Appendices

A. Datasets

Video Understanding. For video action recognition, we pre-train and evaluate the model on UCF-101 [57] which contains 9.5k/3.5k train/val videos. For surgical video action detection, we pre-train and evaluate the model on OR-AR [35] which consists of 820 long videos captured in surgical operating rooms. All videos have 9 temporal workflow phase labels.

Image Understanding. For image understanding tasks, we pre-train the model on ScanNet [20] which contains 2.5M RGB-D frames from 1513 video sequences. For evaluation, we use ScanNet and NYUv2 [49]. NYUv2 contains 1449 densely labeled images from indoor scenes captured with Microsoft Kinect RGB-D camera. We use the official split of 795 images in training set and 654 images in test set.

B. Additional Implementation Details

B.1. Network Architecture

Video Understanding. We follow the network architecture presented in VideoMAE [60] for video based pre-training. The encoder part of the network is vision-transformer base (ViT-B) while the decoder consists of 4 blocks with six multi-head attentions in each. The width of the decoder is set to half of the encoder dimension i.e. 384-d. We use two fully connected layers (one for each modality) on top of the decoder for reconstruction. During fine-tuning, we remove the decoder and add the fully connected layer for prediction. Specifically for surgical video action detection, we follow the fine-tuning method in [35,56]. For evaluation, we fine-tune the model in two stages. In the first stage, we add a fully connected layer on top of the encoder for predicting clip-wise phases. In the second stage, we first extract the features from the encoder and then train a temporal model (Bi-GRU) for detecting phases in a full video.

Image Understanding. We follow MAE [33] for the network design. Our modality-specific encoders are based on ViT-B while the decoder part consists of 8 blocks with 16 multi-head attentions in each block. The width is set to 512. Similarly, we use two fully-connected layers (one for each modality) on top of the decoder for reconstruction. For fine-tuning, we mostly follow MultiMAE [5] for task specific head. More specifically, we use segmentation head based on ConvNeXt architecture [46] and depth-estimation head based on DPT [55].

B.2. Pre-training and Fine-tuning Details

For video understanding, we report the pre-training setting in Table 9 and the fine-tuning setting in Table 10. Moreover, we report the pre-training setting on ScanNet in Table 12 and the transfer setting for semantic segmentation and depth estimation task in Table 13 and Table 14 respectively.

Configuration	OR-AR [35]	UCF-101 [57]
Optimizer	AdamW	
Optimizer betas	{0.9, 0.95}	
Base learning rate	1e-4	1e-3
Weight decay	5e-2	
Learning rate schedule	cosine decay	
gradient clipping	0.02	None
Warmup epochs	40	
Epochs	1600	100 or 800 (from scratch)
Flip augmentation	True	True
Augmentation	MultiScaleCrop	
Num of Frames	16	
sampling rate	4.0	
α	1.0	1.0
β	0.5	0.1
γ	0.2	0.01
η	0.1	0.01

Table 9. Pre-training setting on OR-AR [35] and UCF-101 [57] datasets.

Configuration	OR-AR [35]	UCF-101 [57]
Optimizer	AdamW	
Optimizer betas	{0.9, 0.95}	
Base learning rate	6e-4	1e-3
Weight decay	5e-2	
Learning rate schedule	cosine decay	
Warmup epochs	5	
Epochs	75	100
Flip augmentation	True	True
Mixup	None	0.8
CutMix	None	1.0
drop path	0.1	0.2
drop out	0.0	0.5
Layer-wise lr decay	0.65	0.70
Temporal Model learning rate	1e-3	None
Temporal Model Epochs	15	None

Table 10. Fine-tune setting on OR-AR [35], UCF-101 [57] datasets.

Strategy and Ratio	OR-AR [35]	UCF-101 [57]	ScanNet [20]
RGB Masking strategy	Tube	SurgMAE [41]	Random
RGB Masking ratio	0.9	0.9	0.8
Depth Masking strategy	Tube	Random	Random
Depth Masking ratio	0.9	0.9	0.8

Table 11. Masking strategies during pre-training.

B.3. Masking Strategy

Table 11 shows the different masking strategies for RGB-D modalities during pre-training.

Configuration	ScanNet [20]
Optimizer	AdamW
Optimizer betas	{0.9, 0.95}
Base learning rate	1e-4
Weight decay	5e-2
Learning rate schedule	cosine decay
Stage-1 epochs	20
Stage-2 epochs	100
Augmentation	Gaussian Blur, ColorJitter
α	0.1
β	1.0

Table 12. Pre-training setting on ScanNet [20].

Configuration	ScanNet [20]	NYUv2 [49]
Optimizer	AdamW	
Optimizer betas	{0.9, 0.999}	
Base learning rate	1e-4	
Layer-wise lr decay	0.75	
Weight decay	5e-2	
Learning rate schedule	cosine decay	
Warmup epochs	1	
Warmup learning rate	1e-6	
Drop path	0.1	
Epochs	50	200
Input resolution	240 x 320	640 x 640
Color jitter	✗	✓
RandomGaussianBlur	✓	✗
RandomHorizontalFlip	✓	✗

Table 13. Fine-tune setting on ScanNet [20] and NYUv2 [49] for 2D semantic segmentation.

Configuration	NYUv2 [49]
Optimizer	AdamW
Optimizer betas	{0.9, 0.999}
Base learning rate	1e-4
Weight decay	1e-4
Learning rate schedule	cosine decay
Warmup epochs	100
Warmup learning rate	1e-6
Epochs	2000
Batch Size	128
Layer-wise lr decay	0.75
Input resolution	256 x 256
Augmentation	RandomCrop, Color jitter

Table 14. Fine-tune setting for NYUv2 [49] depth estimation.