Masked Video Distillation: Rethinking Masked Feature Modeling for Self-supervised Video Representation Learning

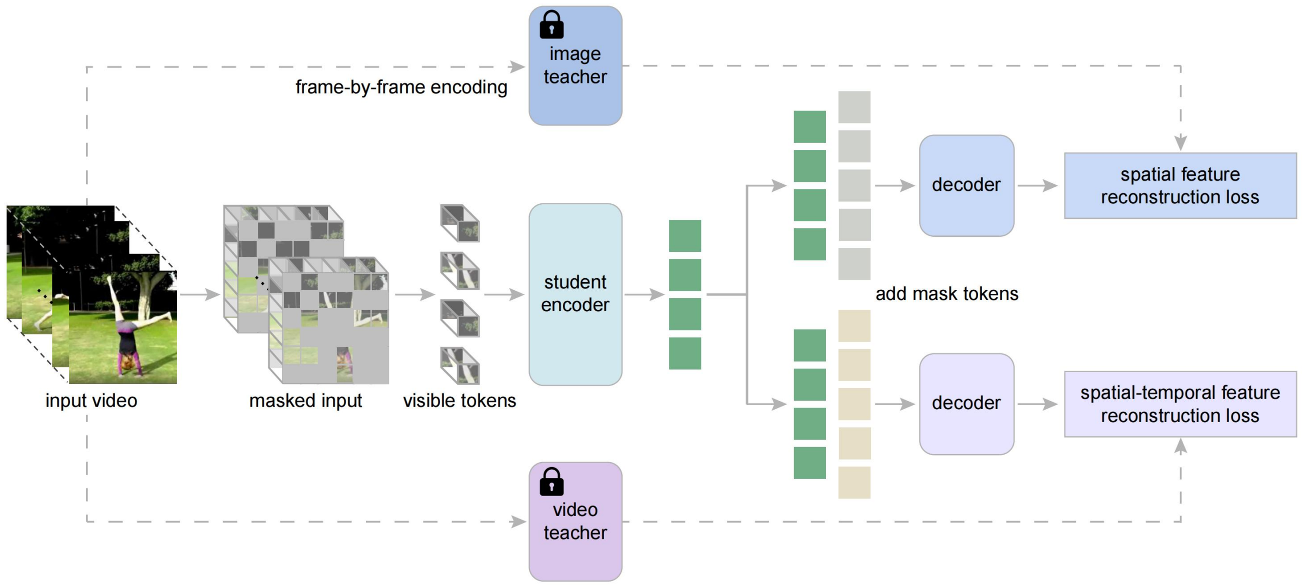
Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, Yu-Gang Jiang

