WED-PM-210

VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking

Limin Wang[1,2,*] Bingkun Huang[1,2,*] Zhiyu Zhao[1,2] Zhan Tong[1]
Yinan He[2] Yi Wang[2] Yali Wang[3,2] Yu Qiao[2,3]

[1] State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] Shanghai AI Lab, China [3] Shenzhen Institute of Advanced Technology, CAS, China

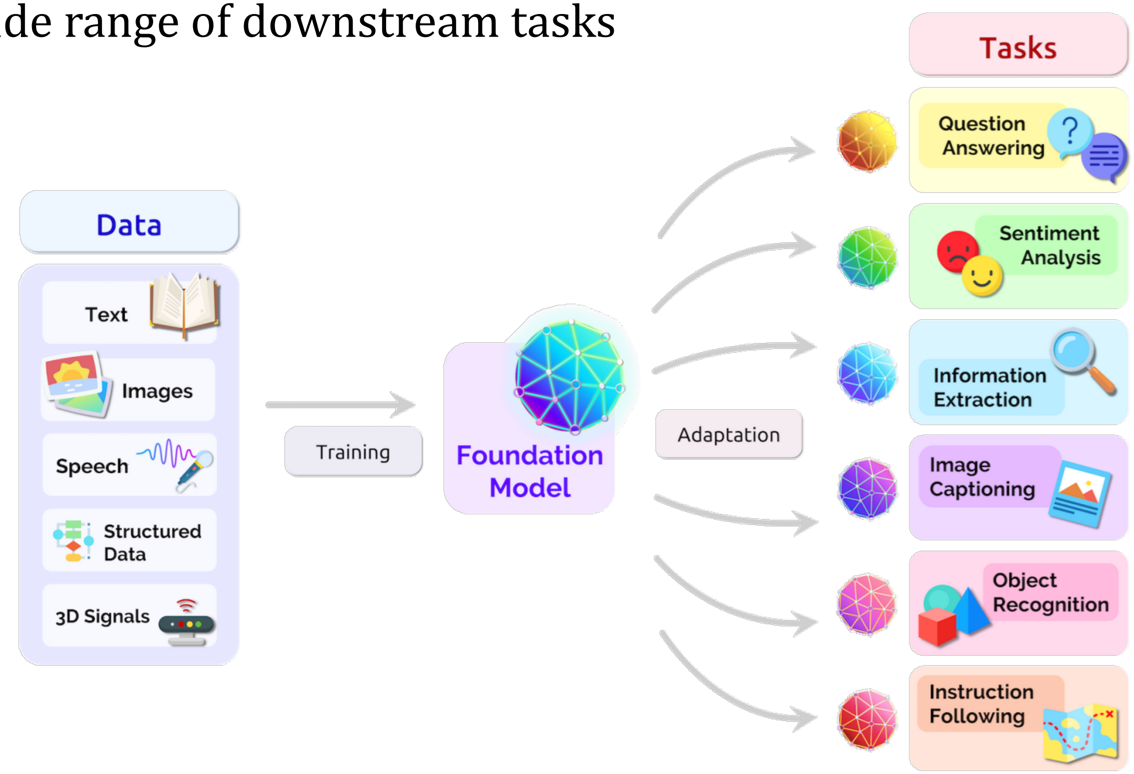
Code & Weights Here!

