Appendix

A Multi-Window Multi-Head Attention

```
def WinAttention(Q, K, V, win_i):
 n, d_k = Q.shape[-2:]
 # partition inputs along patch dimension
 # into non-overlapping windows
 Q = Q.reshape(-1, win_i, d_k)
 K = K.reshape(-1, win_i, d_k)
 V = V.reshape(-1, win_i, d_k)
 # compute self-attention
 X = softmax(Q.(K.transpose()) / sqrt(d_k)).V
 # reshape results
 X = X.reshape(-1, n, d_k)
 return X
```

Figure 5: Pseudocode for WinAttention

B Experimental Details and Hyperparameters

In this section, we provide additional experimental details. Apart from AudioSet, all other datasets are obtained directly from the HEAR [3], where they are pre-processed to 16000 Hz and distributed in a standard format.

Similar to [32], our effective learning rate (lr_{eff}) depends on the base learning rate (lr_{base}) and the batch size as follows: $lr_{\text{eff}} = lr_{\text{base}} * \frac{\text{batch size}}{256}$. In early experiments, we did not find strong augmentations at pre-training time to improve downstream performance, hence no augmentations are used. For more details, refer to Table 3. As previously mentioned, hear-eval-kit[4] was used for downstream experiments, and along with the details provided here should allow for consistent, reproducible downstream experimentation.

Table 3: **Pre-training (PT) and Downstream (FT) hyperparameters**. *: For ViT-L and ViT-H based models, smallest batch size that didn't give OOM was used.

Configuration	AS-5k Pre-training	Downstream
Optimizer	AdamW	Adam
Optimizer momentum	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.95$
Weight decay	0.05	N/A
Base learning rate	0.000015	0.0001
Learning rate schedule	linear-warmup + cosine decay	fixed
Minimum learning rate	0.0	0.0001
Dropout	0.	0.25
Warm-up epochs	10	N/A
Epochs	100	500
Early Stopping	N/A	20
Batch size	1024^*	1024
Accelerators	8x TPU-v3 cores	1 Nvidia-A40

The code for feature extraction and running downstream experiments for our default configurations as well as the corresponding pre-trained weights can be found at https://github.com/10997NeurIPS23/10997_mwmae.

[3] https://hearbenchmark.com/hear-tasks.html

[4] https://github.com/hearbenchmark/hear-eval-kit

C Detailed Ablation Results

Table 4: Results from Patch size ablation experiments. ViT-B encoder was used for all experiments. n denotes total number of patches, and h denotes the number of attention heads in each decoder transformer block.

Model	BO	CD	ESC-50	LC	Mri-S	Mri-T	NS-5h	SC-5h	F50K	VL	$s(m)$
Patch Size=(8$\times$16), n=125, h=4											
MAE	94.9±0.8	70.2±0.3	80.4±0.5	66.0±0.3	97.4±0.1	97.7±0.1	65.9±0.7	88.9±0.5	49.4±0.1	40.6±0.5	85.9±0.3
MW-MAE	95.9±0.5	72.3±0.2	81.2±0.3	68.4±0.3	97.3±0.1	97.8±0.1	67.4±0.8	90.0±0.3	50.8±0.1	41.9±0.5	88.0±0.2
Patch Size=(4$\times$16), n=250, h=8											
MAE	96.2±0.3	72.2±0.2	80.9±0.4	67.3±0.3	97.4±0.1	98.3±0.1	68.3±0.4	89.4±0.3	50.4±0.1	43.1±0.9	88.1±0.2
MW-MAE	96.0±0.5	73.1±0.3	81.2±0.4	68.8±0.2	97.4±0.1	97.9±0.1	69.3±0.6	90.9±0.2	51.2±0.2	44.2±0.9	89.2±0.2
Patch Size=(8$\times$8), n=250, h=8											
MAE	96.1±0.6	72.5±0.2	81.3±0.2	66.0±0.3	97.5±0.1	98.1±0.0	68.5±0.7	89.5±0.4	50.2±0.1	42.3±0.5	87.7±0.2
MW-MAE	96.3±0.4	73.0±0.1	82.6±0.3	69.3±0.3	97.5±0.1	98.1±0.1	70.3±0.8	90.5±0.1	51.4±0.1	42.3±0.5	89.4±0.1
Patch Size=(4$\times$8), n=500, h=12											
MAE	96.7±0.2	71.3±0.3	79.0±0.4	67.8±0.3	97.7±0.0	98.5±0.0	68.7±0.4	89.0±0.4	49.8±0.2	39.2±0.7	87.2±0.1
MW-MAE	95.6±0.7	74.1±0.2	81.9±0.3	70.1±0.3	97.6±0.1	98.2±0.1	72.0±0.7	91.2±0.3	51.6±0.1	44.0±0.8	90.3±0.2
Patch Size=(5$\times$5), n=640, h=16											
MAE	96.0±0.4	70.9±0.2	80.9±0.4	67.6±0.4	97.6±0.1	98.4±0.0	69.3±0.4	88.4±0.3	49.3±0.2	37.7±0.6	86.8±0.2
MW-MAE	96.6±0.4	73.8±0.4	82.0±0.3	70.1±0.4	97.5±0.1	98.3±0.1	72.9±0.5	91.7±0.2	51.3±0.1	44.2±0.6	90.6±0.1

