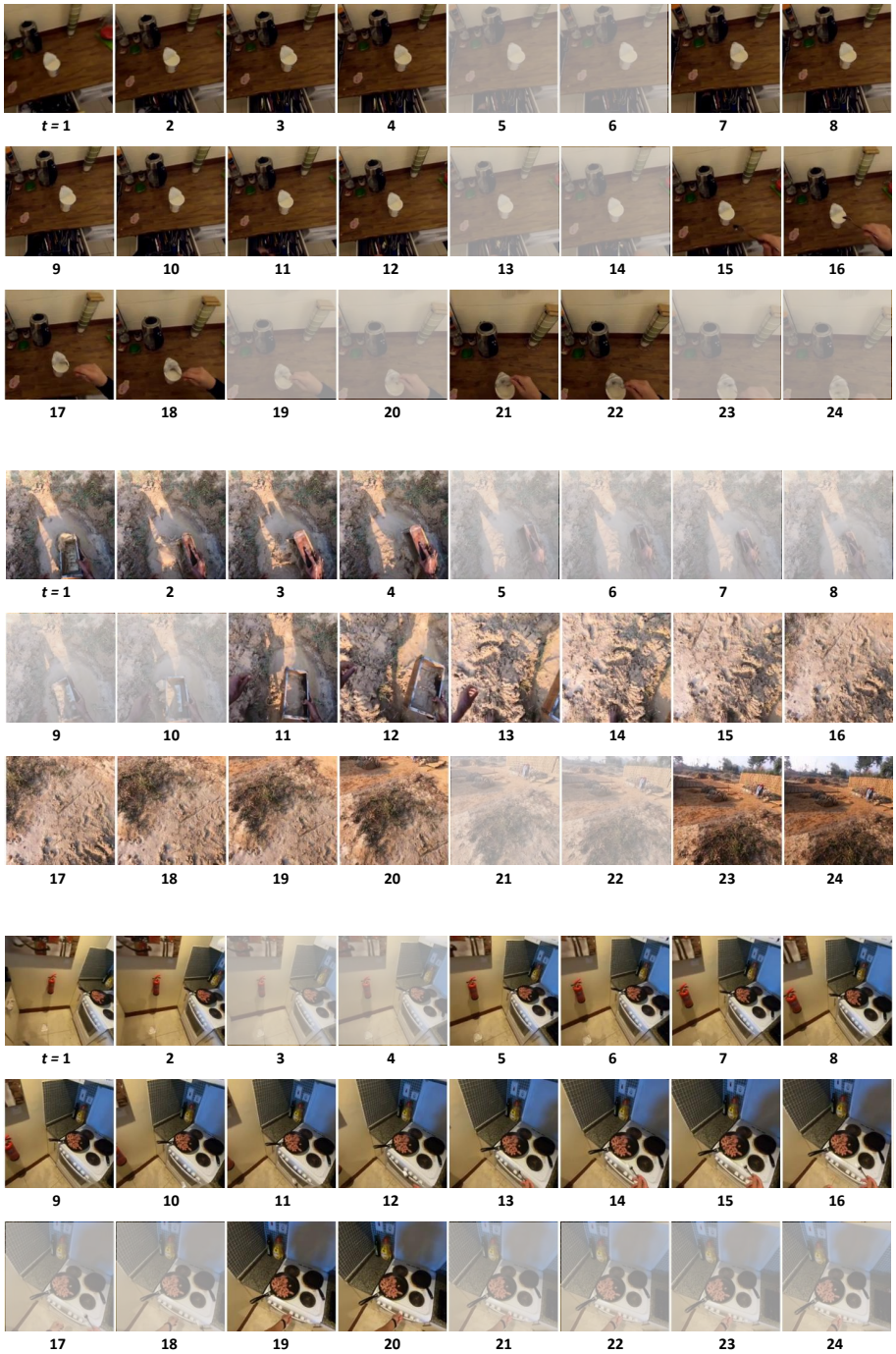


Figure 10: **Examples of on-the-fly information-intensive frame sampling based on our EVEREST.** Among 24 frames in the given video clip, our EVEREST adaptively selects frames containing rich motion information based on temporal correlation. Non-blurred frames are sampled through our method and used in the pre-training phase.