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Background

- On the Opportunities and Risks of Foundation Models, arXiv 2022

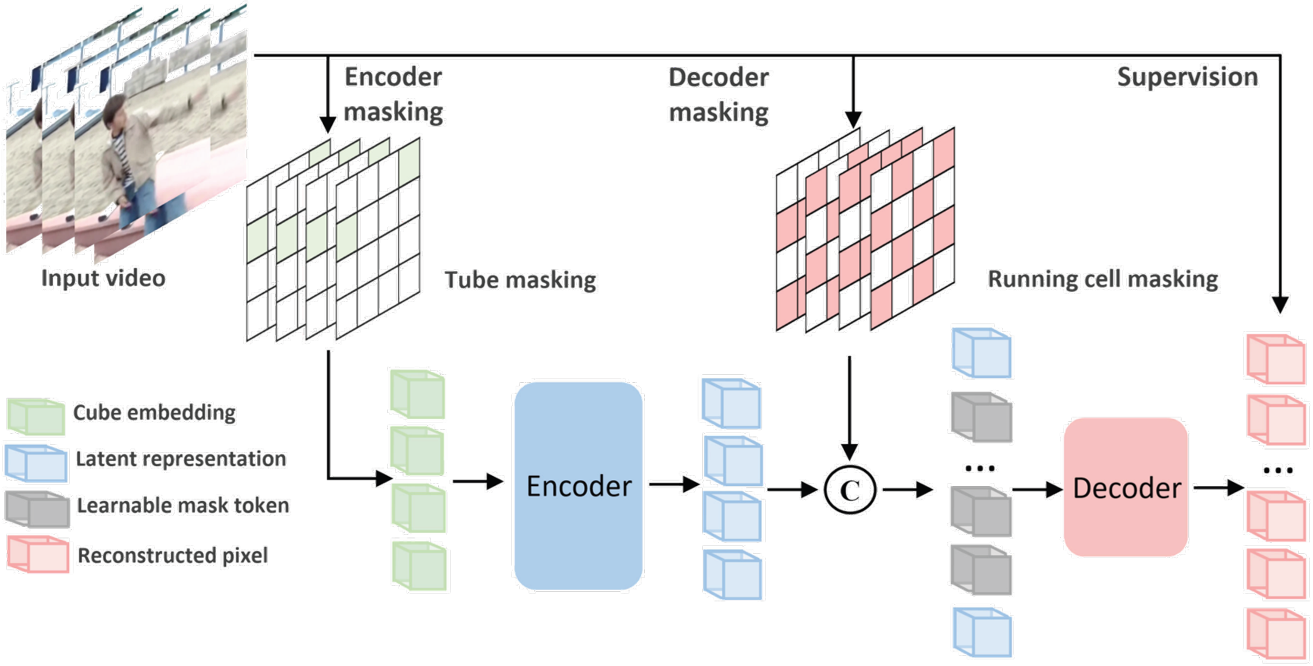
- Foundation Model
  - trained on broad data (generally using self-supervision at scale)
  - can be adapted to a wide range of downstream tasks



VideoMAE V2

- Aims to
  - study the scaling property of video masked autoencoder
  - push its performance limit on video downstream tasks

- Methods
  - Dual masking
  - Model scaling
  - Data Scaling
  - Progressive training

