A1. More Implementation Details

More Backbone Details. For our AdaMV-MoE, the patch sizes are 8 and 16 for classification and object detection & instance segmentation tasks, respectively. We follow [16] to use a constant window scale of 2 on the OD & IS tasks to save computational outlay. The patches are extracted from images in an overlapping way [75] with a size of 2 for all tasks. To extract patches, we prepend convolutional layers before the Transformer's encoder,

Furthermore, similar to [58], the stochastic depth technique [35] is adopted with a probability of 0.1 during training. In addition, Table A6 summarizes the model configurations of *Large Dense* baselines.

Table A6. Detailed model sizes of (*Large Dense*) ViT variants.

Backbones	# Transformer Layers	# Attention Heads	Hidden Dimension	MLP Dimension
ViT-Small*	6	6	576	2304
ViT-Small	12	6	576	2304
ViT-Base	12	6	900	3600
UViT-Base	18	6	576	2304

A2. More Experiment Results

Extra Studies of AES. More comparisons between using a fixed number of activated experts and our AES are conducted with the ViT-Small* backbone. The multi-task vision recognition performance is collected in Table A7. We see that AES yields overall better results since it customizes network capacity for different tasks in MTL.

Table A7. Multi-task vision recognition performance of ViT-Small* with a different number of selected experts or our proposed Adaptive Expert Selection (AES).

# Experts Activated	Classification	Object Detection	Instance Segmentation
	Accuracy(%)	AP(%)	AP^{mask}(%)
$k = 2$	72.04	38.61	35.23
$k = 3$	72.66	39.03	35.68
$k = 4$	72.90	38.98	35.74
AES	**72.99**	**39.04**	**35.76**

Extra Studies on Training Iterations. We vary the number of training iterations for AdaMV-MoE and MTL-ViT, and report the performance in Table A8. We see that AdaMV-MoE enjoys a better convergence.

More Specialization Results. Similar to Figure 4, we provide the class-wise expert usage for object detection and instance segmentation tasks in Figure A8. We see the expert 4 is most frequently used, while other experts' activation is more correlated with class types.

Table A8. Multi-task vision recognition performance of ViT-Small* trained with a different number of iterations.

# Iterations	Methods	Classification	Object Detection	Instance Segmentation
		Accuracy(%)	AP(%)	AP^{mask}(%)
100K	MTL-ViT	54.61	35.17	33.18
	AdaMV-MoE	58.84	36.65	34.13
300K	MTL-ViT	68.36	36.46	34.25
	AdaMV-MoE	69.71	37.50	34.93
500K	MTL-ViT	68.71	36.98	34.60
	AdaMV-MoE	72.99	39.04	35.76

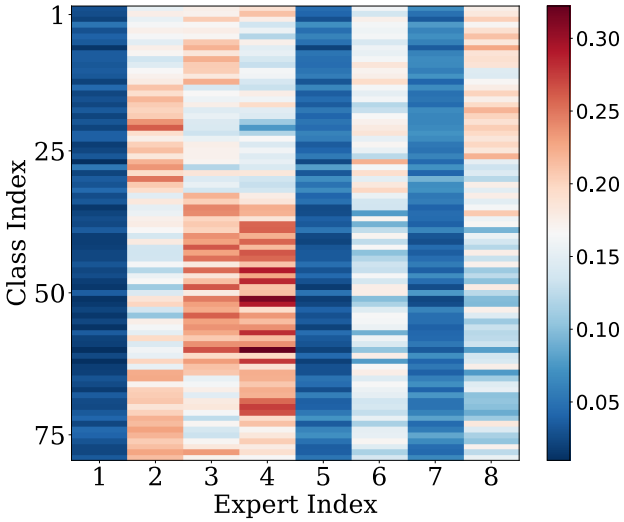

Figure A8. Analysis on the class-level routing specialization of OD & IS, produced by AdaMV-MoE with ViT-Small*.