NEURAL INFORMATION PROCESSING SYSTEMS

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

Zhan Tong[1,2] Yibing Song[2] Jue Wang[2] Limin Wang[1,3]

[1] State Key Laboratory for Novel Software Technology, Nanjing University

[2] Tencent AI Lab [3] Shanghai AI Lab

南京大學 NANJING UNIVERSITY

MCG MULTIMEDIA COMPUTING GROUP 媒体计算研究组

Tencent AI Lab 腾讯人工智能实验室

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

Motivation

- **Transformer improves a series of computer vision tasks**
  - include **fewer** inductive biases
  - e.g., classification, detection, segmentation and **video understanding**
- **Challenges for video understanding**
  - temporal **redundancy** and **correlation**
  - **higher** computational consumption for video
- **Challenges for training video transformer**
  - need extra **large-scale image/video** data
  - **heavily** depend on **pre-trained models**

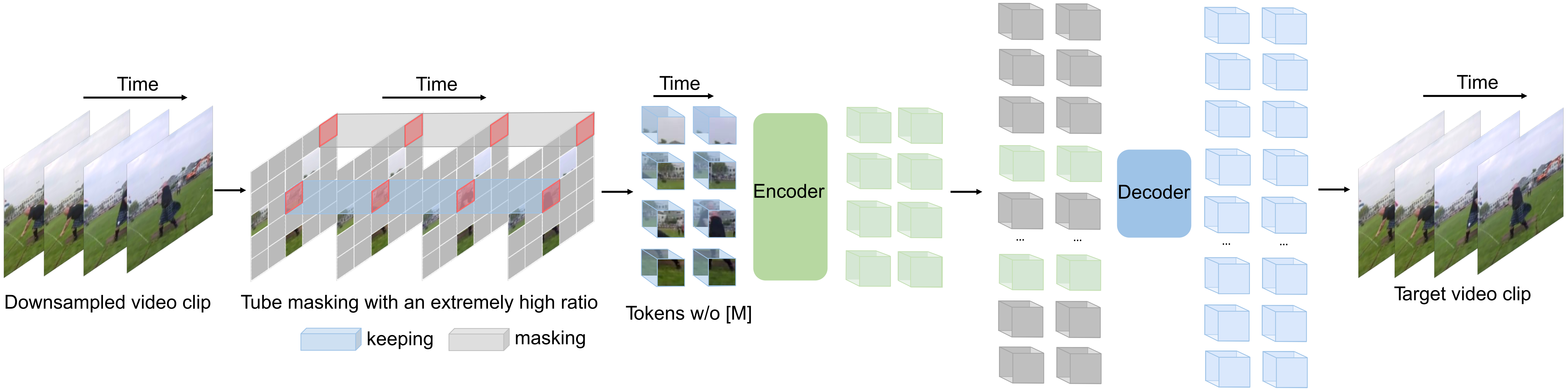
Motivation

NEURAL INFORMATION PROCESSING SYSTEMS

How to efficiently train a vanilla ViT on the video dataset itself without using any pre-trained model or extra data?