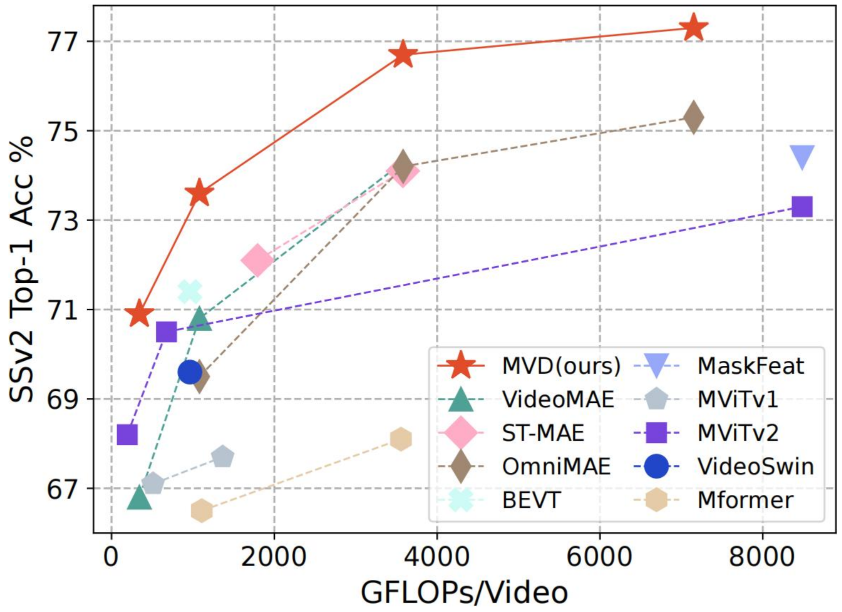
Poster: TUE-PM-209

Code repo: https://github.com/ruiwang2021/mvd

Overview

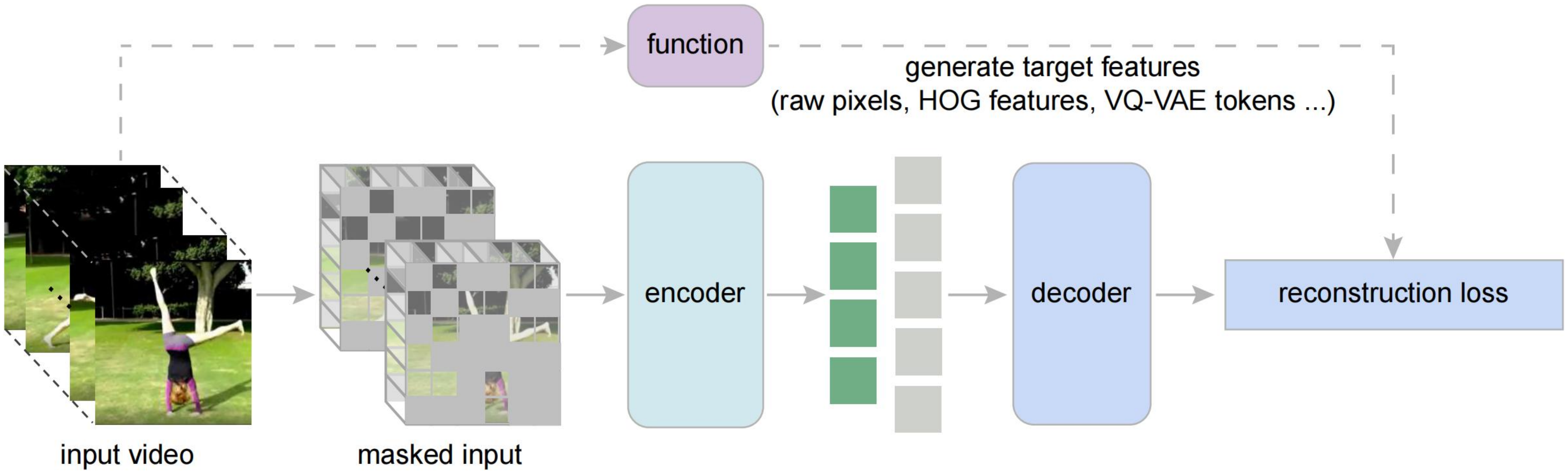
- MVD is a self-supervised video representation learning method.
- We study how to design better target features for masked video modeling.
- MVD reconstructs high-level features encoded by pretrained image&video models.
- MVD achieves SoTA on various video downstream tasks.





Masked Video Modeling (MVM)

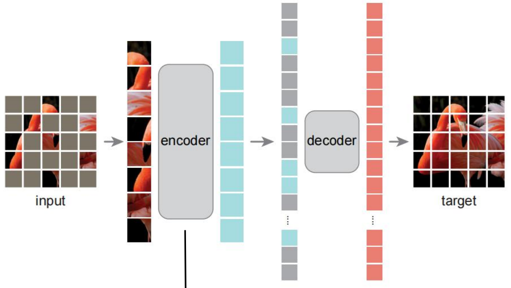
- A paradigm for self-supervised learning:
  - reconstructs features of masked input regions (patches).
- Previous works: reconstruct low-level features of masked patches.
  - raw pixels, HOG features ...



How to design better reconstruction targets for MVM?

