A. Additional Results

A.1. Reults of AVA Action Detection

As shown in Table 10, when transferred to the more complicated action detection task (AVA v2.2), MVD still shows remarkable improvement compared with previous methods. For example, without additional labels of K400, MVD with ViT-L outperforms VideoMAE by 3.4 to achieve 37.7 mAP. When we intermediately finetune the pretrained models on K400, MVD with ViT-L also achieves significant performance improvement (*i.e.*, 1.7 mAP) compared with VideoMAE. Finally, with a ViT-Huge model, MVD achieves 41.1 mAP, improving 1.6 over the prior state-of-the-art method.

method	extra data	extra labels	mAP	GFLOPs	Param
supervised					
SlowFast R101 [7]	K400	✓	23.8	138	53
MViTv2-B [11]	K400	✓	29.0	225	51
MViTv2-L [11]	IN21K+K700	✓	34.4	2828	213
self-supervised					
MaskFeat MViT-L [19]	K400	✓	37.5	2828	218
VideoMAE ViT-B [16]	K400	✗	26.7	180	87
VideoMAE ViT-B [16]	K400	✓	31.8	180	87
VideoMAE ViT-L [16]	K400	✗	34.3	597	305
VideoMAE ViT-L [16]	K400	✓	37.0	597	305
VideoMAE ViT-H [16]	K400	✗	36.5	1192	633
VideoMAE ViT-H [16]	K400	✓	39.5	1192	633
ST-MAE ViT-L [6]	K400	✓	35.7	598	304
ST-MAE ViT-H [6]	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

Table 10. **Comparison with previous works on AVA v2.2**. "Extra labels" denotes whether the pretrained models are intermediately finetuned on the pretraining video dataset using labels before transferred to AVA.

A.2. Ablation Study

Comparison with feature distillation. In our paper, we use a baseline method named per-token distillation based on previous feature distillation methods. In Table 11, per-token distillation with different inputs is compared with masked feature reconstruction. The results demonstrate that our MVD outperforms per-token distillation with full input or masked input on both K400 and SSv2.

method	masked input	top-1 accuracy	
		K400	SSv2
per-token distillation	✗	80.4	69.9
per-token distillation	✓	80.9	70.5
masked reconstruction	✓	**82.1**	**71.8**

Table 11. **Comparison with feature distillation without masked reconstruction.** "Masked input" denotes that tube masking with a masking ratio of 90% is applied on the input and only the visible tokens are fed to the student model. We distill ViT-B models from the video teacher for 400 epochs here.

Pretraining time comparison. In our paper, we find that MVD with a single video teacher achieves better accuracy with

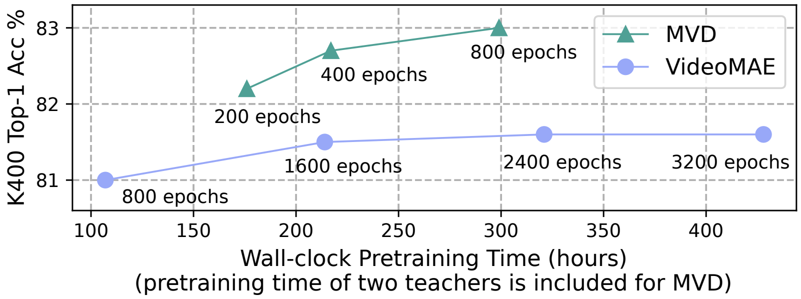


Figure 4. **Pretraining time comparison between MVD and VideoMAE.**

less training time, compared to VideoMAE trained for 1600 epochs. In Figure 4, we also compare the training time between VideoMAE and MVD where we train two teachers for 800 epochs. With distilling for 200 epochs, MVD outperforms VideoMAE trained for 1600 epochs by 0.7% with 38h less training time. For more epochs, MVD also achieves a better accuracy-time curve.

Ablation on the initialization of the students. We study whether to initialize the student model with weights from the teachers. As shown in Table 12, we observe that initializing the student with weights from the teachers does not bring significant improvements (even worse on SSv2). Therefore, we train the student model from scratch in MVD.