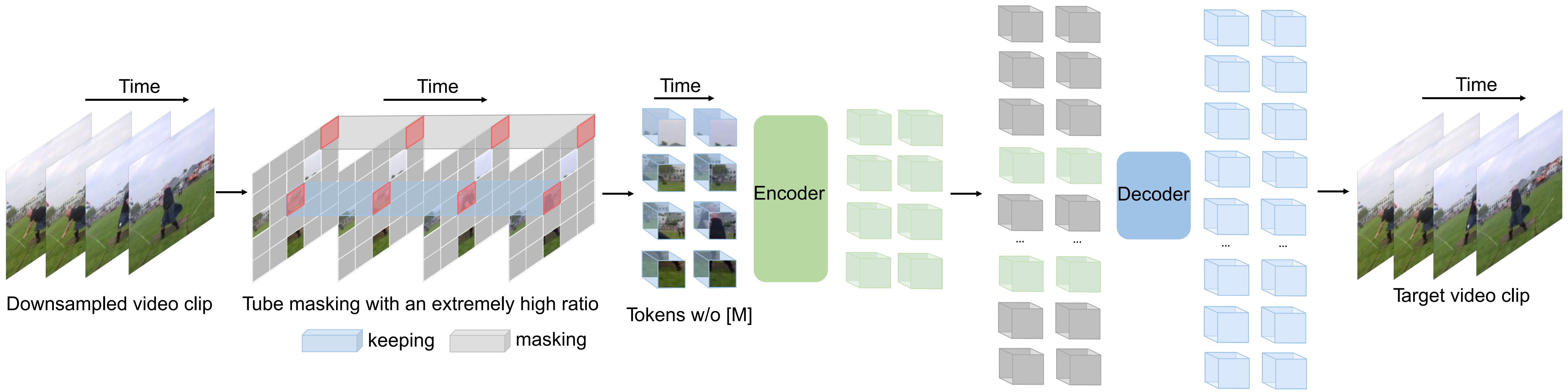
VideoMAE

- Our VideoMAE attempts to solve it in two aspects
  - **Self-supervised** pre-training with **masked autoencoder**
  - A new masking strategy: **tube masking** with an **extremely high** ratio



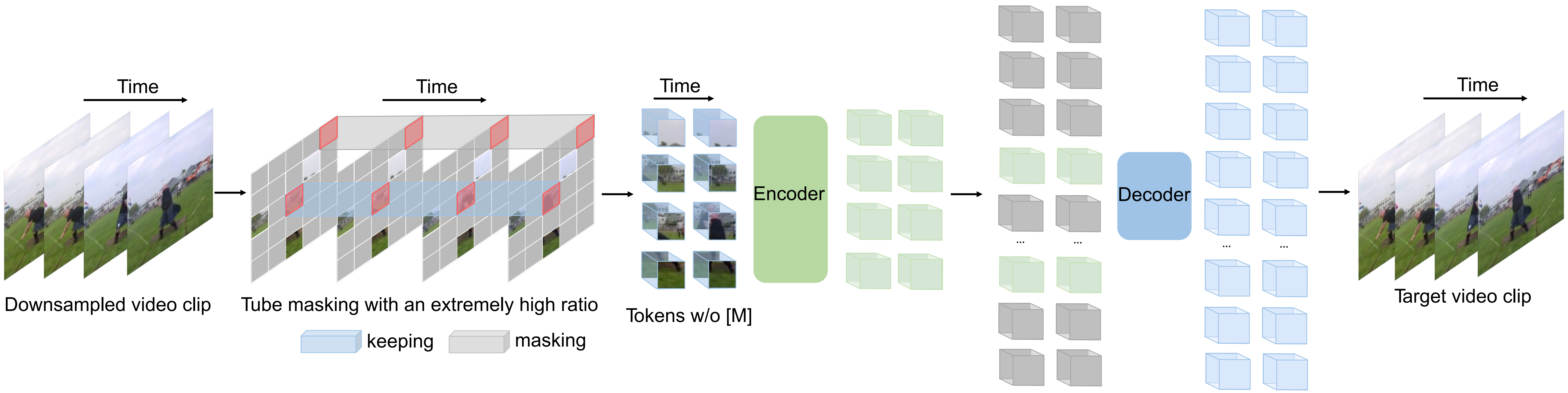
VideoMAE

NEURAL INFORMATION PROCESSING SYSTEMS



- **Self-supervised pre-training with masked autoencoder**
  - **a simple but effective masking and reconstruction proxy task**
  - **an efficient pre-training process with only unmasked tokens into the encoder.**

VideoMAE



- **A new masking strategy:**
  - **tube masking with an extremely high ratio**
  - **makeing video reconstruction a more challenging self-supervision task**

Overall VideoMAE

→ **and eventually, VideoMAE is**

- a **simple**, **data-efficient** method for **self-supervised video pre-training**
  - with **high** performance and **no** extra data **required**