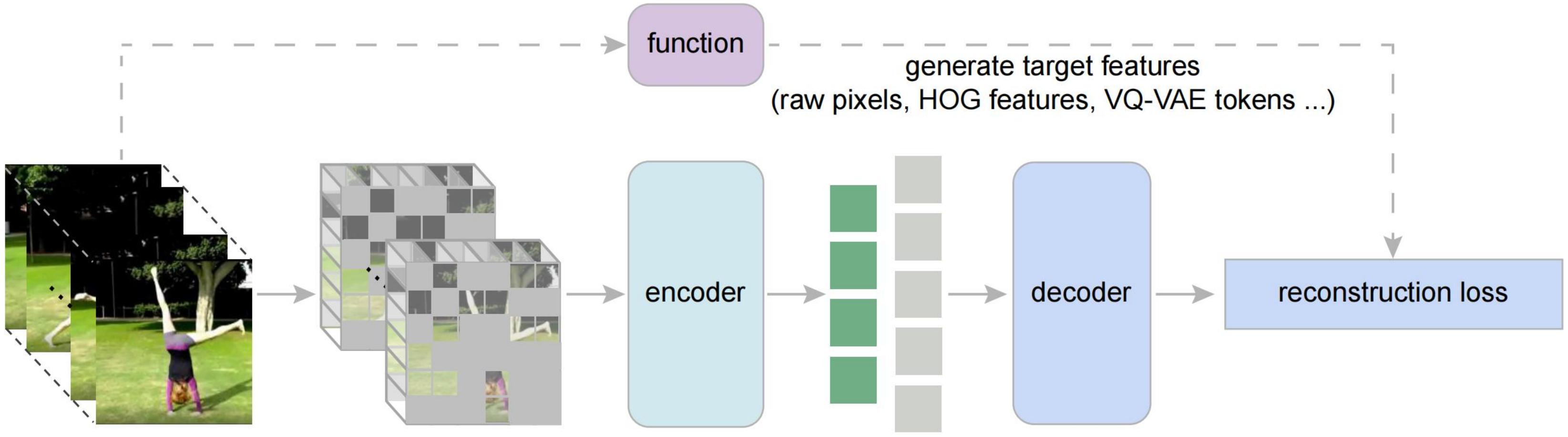
Masked Video Distillation

First Stage: MIM or MVM pretraining



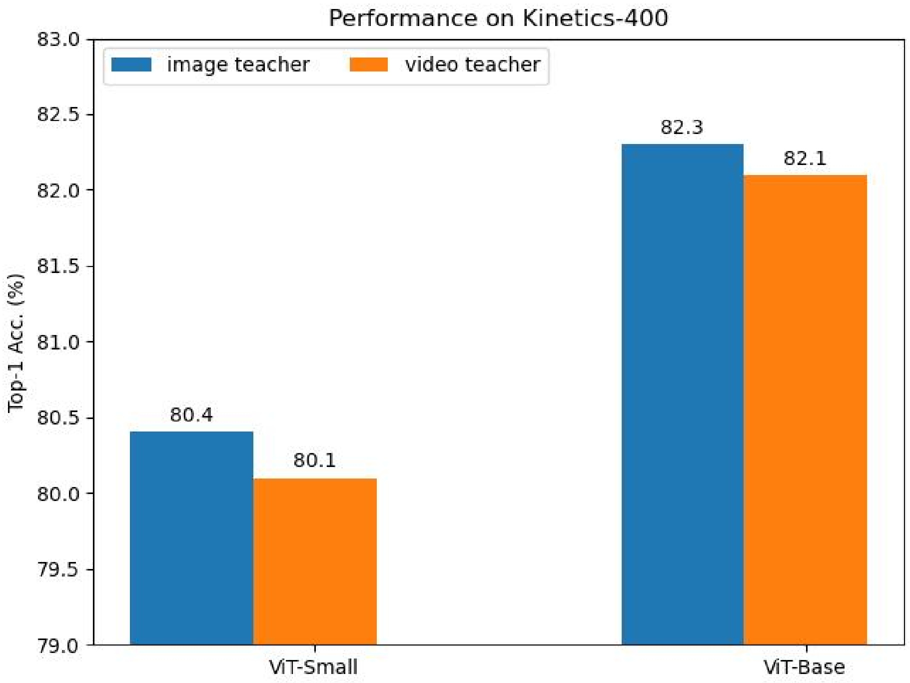
Pretrained Encoder as a **teacher**

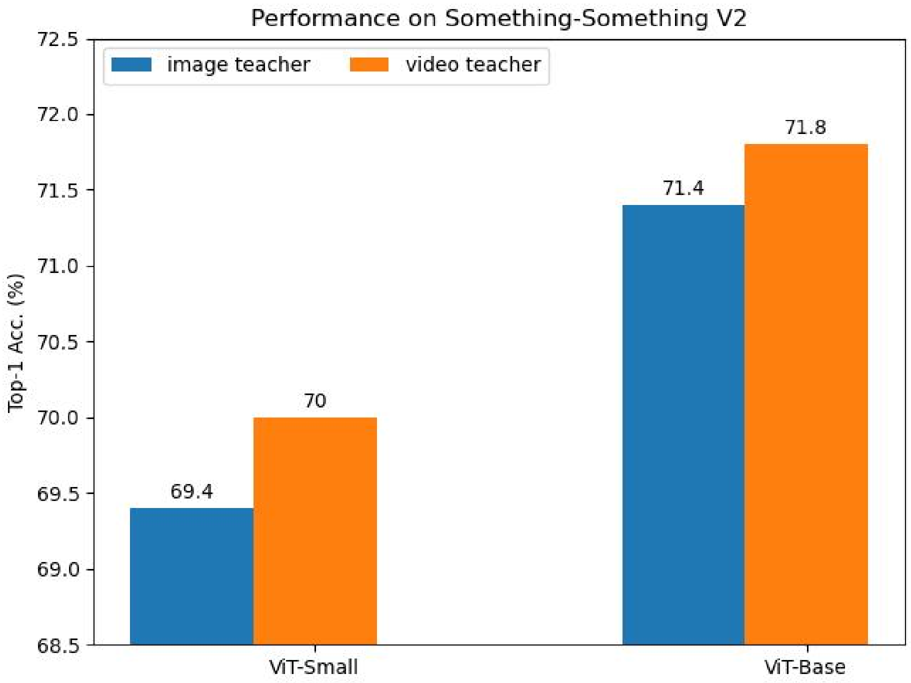
Second Stage:



High-level Features as targets of MVM

Comparison between different teachers

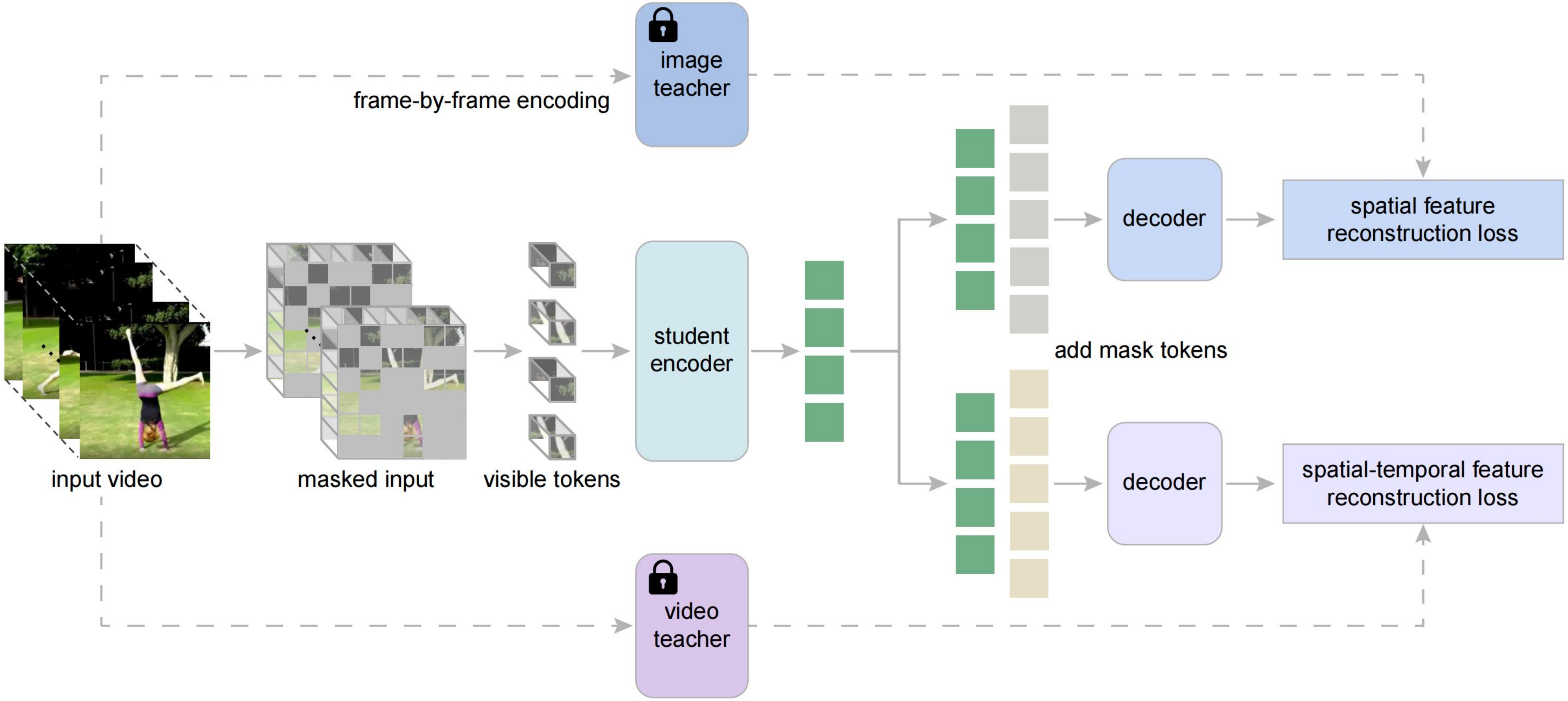




- Image teachers: ViT pretrained on IN-1K with Masked Image Modeling (MIM)
- Video teachers: ViT pretrained on K400 with Masked Video Modeling (MVM)

