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] *: Equal Contribution

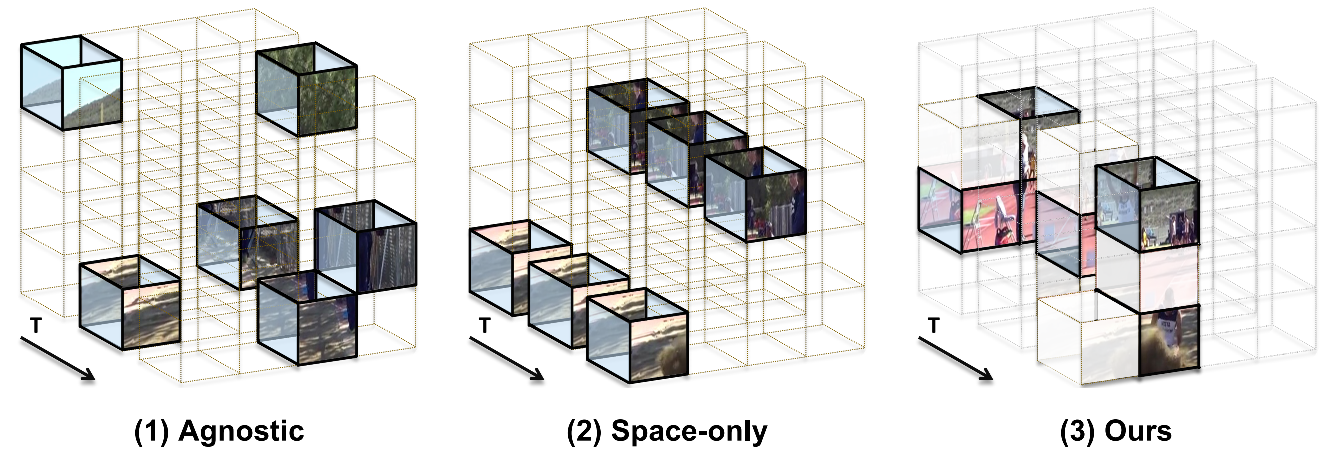
Korea Military Academy[1] *UNC-Chapel Hill*[2] *KAIST*[3] *ETRI*[4] *DeepAuto.ai*[5]

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

KAIST

Motivation

❑ Comparison of Masked Video Autoencoders



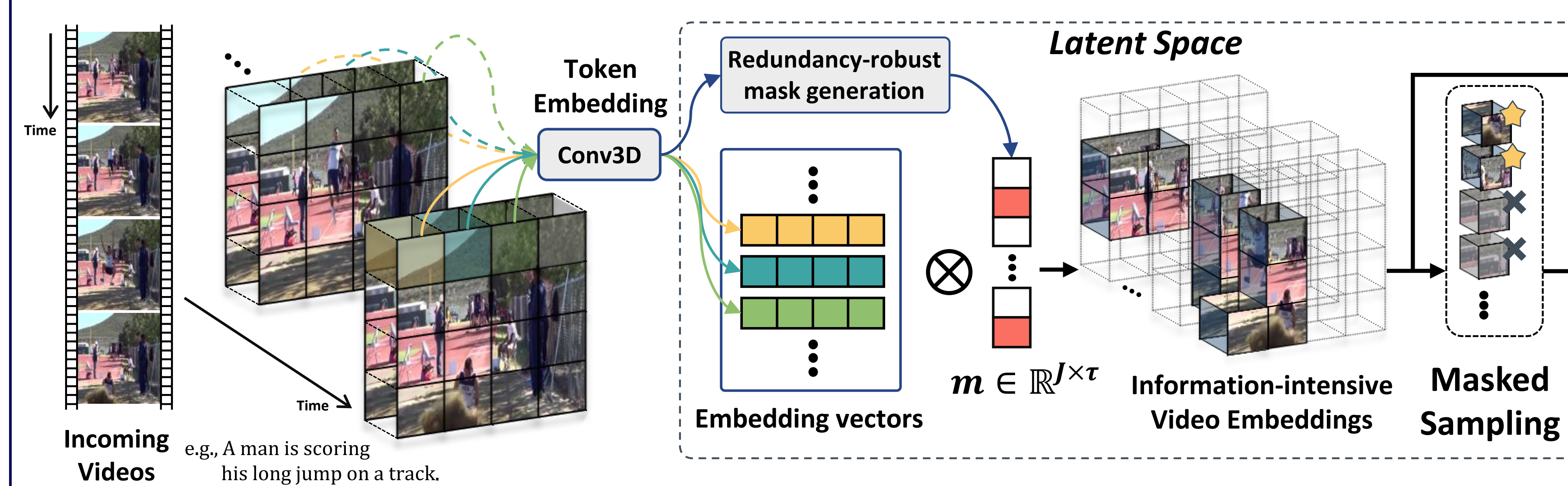
- ✓ Recently proposed Masked Video Autoencoders reconstruct randomly masked spatiotemporal regions in video clips.
- ✓ However, **tokens** (a pair of two temporally successive patches in the same space) in videos **are not equally valuable** to reconstruct.
- ✓ Moreover, learning representations from videos is infeasible without **a huge computing budget**.