Key Ablation Study

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

Masking strategy

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

Pre-training strategy

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

Pre-training dataset

Main Results and Analysis

⇥ **VideoMAE is a data-efficient learner**

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

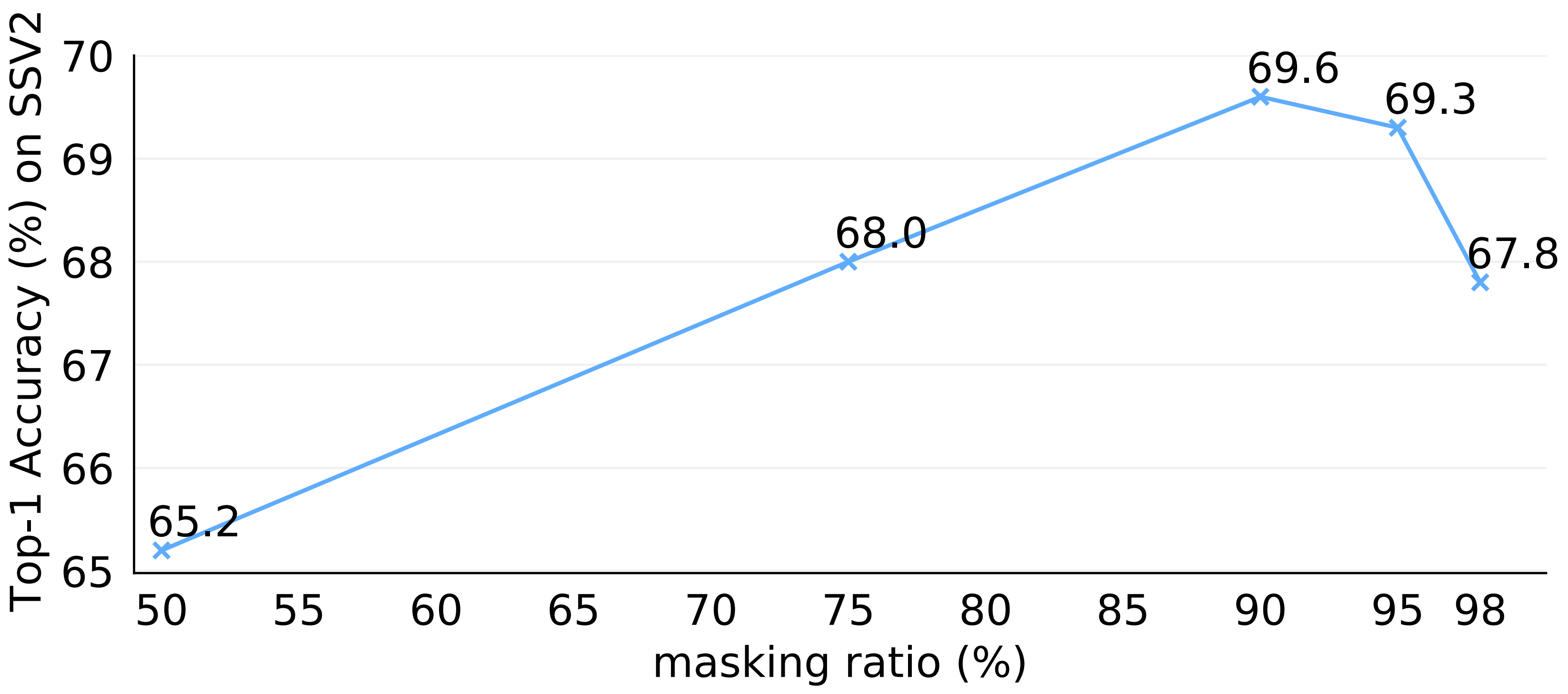
Performance on video datasets of **different scales**