Student Init.	top-1 accuracy	
	K400	SSv2
from scratch	82.7	72.1
image teacher	82.9	71.9
video teacher	82.4	71.7

Table 12. **Ablations on the initialization of the student models.**

Ablation on the temporal size of 3D patches in the video teacher. Following previous works [1, 16, 18], we utilize the temporal patch size of 2 for both video teachers and students. In Table 13, we try to increase the temporal patch size of the video teacher and find that a temporal patch size 2 performs better.

patch size	top-1 accuracy	
	K400	SSv2
2×16 ×16	82.5	71.4
4×16 ×16	81.9	70.7

Table 13. **Ablation on the temporal size of 3D patches in the video teacher.**

A.3. Masked Feature Modeling for Image Models

We perform masked reconstruction of high-level features for the image ViT on ImageNet-1K. For masked feature modeling on the image data, only the image teacher in MVD can be used. As the results shown in Table 14, compared with the MAE baseline, masked feature distillation achieves 0.4% Top-1 accuracy gain on ImageNet-1K. When comparing the performance improvement against masked reconstruction of pixels between image models and video models, we observe that MVD achieves greater performance gains on video downstream tasks.

B. Implementation Details

B.1. Pretraining Experiments

We pretrain image teacher models on ImageNet-1K following the strategy in [8], and pretrain video teacher models on Kinetics-400 following the strategy in [16]. For the distillation stage in MVD, we distill student models with teacher models for 400 epochs on Kinetics-400 unless otherwise stated. The length of input videos is 16 frames during pretraining. We adopt

target	epoch	top-1 accuracy		
		IN-1K	K400	SSv2
pixels	1600	83.6	81.5	69.7
features	400	84.0 ↑**0.4**	82.7 ↑**1.2**	72.5 ↑**2.8**

Table 14. **Comparison with masked feature modeling for image models.** We distill ViT-B for 400 epochs here.

config	Kinetics-400
optimizer	AdamW [13]
base learning rate	1.5e-4
weight decay	0.05
optimizer momentum	β_1,β_2=0.9,0.95 [4]
batch size	1024 (S,B), 512 (L,H)
learning rate schedule	cosine decay [12]
warmup epochs	40
augmentation	MultiScaleCrop [17]
drop path	0.1 (S,B), 0.2(L,H)

Table 15. **Pretraining setting of MVD.**

config	Sth-Sth V2	Kinetics-400	UCF101	HMDB51
optimizer	AdamW			
base learning rate	1e-3(S), 5e-4(B,L)	1e-3	5e-4	1e-3
weight decay	0.05			
optimizer momentum	β_1, β_2=0.9, 0.999			
batch size	512	512	128	128
learning rate schedule	cosine decay			
warmup epochs	5			
training epochs	40 (S), 30 (B,L)	150 (S), 75 (B), 50 (L)	100	50
repeated augmentation	2			
flip augmentation	*no*	*yes*	*yes*	*yes*
RandAug [5]	(9, 0.5)			
label smoothing [14]	0.1			
mixup [21]	0.8			
cutmix [20]	1.0			
drop path [10]	0.1 (S,B), 0.3 (L,H)		0.2	0.2
dropout [9]	0.5 (L,H)	0.5 (L,H)	0.5	0.5
layer-wise lr decay [2]	0.7 (S),0.75 (B,L,H)	0.75	0.7	0.7

Table 16. **Fine-tuning setting of MVD.**

tube masking in [16] and the masking ratio in the distillation stage is 90%. We conduct experiments of pretraining on 32 NVIDIA V100 GPUs. The default setting of pretraining is presented in Table 15.

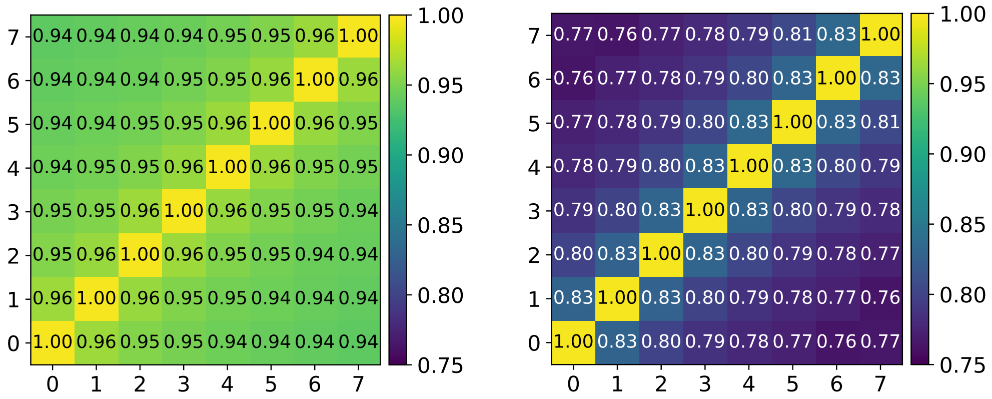
B.2. Finetuning Experiments

We transfer models pretrained by MVD on Kinetics-400 to video downstream tasks with the default setting in Table 16.