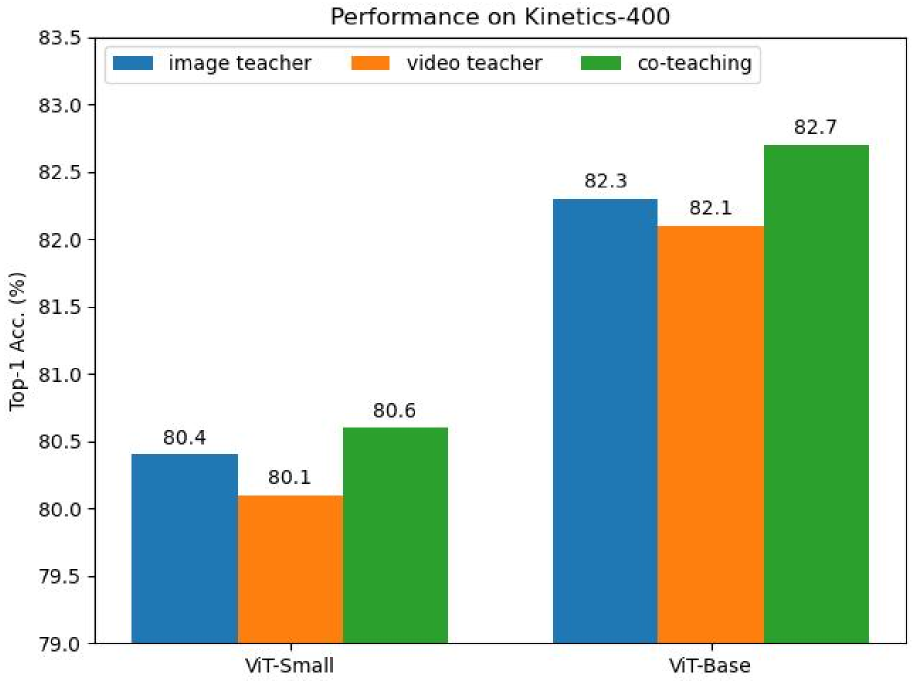
Students distilled with different teachers exhibit different properties