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

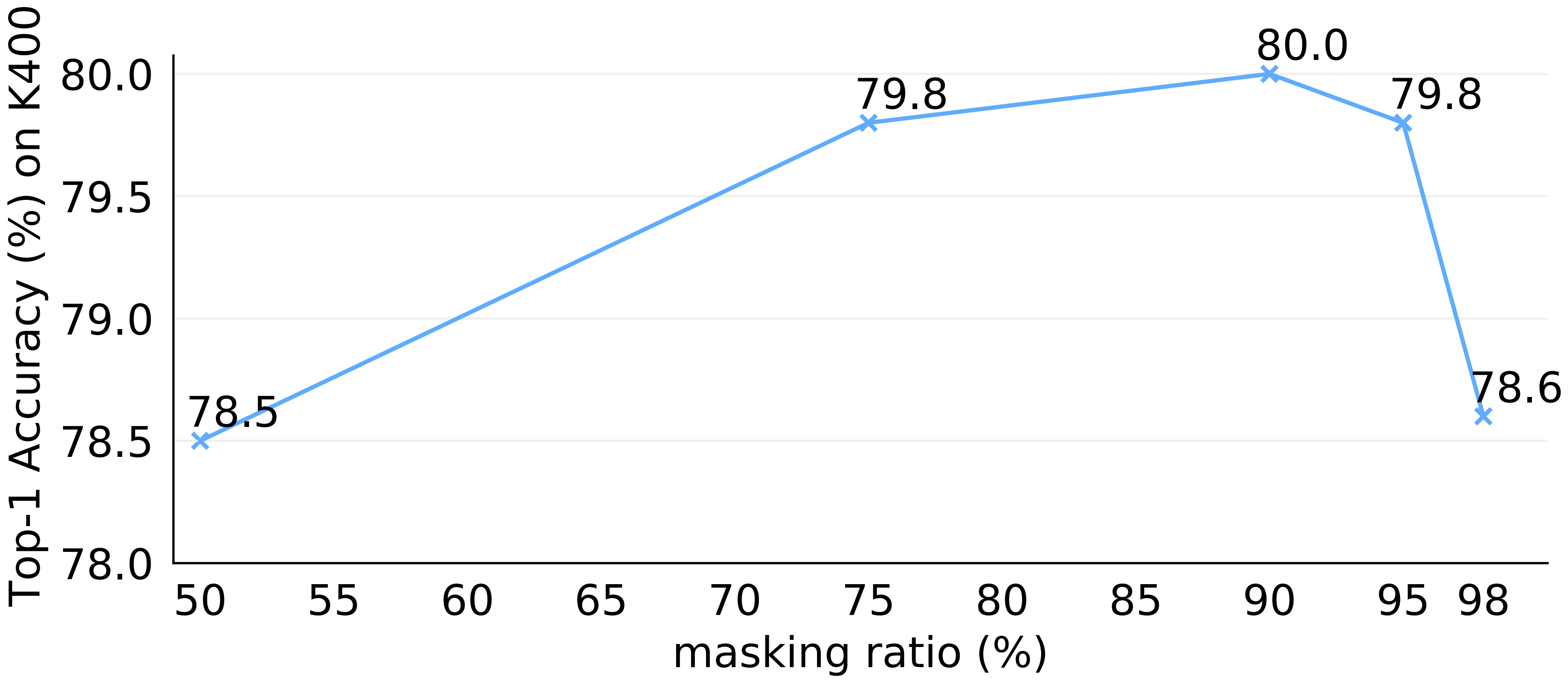
Efficiency and effectiveness on Something-Something V2