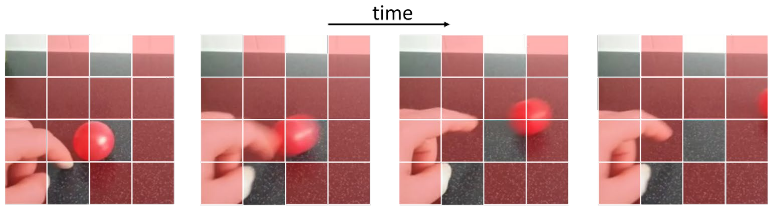


- Results
  - 6 SOTA on video tasks

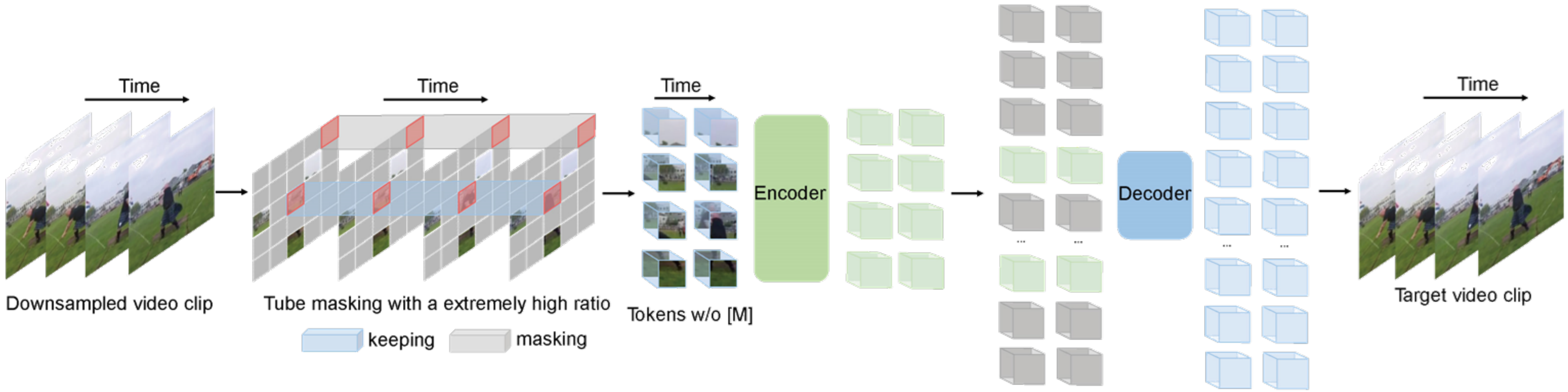
Revisit VideoMAE

- Asymmetric encoder-decoder architecture
- Simple but effective masking and reconstruction pretext task

- Tube masking with extremely high mask ratio



(d) tube masking



Challenges of scaling VideoMAE

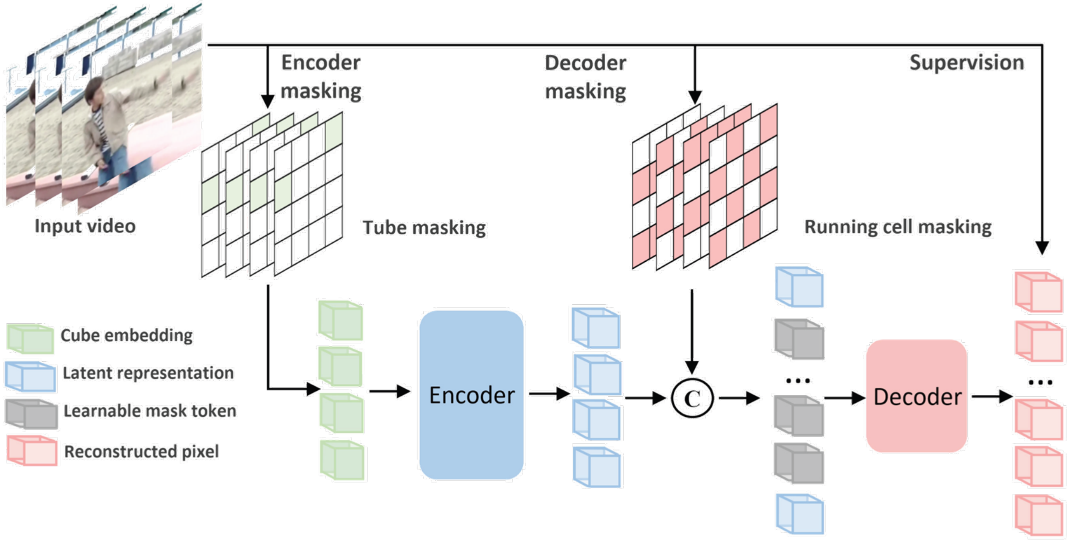
- **Bottleneck** of computational cost and memory consumption
  - Dual Masking
    - 90% Tube Masking for Encoder
    - 50% Running Cell Masking for Decoder

- **Limited availability** of public video datasets
  - UnlabeledHybrid
    - Kinetics, SSv2, AVA, WebVid, self-collected Instagram videos

- **Uncertainty in adapting** the billion-level pre-trained model
  - Progressive training
    - Pre-training → Post-pre-training → Specific Fine-tunning

Bottleneck of computational cost and memory consumption

- Dual Masking
  - 90% Tube Masking for Encoder
  - 50% Running Cell Masking for Decoder



Decoder Masking	ρ^d	Top-1	FLOPs
None	0%	**70.28**	35.48G
Frame	50%	69.76	25.87G
Random	50%	64.87	25.87G
Running cell [1]	50%	66.74	25.87G
Running cell [2]	25%	70.22	31.63G
Running cell [2]	50%	70.15	25.87G
Running cell [2]	75%	70.01	21.06G