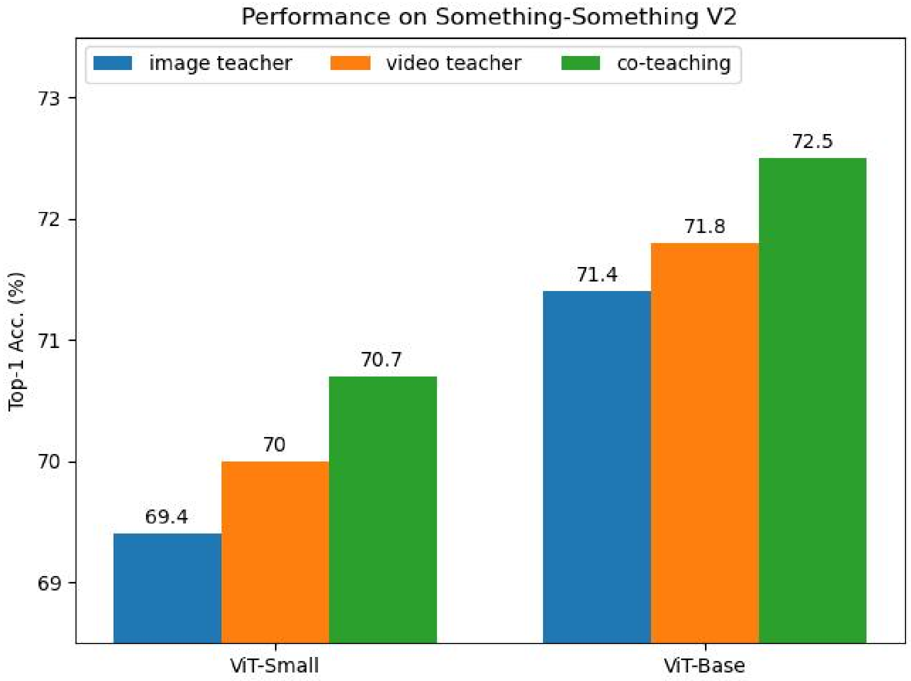
Spatial-temporal Co-teaching



Masked Video Distillation (MVD) Framework

Comparison between different teachers



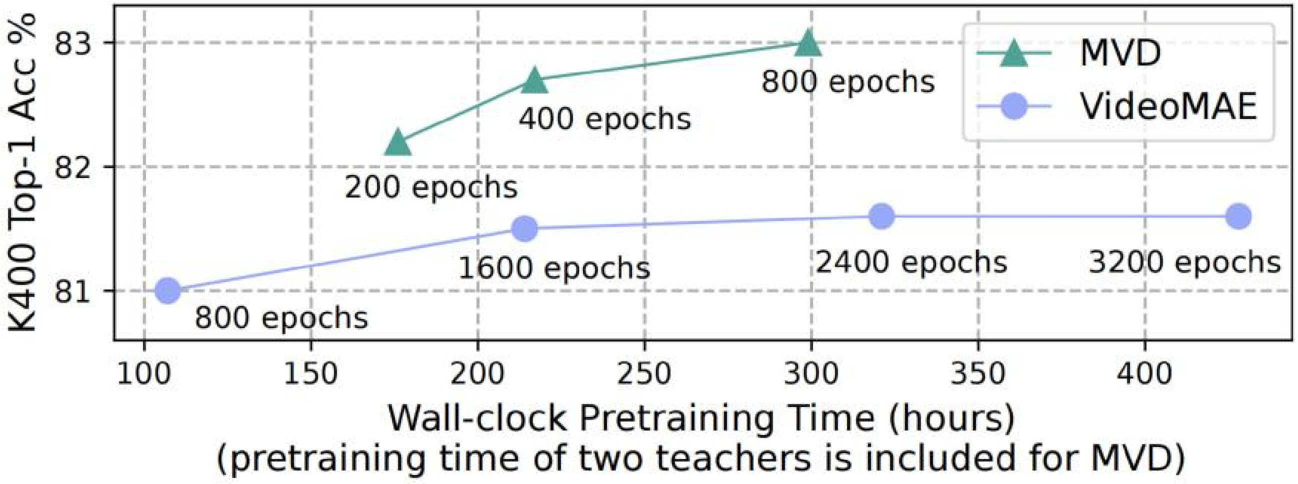


- Image teachers: ViT pretrained on IN-1K with Masked Image Modeling (MIM)
- Video teachers: ViT pretrained on K400 with Masked Video Modeling (MVM)

Co-teaching outperforms distillation with one single teacher in MVD

Comparison with VideoMAE

student	teacher	K400 top-1		SSv2 top-1	
		ViMAE	MVD	ViMAE	MVD
ViT-S	ViT-B	79.0	80.6 ↑**1.6**	66.4	70.7 ↑**4.3**
ViT-S	ViT-L	79.0	81.0 ↑**2.0**	66.4	70.9 ↑**4.5**
ViT-B	ViT-B	81.5	82.7 ↑**1.2**	69.7	72.5 ↑**2.8**
ViT-B	ViT-L	81.5	83.4 ↑**1.9**	69.7	73.7 ↑**4.0**
ViT-L	ViT-L	85.2	86.0 ↑**0.8**	74.0	76.1 ↑**2.1**