Main Results and Analysis

→ The effect of an **extremely high masking ratio**



(a) Performance on Something-Something V2



(b) Performance on Kinetics-400

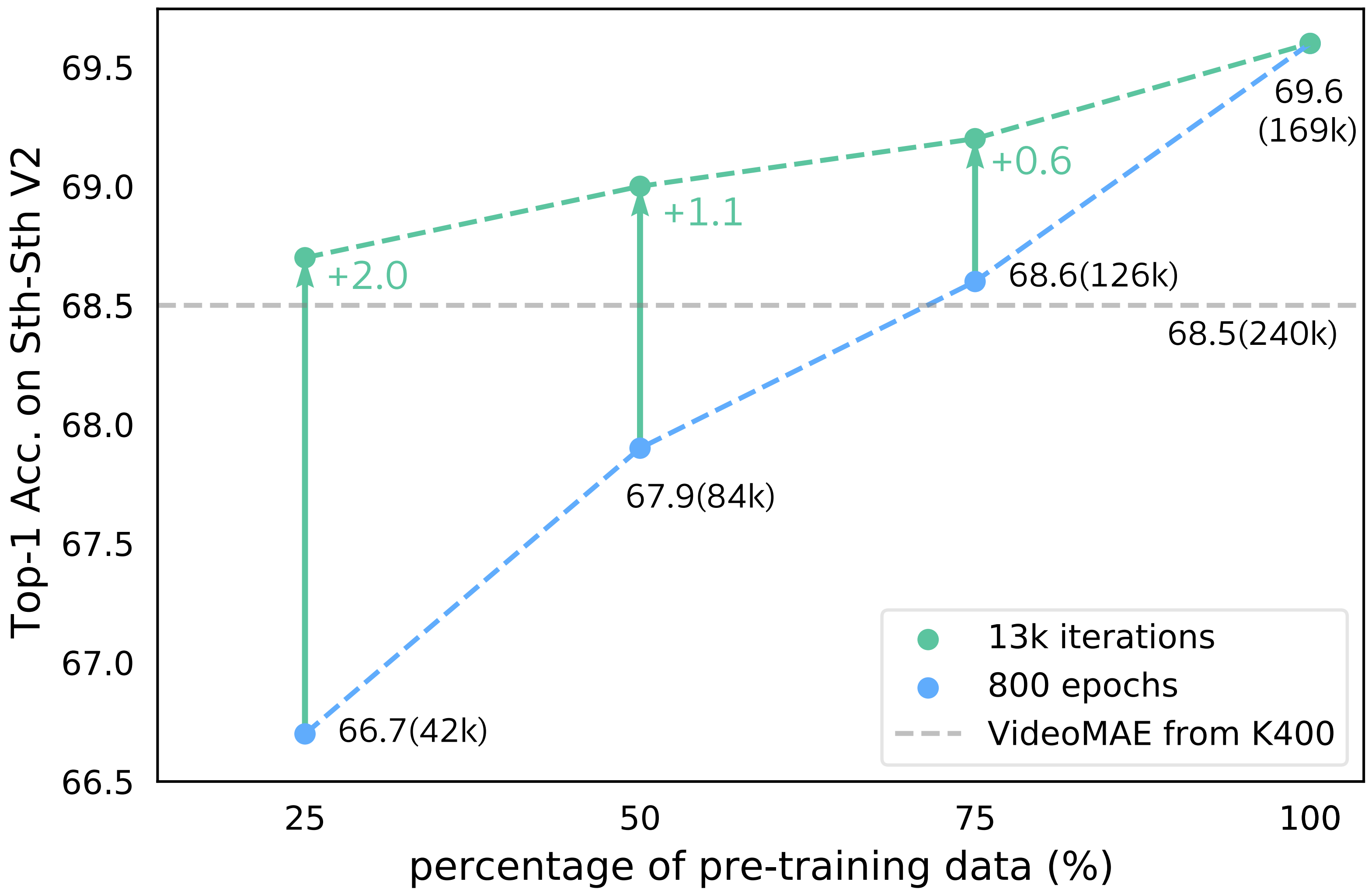
Main Results and Analysis

NEURAL INFORMATION PROCESSING SYSTEMS

⇥ Transfer learning: quality vs. quantity

method	K400 → SSV2	K400 → UCF	K400 → HMDB
MoCo v3	62.4	93.2	67.9
VideoMAE	**68.5**	**96.1**	**73.3**