Table 5: Effect of encoder size on performance. Patch size of 4$\times$16 was used for all experiments.

Model	BO	CD	ESC-50	LC	Mri-S	Mri-T	NS-5h	SC-5h	F50K	VL	$s(m)$
Encoder=ViT-T											
MAE	95.6±0.5	63.2±0.2	70.1±0.5	64.6±0.3	97.1±0.1	97.4±0.1	66.4±0.7	74.3±0.8	41.6±0.1	26.4±0.6	77.6±0.3
MW-MAE	93.3±1.0	64.4±0.2	71.9±0.5	65.5±0.3	97.1±0.1	97.6±0.1	68.1±0.4	77.0±0.6	43.4±0.1	28.6±1.1	79.0±0.3
Encoder=ViT-M											
MAE	95.2±0.7	69.5±0.2	77.8±0.3	67.4±0.3	97.4±0.0	98.0±0.1	66.6±0.7	88.0±0.4	48.1±0.1	38.3±0.8	85.3±0.2
MW-MAE	95.9±0.3	71.8±0.3	80.3±0.4	69.7±0.1	97.2±0.1	97.8±0.1	68.1±0.5	88.8±0.6	49.6±0.1	39.8±0.8	87.5±0.2
Encoder=ViT-B											
MAE	96.2±0.3	72.2±0.2	80.9±0.4	67.3±0.3	97.4±0.1	98.3±0.1	68.3±0.4	89.4±0.3	50.4±0.1	43.1±0.9	88.1±0.2
MW-MAE	96.0±0.5	73.1±0.3	81.2±0.4	68.8±0.2	97.4±0.1	97.9±0.1	69.3±0.6	90.9±0.2	51.2±0.2	44.2±0.9	89.2±0.2
Encoder=ViT-L											
MAE	95.8±0.6	72.4±0.1	79.7±0.3	66.8±0.4	97.5±0.1	98.2±0.1	69.5±0.6	90.9±0.2	50.7±0.1	43.6±0.4	88.3±0.2
MW-MAE	95.7±0.5	75.5±0.2	82.5±0.5	70.1±0.3	97.4±0.0	98.1±0.1	70.7±0.6	93.2±0.1	53.3±0.1	51.9±0.8	92.3±0.2
Encoder=ViT-H											
MAE	96.8±0.2	71.1±0.2	78.3±0.4	67.1±0.2	97.5±0.0	98.5±0.0	67.6±0.6	89.6±0.1	49.5±0.2	40.0±0.7	86.9±0.1
MW-MAE	96.8±0.2	74.8±0.1	81.6±0.4	69.5±0.4	97.4±0.0	98.2±0.1	70.8±0.5	92.4±0.2	52.1±0.1	47.5±0.6	91.1±0.2

Table 6: Effect of decoder depth on downstream performance. ViT-B encoder, patch size of 4$\times$16 were used for each experiment.