Kinetics experiments. When finetuning on Kinetics-400, we adopt the dense sampling following [3,7] and the default length of input videos is 16 frames. For inference, we use 3 spatial crops × 5 temporal clips.

Something-Something v2 experiments. During finetuning on Something-Something v2, we adopt the uniform sampling following [17] and the default length of input videos is 16 frames. For inference, we use 3 spatial crops × 2 temporal clips.

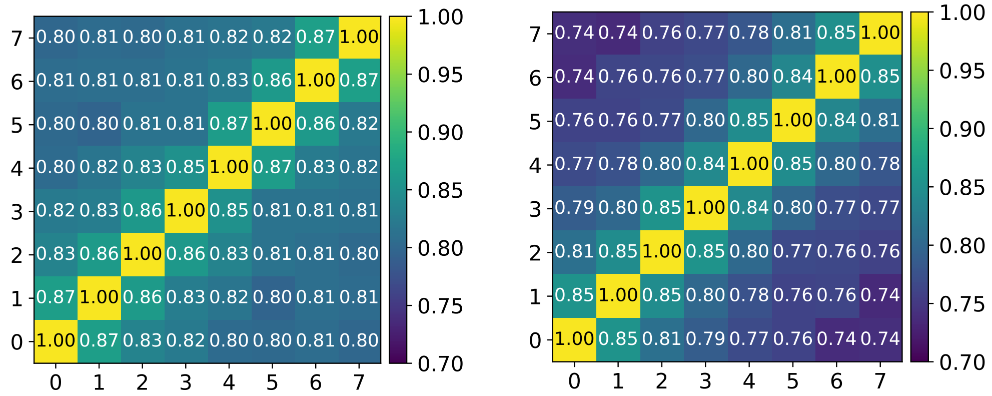
UCF101 and HMDB51 experiments. For finetuning on UCF101 and HMDB51, we adopt the dense sampling and the default length of input videos is 16 frames. For inference, we use 3 spatial crops × 5 temporal clips. On UCF101 and



(a) image teacher.

(b) video teacher.

Figure 5. **Feature similarity across different frames for different teacher models.** Similarity matrices are computed on the Kinetics-400 validation set. The numbers in the grid are the values of cosine similarity between two frame features.



(a) student distill from the image teacher.

(b) student distill from the video teacher.

Figure 6. **Feature similarity across different frames for student models distilled from different teacher models.** Similarity matrices are computed on the Kinetics-400 validation set.

HMDB51, we follow the commonly used protocols and evaluate our method across all 3 train/val splits.

AVA experiments. When finetuning on AVA v2.2, following [16], we adopt the detection architecture in [7] and the detected person boxes from AIA [15]. The default length of input videos is 16 frames. We also use the default finetuning setting in [16] for a fair comparison.

C. Visualization

C.1. Analysis of temporal dynamics

In our paper, to quantify the temporal dynamics that models capture from the input video, we study the similarity between feature maps across different frames of each input video clip via the cosine similarity.

Analysis of features encoded by different teachers. The properties of target features generated by different teachers may influence the performance of students on different downstream tasks. As similarity matrices shown in Figure 5, for image teachers, the feature maps of different frames are almost the same. However, for video teachers, the features of different frames have larger differences. This indicates that video teachers capture more temporal difference. Therefore, students distilled from video teachers can learn stronger temporal dynamics and perform better on temporally-heavy downstream tasks.

Analysis of features encoded by students distilled from different teachers. To study what students learn from different teachers, we visualize the feature similarity across different frames for student models. As results shown in Figure 6, we observe that (a) for the student distilled from the image teacher, the features of different frames have larger differences compared with those encoded by the image teacher. This indicates that students can also learn some temporal dynamics from the masked reconstruction of spatial features on videos. (b) For the student distilled from the video teacher, the features of different frames have larger differences compared with those encoded by the student distilled from the image teacher. This demonstrates that students learn stronger temporal dynamics from video teachers.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, 2021. 2

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT pre-training of image transformers. In *ICLR*, 2022. 3

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 3

[4] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 3

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR workshops*, 2020. 3

[6] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022. 1

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 1, 3, 4

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021. 2

[9] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 3

[10] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 3

[11] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 2022. 1

[12] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 3

[13] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 3

[14] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3

[15] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *ECCV*, 2020. 4

[16] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022. 1, 2, 3, 4

[17] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE TPAMI*, 2018. 3

[18] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *CVPR*, 2022. 2

[19] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 1

[20] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 3

[21] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 3