feature **transferability** on smaller datasets



data quality is more important than data quantity

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

⇥ **Leading performance on Something-Something V2**

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
TEINet$_{En}$ [39]	ResNet50$_{\times 2}$	ImageNet-1K	✓	8+16	99×10×3	50	66.5	N/A
TANet$_{En}$ [40]	ResNet50$_{\times 2}$		✓	8+16	99×2×3	51	66.0	90.1
TDN$_{En}$ [74]	ResNet101$_{\times 2}$		✓	8+16	198×1×3	88	69.6	92.2
SlowFast [22]	ResNet101	Kinetics-400	✓	8+32	106×1×3	53	63.1	87.6
MViTv1 [21]	MViTv1-B		✓	64	455×1×3	37	67.7	90.9
TimeSformer [6]	ViT-B	ImageNet-21K	✓	8	196×1×3	121	59.5	N/A
TimeSformer [6]	ViT-L		✓	64	5549×1×3	430	62.4	N/A
ViViT FE [3]	ViT-L	IN-21K+K400	✓	32	995×4×3	N/A	65.9	89.9
Motionformer [50]	ViT-B		✓	16	370×1×3	109	66.5	90.1
Motionformer [50]	ViT-L		✓	32	1185×1×3	382	68.1	91.2
Video Swin [38]	Swin-B		✓	32	321×1×3	88	69.6	92.7
VIMPAC [64]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	68.1	N/A
BEVT [76]	Swin-B	IN-1K+K400+DALLE	✗	32	321×1×3	88	70.6	N/A
MaskFeat$\uparrow$312 [79]	MViT-L	Kinetics-600	✓	40	2828×1×3	218	75.0	95.0
VideoMAE	ViT-B	Kinetics-400	✗	16	180×2×3	87	69.7	92.3
VideoMAE	ViT-L	Kinetics-400	✗	16	597×2×3	305	74.0	94.6
VideoMAE	ViT-S	*no external data*	✗	16	57×2×3	22	66.8	90.3
VideoMAE	ViT-B		✗	16	180×2×3	87	70.8	92.4
VideoMAE	ViT-L		✗	16	597×2×3	305	74.3	94.6
VideoMAE	ViT-L		✗	32	1436×1×3	305	**75.4**	**95.2**

Experiments

→ **Leading performance on Kinetics-400**

Method	Backbone	Extra data	Ex. labels	Frames	GFLOPs	Param	Top-1	Top-5
NL I3D [77]	ResNet101	ImageNet-1K	✓	128	359×10×3	62	77.3	93.3
TANet [40]	ResNet152	ImageNet-1K	✓	16	242×4×3	59	79.3	94.1
TDN_{En} [74]	ResNet101	ImageNet-1K	✓	8+16	198×10×3	88	79.4	94.4
TimeSformer [6]	ViT-L	ImageNet-21K	✓	96	8353×1×3	430	80.7	94.7
ViViT FE [3]	ViT-L	ImageNet-21K	✓	128	3980×1×3	N/A	81.7	93.8
Motionformer [50]	ViT-L	ImageNet-21K	✓	32	1185×10×3	382	80.2	94.8
Video Swin [38]	Swin-L	ImageNet-21K	✓	32	604×4×3	197	83.1	95.9
ViViT FE [3]	ViT-L	JFT-300M	✓	128	3980×1×3	N/A	83.5	94.3
ViViT [3]	ViT-H	JFT-300M	✓	32	3981×4×3	N/A	84.9	95.8
VIMPAC [64]	ViT-L	HowTo100M+DALLE	✗	10	N/A×10×3	307	77.4	N/A
BEVT [76]	Swin-B	IN-1K+DALLE	✗	32	282×4×3	88	80.6	N/A
$\text{MaskFeat}_{\uparrow 352}$ [79]	MViT-L	Kinetics-600	✗	40	3790×4×3	218	87.0	97.4
ip-CSN [68]	ResNet152	*no external data*	✗	32	109×10×3	33	77.8	92.8
SlowFast [22]	R101+NL	*no external data*	✗	16+64	234×10×3	60	79.8	93.9
MViTv1 [21]	MViTv1-B	*no external data*	✗	32	170×5×1	37	80.2	94.4
MaskFeat [79]	MViT-L	*no external data*	✗	16	377×10×1	218	84.3	96.3
VideoMAE	ViT-S	*no external data*	✗	16	57×5×3	22	79.0	93.8
VideoMAE	ViT-B	*no external data*	✗	16	180×5×3	87	81.5	95.1
VideoMAE	ViT-L	*no external data*	✗	16	597×5×3	305	85.2	96.8
VideoMAE	ViT-H	*no external data*	✗	16	1192×5×3	633	**86.6**	**97.1**
$\textbf{VideoMAE}_{\uparrow 320}$	ViT-L	*no external data*	✗	32	3958×4×3	305	86.1	97.3
$\textbf{VideoMAE}_{\uparrow 320}$	ViT-H	*no external data*	✗	32	7397×4×3	633	**87.4**	**97.6**

Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

→ **Leading performance** on AVA v2.2

Method	Backbone	Pre-train Dataset	Extra Labels	$T \times \tau$	GFLOPs	Param	mAP
supervised [22]	SlowFast-R101	Kinetics-400	✓	8×8	138	53	23.8
CVRL [53]	SlowOnly-R50	Kinetics-400	✗	32×2	42	32	16.3
ρBYOL$_{\rho=3}$ [23]	SlowOnly-R50	Kinetics-400	✗	8×8	42	32	23.4
ρMoCo$_{\rho=3}$ [23]	SlowOnly-R50	Kinetics-400	✗	8×8	42	32	20.3
MaskFeat$\uparrow$312 [79]	MViT-L	Kinetics-400	✓	40×3	2828	218	37.5
MaskFeat$\uparrow$312 [79]	MViT-L	Kinetics-600	✓	40×3	2828	218	38.8
VideoMAE	ViT-S	Kinetics-400	✗	16×4	57	22	22.5
VideoMAE	ViT-S	Kinetics-400	✓	16×4	57	22	28.4
VideoMAE	ViT-B	Kinetics-400	✗	16×4	180	87	26.7
VideoMAE	ViT-B	Kinetics-400	✓	16×4	180	87	31.8
VideoMAE	ViT-L	Kinetics-400	✗	16×4	597	305	34.3
VideoMAE	ViT-L	Kinetics-400	✓	16×4	597	305	37.0
VideoMAE	ViT-H	Kinetics-400	✗	16×4	1192	633	**36.5**
VideoMAE	ViT-H	Kinetics-400	✓	16×4	1192	633	**39.5**
VideoMAE	ViT-L	Kinetics-700	✗	16×4	597	305	**36.1**
VideoMAE	ViT-L	Kinetics-700	✓	16×4	597	305	**39.3**