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**

- ✓ 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.

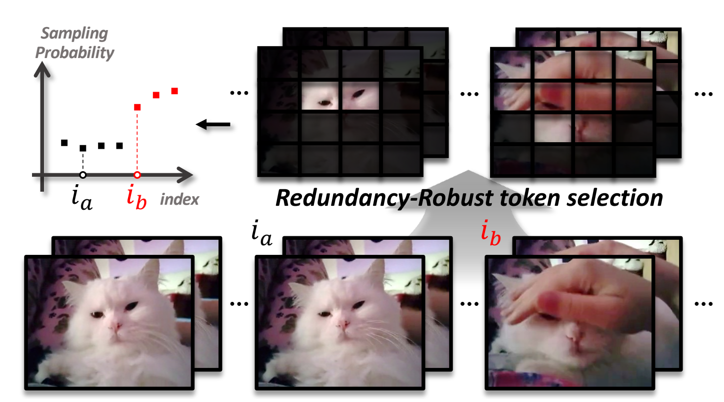
Methodology

❑ Redundancy-robust(ReRo) Masking Generation



Our **ReRo 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**.

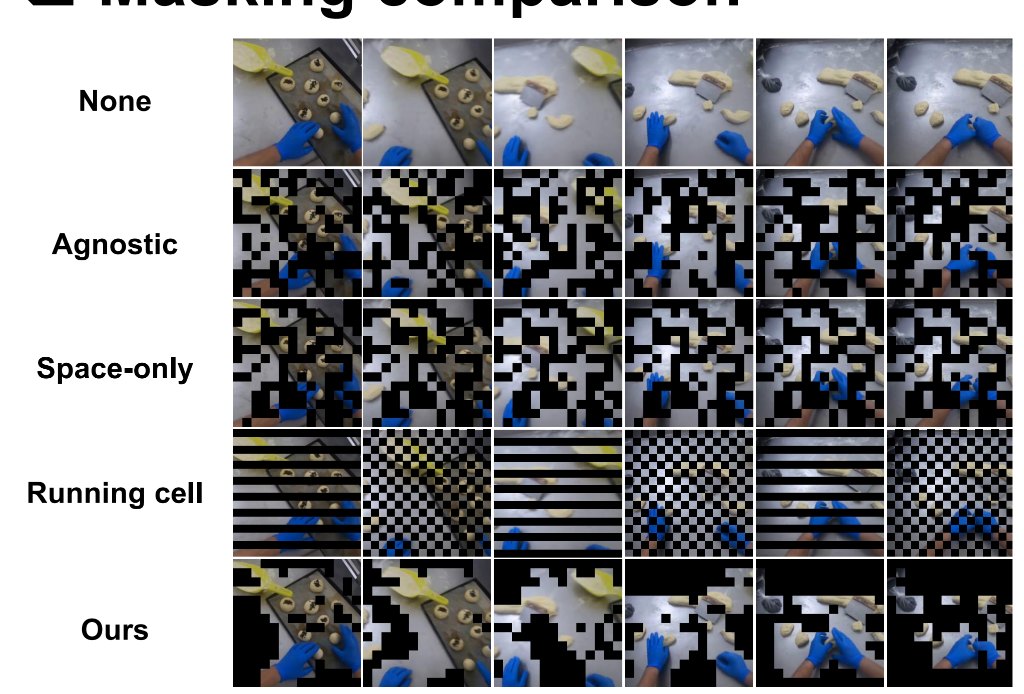
❑ On-the-fly Information-intensive Frame Selection