Model	BO	CD	ESC-50	LC	Mri-S	Mri-T	NS-5h	SC-5h	F50K	VL	$s(m)$
***depth*=1**											
MAE	96.4±0.2	69.8±0.3	78.9±0.3	67.4±0.3	97.4±0.1	97.9±0.1	66.4±0.8	88.5±0.2	49.4±0.2	39.0±1.1	86.1±0.2
MW-MAE	96.6±0.5	72.4±0.2	79.0±0.4	68.7±0.3	97.5±0.1	98.0±0.1	68.8±0.5	90.2±0.3	50.6±0.1	39.1±0.8	87.8±0.2
***depth*=2**											
MAE	96.8±0.3	71.3±0.3	78.8±0.2	68.8±0.2	97.4±0.1	98.2±0.0	67.2±0.6	90.0±0.2	49.6±0.2	39.4±0.7	87.3±0.1
MW-MAE	96.0±0.7	73.1±0.2	79.4±0.3	69.2±0.3	97.4±0.1	98.2±0.1	69.0±0.6	90.6±0.2	50.7±0.2	40.1±0.6	88.3±0.3
***depth*=4**											
MAE	96.2±0.3	72.2±0.2	80.9±0.4	67.3±0.3	97.4±0.1	98.3±0.1	68.3±0.4	89.4±0.3	50.4±0.1	43.1±0.9	88.1±0.2
MW-MAE	96.0±0.5	73.1±0.3	81.2±0.4	68.8±0.2	97.4±0.1	97.9±0.1	69.3±0.6	90.9±0.2	51.2±0.2	44.2±0.9	89.2±0.2
***depth*=8**											
MAE	96.3±0.3	71.7±0.3	81.6±0.4	67.4±0.3	97.4±0.0	98.1±0.1	67.8±0.7	89.9±0.3	50.8±0.2	43.4±0.6	88.2±0.1
MW-MAE	96.2±0.5	73.2±0.2	82.2±0.4	69.7±0.3	97.3±0.0	98.1±0.1	69.4±0.5	91.3±0.2	52.0±0.2	44.7±0.8	89.9±0.2

Table 7: Amount of pre-training dataset used v/s downstream performance.

Model	BO	CD	ESC-50	LC	Mri-S	Mri-T	NS-5h	SC-5h	F50K	VL	$s(m)$
10% of AS-5k											
MAE	93.6±0.7	51.3±0.2	49.5±0.3	48.4±0.4	97.1±0.1	96.4±0.1	61.1±0.7	70.4±0.9	29.7±0.2	17.3±0.5	63.3±0.2
MW-MAE	94.1±0.3	63.9±0.3	67.1±0.3	60.5±0.2	97.3±0.1	97.6±0.0	64.4±0.5	82.0±0.4	40.9±0.2	30.1±1.1	77.2±0.3
25% of AS-5k											
MAE	96.2±0.6	57.5±0.3	64.9±0.4	56.9±0.3	97.4±0.1	97.5±0.1	65.0±0.6	79.3±0.4	39.2±0.1	24.2±0.7	73.6±0.2
MW-MAE	96.1±0.5	68.0±0.2	75.5±0.4	67.2±0.3	97.3±0.1	98.0±0.1	65.9±0.4	86.5±0.2	46.4±0.1	35.7±0.6	83.8±0.2
50% of AS-5k											
MAE	97.2±0.3	65.5±0.3	74.1±0.3	64.3±0.3	97.5±0.1	98.1±0.1	67.0±0.6	85.3±0.6	45.1±0.1	32.4±0.8	81.9±0.2
MW-MAE	95.9±0.5	70.9±0.2	79.1±0.3	69.1±0.4	97.4±0.1	98.1±0.1	68.4±0.7	88.5±0.2	49.1±0.1	39.5±0.5	87.0±0.2
75% of AS-5k											
MAE	95.3±0.5	70.2±0.2	79.0±0.3	67.4±0.2	97.4±0.1	98.1±0.1	67.4±0.6	88.8±0.3	49.2±0.1	39.5±0.7	86.2±0.2
MW-MAE	96.0±0.5	72.6±0.3	80.5±0.4	69.5±0.3	97.4±0.1	97.9±0.1	68.3±0.4	89.9±0.2	50.5±0.1	41.7±0.8	88.4±0.2
100% of AS-5k											
MAE	96.2±0.3	72.2±0.2	80.9±0.4	67.3±0.3	97.4±0.1	98.3±0.1	68.3±0.4	89.4±0.3	50.4±0.1	43.1±0.9	88.1±0.2
MW-MAE	96.0±0.5	73.1±0.3	81.2±0.4	68.8±0.2	97.4±0.1	97.9±0.1	69.3±0.6	90.9±0.2	51.2±0.2	44.2±0.9	89.2±0.2