Masking	Backbone	pre-training dataset	FLOPs	Mems	Time	Speedup	Top-1
Encoder masking	ViT-B	Something-Something V2	35.48G	631M	28.4h	-	70.28
Dual masking	ViT-B	Something-Something V2	25.87G	328M	15.9h	**1.79**×	70.15
Encoder masking	ViT-g	UnlabeledHybrid	263.93G	1753M	356h [1]	-	-
Dual masking	ViT-g	UnlabeledHybrid	241.61G	1050M	241h	**1.48**×	77.00

Limited availability of public video datasets

- UnlabeledHybrid
  - Kinetics, SSv2, AVA, WebVid, self-collected Instagram videos
  - 1.35 million video clips

method	pre-train data	data size	epoch	ViT-B	ViT-L	ViT-H	ViT-g
MAE-ST [18]	Kinetics400	0.24M	1600	81.3	84.8	85.1	-
MAE-ST [18]	IG-uncurated	1M	1600	-	84.4	-	-
VideoMAE V1 [63]	Kinetics400	0.24M	1600	**81.5**	85.2	86.6	-
VideoMAE V2	UnlabeledHybrid	1.35M	1200	**81.5** (77.0)	**85.4** (81.3)	**86.9** (83.2)	**87.2** (83.9)
$\Delta Acc.$ with V1	-	-	-	*+ 0%*	*+ 0.2%*	*+ 0.3%*	-

Results on the Kinetics-400 dataset

method	pre-train data	data size	epoch	ViT-B	ViT-L	ViT-H	ViT-g
MAE-ST [18]	Kinetics400	0.24M	1600	-	72.1	74.1	-
MAE-ST [18]	Kinetics700	0.55M	1600	-	73.6	75.5	-
VideoMAE V1 [63]	Something-Something V2	0.17M	2400	70.8	74.3	74.8	-
VideoMAE V2	UnlabeledHybrid	1.35M	1200	**71.2** (69.5)	**75.7** (74.00)	**76.8** (75.5)	**77.0** (75.7)
$\Delta Acc.$ with V1	-	-	-	*+ 0.4%*	*+ 1.4%*	*+ 2.0%*	-

Results on the Something-Something V2 dataset

Uncertainty in adapting the billion-level pre-trained model

- Progressive training
  - Pre-training on UnlabeledHybrid
  - Post-pre-training on LabeledHybrid (Kinetics 710)
  - Specific Fine-tunning on downstream dataset

method	extra supervision	ViT-H	ViT-g
MAE-ST [18]	K600	86.8	-
VideoMAE V1 [63]	K710	88.1 (84.6)	-
VideoMAE V2	-	86.9 (83.2)	87.2 (83.9)
VideoMAE V2	K710	**88.6** (85.0)	**88.5** (85.6)
Δ*Acc.* with V1	K710	***+ 0.5%***	-

Study on progressive pre-training

Powerful VideoMAE V2-g

- ViT-giant with 1.01 billion parameters
- Performance Ranks

✓ SOTA	AVA-Kinetics	43.9
✓ SOTA	AVA v2.2	42.6
✓ SOTA	FineAction	18.2
✓ SOTA	THUMOS'14	69.6
✓ SOTA	HMDB-51	88.1
✓ SOTA	UCF101	99.6
✓ RANK #2	SSv1	68.7
✓ RANK #3	SSv2	77.0
✓ RANK #5	Kinetics-400	90.0
✓ RANK #9	Kinetics-600	89.9

Code & Weights Here!

- Distillation

Model	Dataset	Teacher Model	#Frame	K710 Top-1	K400 Top-1	K600 Top-1	Checkpoint
ViT-small	K710	vit_g_hybrid_pt_1200e_k710_ft	16x5x3	77.6	83.7	83.1	vit_s_k710_dl_from_giant.pth
		fine-tuning accuracy	16x7x3	--	84.0	84.6	--
ViT-base	K710	vit_g_hybrid_pt_1200e_k710_ft	16x5x3	81.5	86.6	85.9	vit_b_k710_dl_from_giant.pth
		fine-tuning accuracy	16x7x3	--	87.1	87.4	