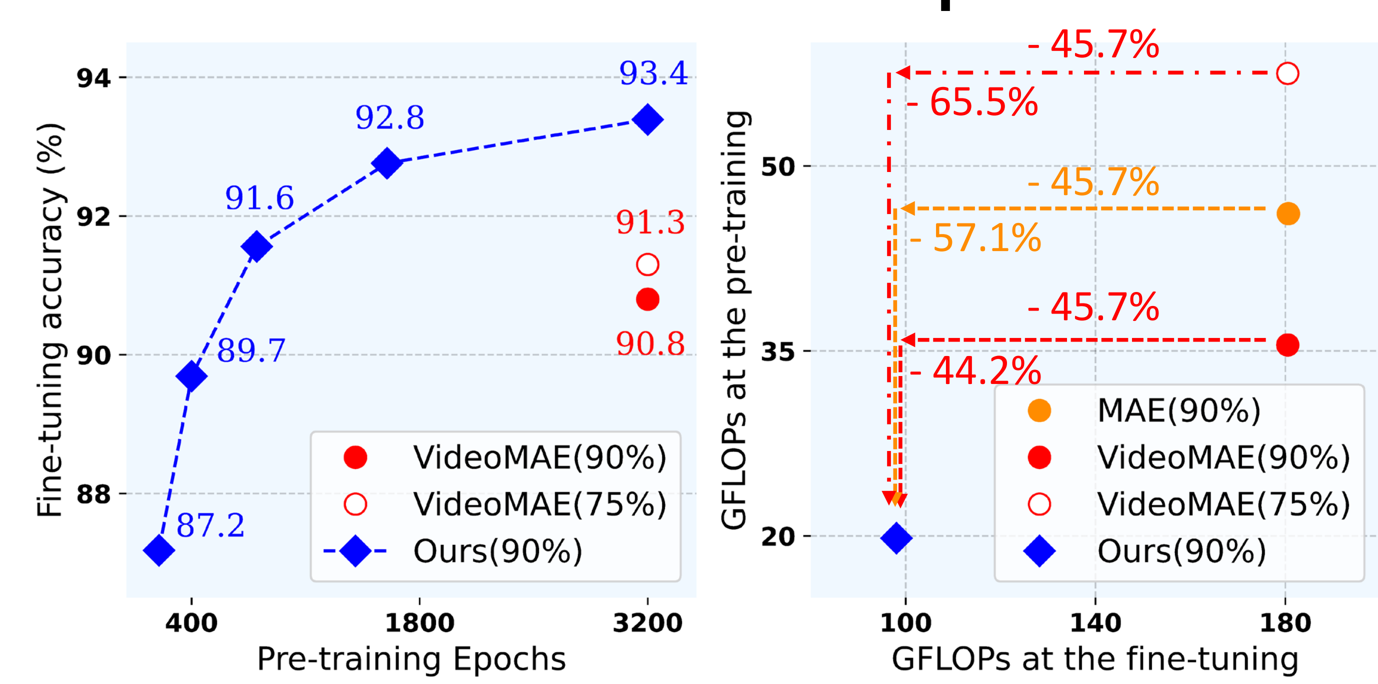
We **adaptively select frames** based on the ReRo token frequency, which indicates **significance compared to the other frames**. Our frame selection method is crucial to **better capture causality** in the arrival video, as the model can observe longer video fragments while **avoiding redundant frames**.

Experiments

❑ Masking comparison

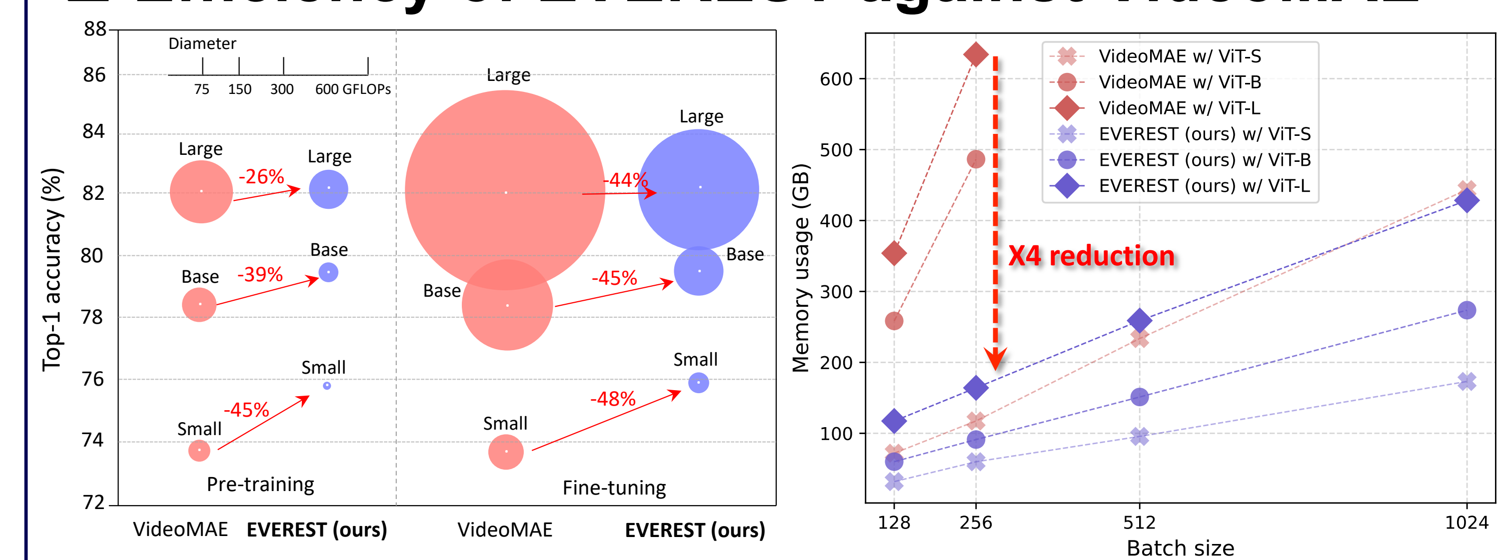


❑ Performance & GFLOPs comparison



Experiments

❑ Efficiency of EVEREST against VideoMAE



❑ Performance comparison on K400

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**

❑ EVEREST-Finetuning with other MVAs on K400

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

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 empirically demonstrate that our method is significantly **more efficient in computations, memory, and training time** than strong baselines.