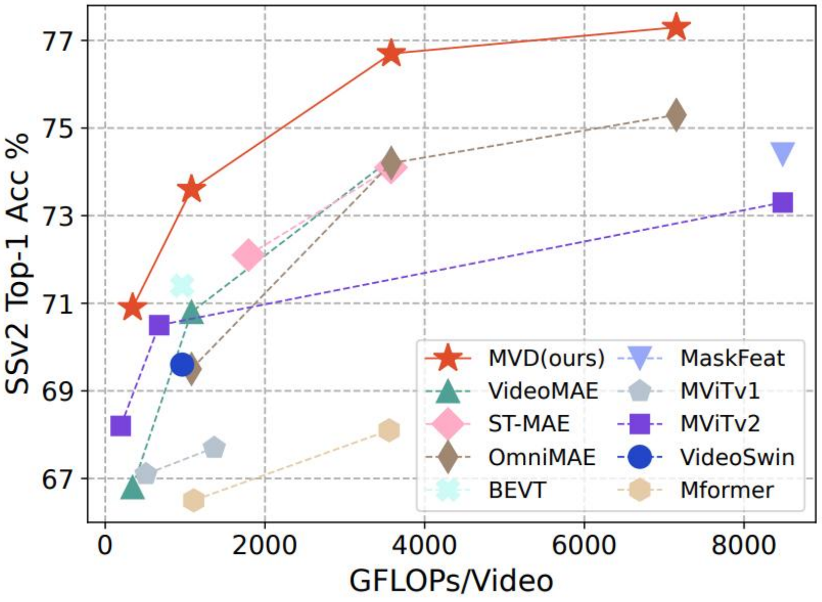


MVD outperforms VideoMAE by clear margins across different model scales

Comparison to State-of-the-art

method	extra data	top-1	top-5	GFLOPs	Param
supervised					
NL I3D R101 [65]	-	77.3	93.3	359×30	62
ip-CSN-152 [59]	-	77.8	92.8	109×30	33
SlowFast NL [22]	-	79.8	93.9	234×30	60
X3D-XL [20]	-	79.1	93.9	48×30	11
MViTv1-B [18]	-	80.2	94.4	170×5	37
VideoSwin-B [42]	IN-1K	80.6	94.6	282×12	88
Uniformer-B [36]	IN-1K	83.0	95.4	259×12	50
TimeSformer [4]	IN-21K	80.7	94.7	2380×3	121
Mformer-B [47]	IN-21K	79.7	94.2	370×30	109
Mformer-L [47]	IN-21K	80.2	94.8	1185×30	382
ViViT-L FE [1]	IN-21K	81.7	93.8	3980×3	N/A
VideoSwin-L [42]	IN-21K	83.1	95.9	604×12	197
self-supervised					
VIMPAC ViT-L [55]	HowTo100M	77.4	N/A	N/A×30	307
BEVT Swin-B [63]	IN-1K	81.1	N/A	282×12	88
MaskFeat MViT-S [67]	-	82.2	95.1	71×10	36
VideoMAE ViT-S [57]	-	79.0	93.8	57×15	22
VideoMAE ViT-B [57]	-	81.5	95.1	180×15	87
VideoMAE ViT-L [57]	-	85.2	96.8	597×15	305
VideoMAE ViT-H [57]	-	86.6	97.1	1192×15	633
ST-MAE ViT-B [21]	-	81.3	94.9	180×21	87
ST-MAE ViT-L [21]	-	84.8	96.2	598×21	304
ST-MAE ViT-H [21]	-	85.1	96.6	1193×21	632
OmniMAE ViT-B [25]	IN-1K	80.8	N/A	180×15	87
OmniMAE ViT-L [25]	IN1K+SSv2	84.0	N/A	597×15	305
OmniMAE ViT-H [25]	IN1K+SSv2	84.8	N/A	1192×15	633
MVD-S (Teacher-B)	IN-1K	80.6	94.7	57×15	22
MVD-S (Teacher-L)	IN-1K	81.0	94.8	57×15	22
MVD-B (Teacher-B)	IN-1K	82.7	95.4	180×15	87
MVD-B (Teacher-L)	IN-1K	83.4	95.8	180×15	87
MVD-L (Teacher-L)	IN-1K	86.0	96.9	597×15	305
MVD-L (Teacher-L) †	IN-1K	**86.4**	**97.0**	597×15	305
MVD-H (Teacher-H) †	IN-1K	**87.3**	**97.4**	1192×15	633