D High Resolution PWCCA Visualizations for better viewing

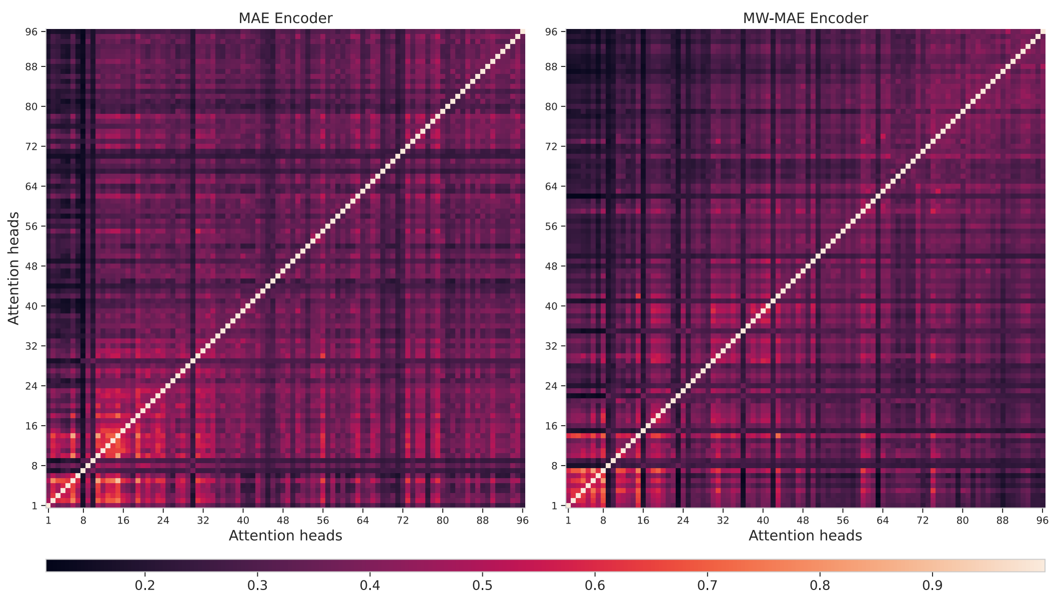

Figure 6: Encoder PWCCA correlation matrices

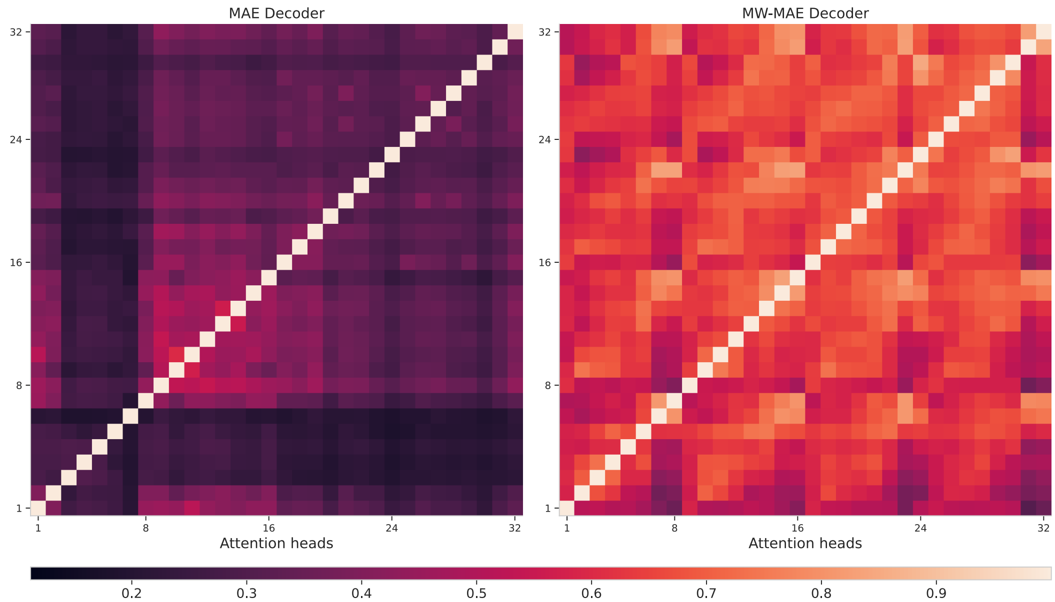

Figure 7: Decoder PWCCA correlation matrices

E Limitations

The direct limitations of our work are:

1. Pre-training data scale: As opposed to text corpus used in NLP [13] as well as speech representations [10, 14], AudioSet is several order of magnitudes smaller. While MW-MAEs demonstrate good performance characteristics in low-data scenarios, analysis on larger scales of data is definitely warranted.
2. Computational demands: transformer based models are computationally expensive to train, and despite their favourable generalization characteristics, MW-MAEs are no different. MW-MAEs and as well as previous works [31, 32] have showed the efficacy of MAEs when pretrained with AudioSet, however, training on longer duration audio data is still a challenge.