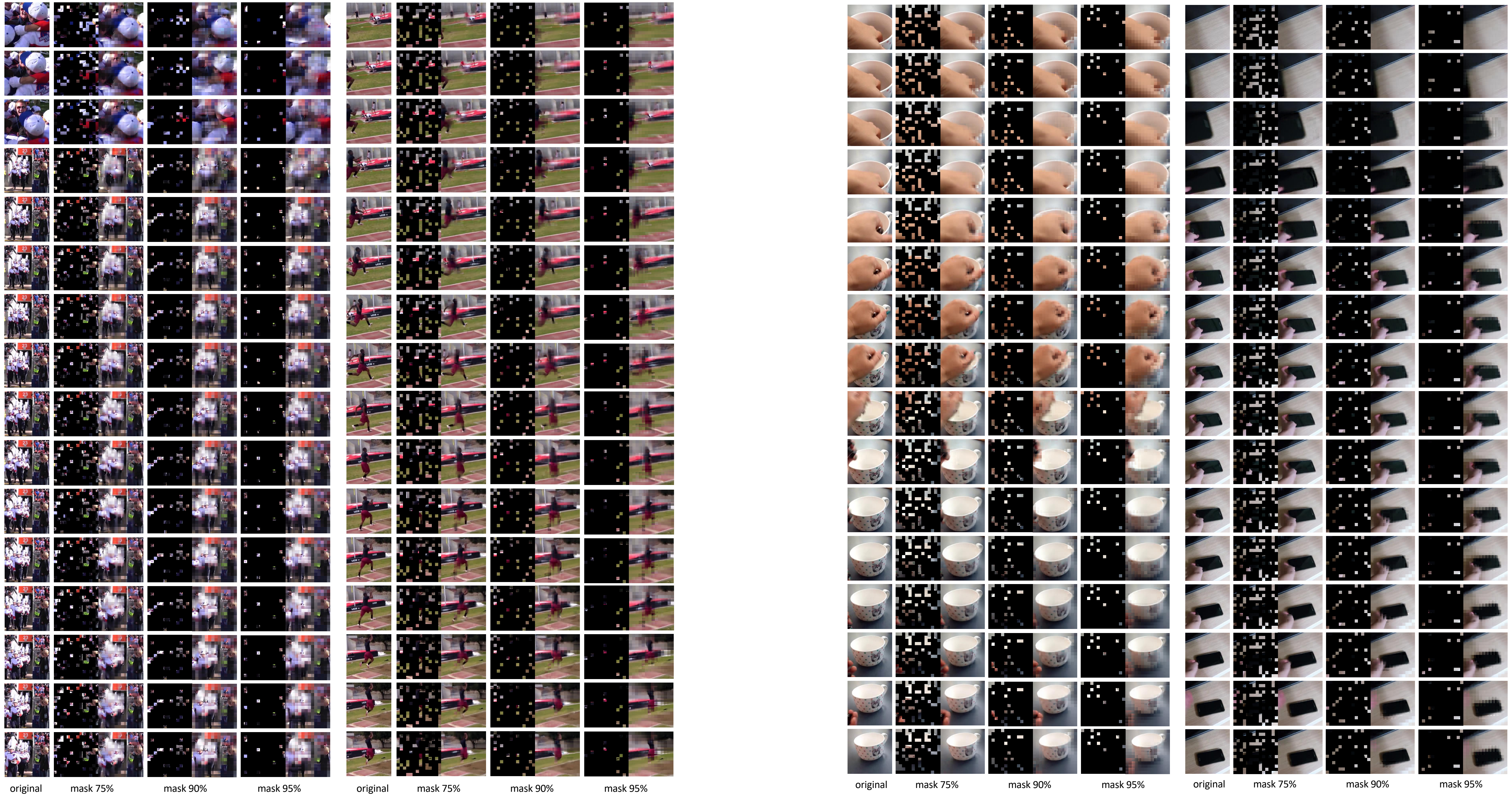
Experiments

NEURAL INFORMATION PROCESSING SYSTEMS

→ **Leading performance** on UCF101 and HMDB51

Method	Backbone	Extra data	Frames	Param	Modality	UCF101	HMDB51
OPN [35]	VGG	UCF101	N/A	N/A	V	59.6	23.8
VCOP [82]	R(2+1)D	UCF101	N/A	N/A	V	72.4	30.9
CoCLR [29]	S3D-G	UCF101	32	9M	V	81.4	52.1
Vi^2CLR [18]	S3D	UCF101	32	9M	V	82.8	52.9
VideoMAE	ViT-B	*no external data*	16	87M	V	**91.3**	**62.6**
SpeedNet [5]	S3D-G	Kinetics-400	64	9M	V	81.1	48.8
VTHCL [84]	SlowOnly-R50	Kinetics-400	8	32M	V	82.1	49.2
Pace [73]	R(2+1)D	Kinetics-400	16	15M	V	77.1	36.6
MemDPC [28]	R-2D3D	Kinetics-400	40	32M	V	86.1	54.5
CoCLR [29]	S3D-G	Kinetics-400	32	9M	V	87.9	54.6
RSPNet [12]	S3D-G	Kinetics-400	64	9M	V	93.7	64.7
VideoMoCo [45]	R(2+1)D	Kinetics-400	16	15M	V	78.7	49.2
Vi^2CLR [18]	S3D	Kinetics-400	32	9M	V	89.1	55.7
CVRL [53]	SlowOnly-R50	Kinetics-400	32	32M	V	92.9	67.9
CVRL [53]	SlowOnly-R50	Kinetics-600	32	32M	V	93.6	69.4
CVRL [53]	Slow-R152 (2×)	Kinetics-600	32	328M	V	94.4	70.6
CORP_f [32]	SlowOnly-R50	Kinetics-400	32	32M	V	93.5	68.0
$\rho\text{SimCLR}_{\rho=2}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	88.9	N/A
$\rho\text{SwAV}_{\rho=2}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	87.3	N/A
$\rho\text{MoCo}_{\rho=2}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	91.0	N/A
$\rho\text{BYOL}_{\rho=2}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	92.7	N/A
$\rho\text{BYOL}_{\rho=4}$ [23]	SlowOnly-R50	Kinetics-400	8	32M	V	94.2	72.1
VideoMAE(Ours)	ViT-B	Kinetics-400	16	87M	V	**96.1**	**73.3**

Visualizations



Recap

NEURAL INFORMATION PROCESSING SYSTEMS

- **VideoMAE, a data-efficient learner, enjoys**
  - **masked video modeling for video pre-training**
  - **a simple, efficient and strong baseline for SSVP**
  - **leading performance with no extra data required**

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

Code is available at
https://github.com/MCG-NJU/VideoMAE

VideoMAE

NEURAL INFORMATION
PROCESSING SYSTEMS