Kinetics-400



Something-Something v2

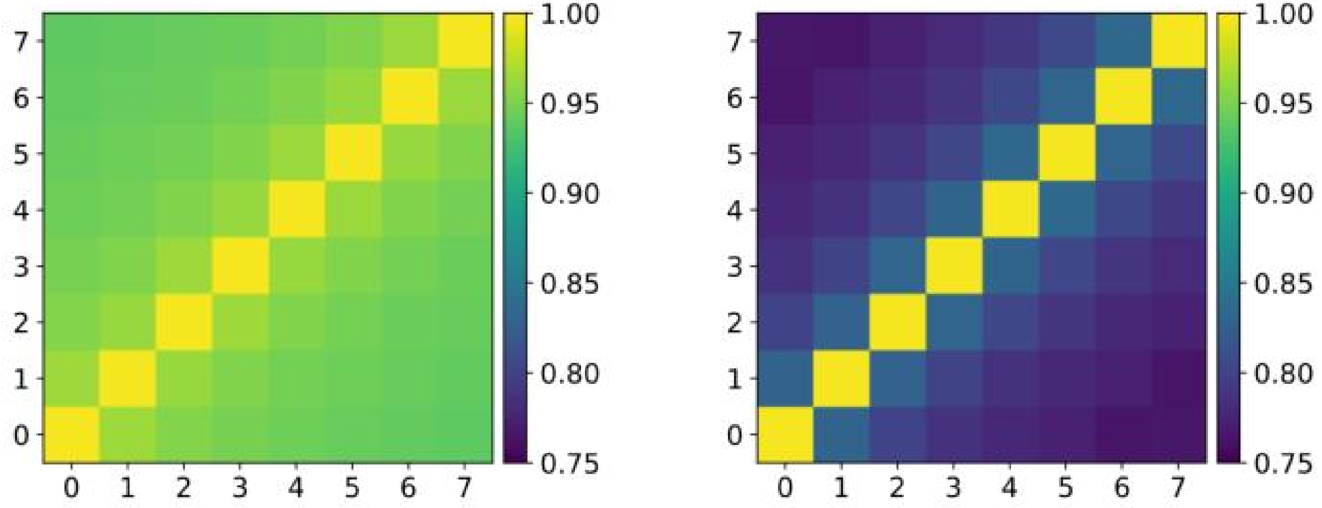
method	extra data	Param	UCF101	HMDB51
VideoMoCo R2+1D [50]	K400	15	78.7	49.2
MemDPC R2D3D [30]	K400	32	86.1	54.5
Vi^{2}CLR S3D [12]	K400	9	89.1	55.7
CORP Slow-R50 [35]	K400	32	93.5	68.0
CVRL Slow-R50 [53]	K400	32	92.9	67.9
CVRL Slow-R152 [53]	K600	328	94.4	70.6
ρBYOL Slow-R50 [24]	K400	32	94.2	72.1
VIMPAC ViT-L [60]	HowTo100M	307	92.7	65.9
VideoMAE ViT-B [61]	K400	87	96.1	73.3
MVD-B (Teacher-B)	IN-1K+K400	87	**97.0**	**76.4**
MVD-B (Teacher-L)	IN-1K+K400	87	**97.5**	**79.7**

UCF-101 & HMDB-51

method	extra data	extra labels	mAP	GFLOPs	Param
supervised					
SlowFast R101 [23]	K400	✓	23.8	138	53
MViTv2-B [41]	K400	✓	29.0	225	51
MViTv2-L [41]	IN-21K+K700	✓	34.4	2828	213
self-supervised					
MaskFeat MViT-L [73]	K400	✓	37.5	2828	218
VideoMAE ViT-B [63]	K400	✗	26.7	180	87
VideoMAE ViT-B [63]	K400	✓	31.8	180	87
VideoMAE ViT-L [63]	K400	✗	34.3	597	305
VideoMAE ViT-L [63]	K400	✓	37.0	597	305
VideoMAE ViT-H [63]	K400	✗	36.5	1192	633
VideoMAE ViT-H [63]	K400	✓	39.5	1192	633
ST-MAE ViT-L [22]	K400	✓	35.7	598	304
ST-MAE ViT-H [22]	K400	✓	36.2	1193	632
MVD-B (Teacher-B)	IN-1K+K400	✗	29.3	180	87
MVD-B (Teacher-B)	IN-1K+K400	✓	33.6	180	87
MVD-B (Teacher-L)	IN-1K+K400	✗	31.1	180	87
MVD-B (Teacher-L)	IN-1K+K400	✓	34.2	180	87
MVD-L (Teacher-L)	IN-1K+K400	✗	37.7	597	305
MVD-L (Teacher-L)	IN-1K+K400	✓	38.7	597	305
MVD-H (Teacher-H)	IN-1K+K400	✗	**40.1**	1192	633
MVD-H (Teacher-H)	IN-1K+K400	✓	**41.1**	1192	633

AVA v2.2

Visualization of teachers' features on videos



(a) image teacher. (b) video teacher.

Figure 3. **Feature similarity across different frames for different teacher models.** Similarity matrices are computed on the Kinetics-400 validation set.

Video teachers capture more temporal difference

Conclusion

- MVD is a two-stage masked video modeling framework.
- Students distilled with different teachers show different properties.
- Combining image and video teachers with co-teaching achieves higher performance.
- MVD outperforms previous MVM methods at different model scales.
- MVD achieves SoTA on various video downstream tasks.

Code & pretrained models are available on GitHub!
https://github.com/ruiwang2021/mvd