							AS-20K (mAP↑)			AS-2M (mAP↑)			VGGSound (Acc.↑)		
Method	Type	Data^{PT}	Par.^{PT}	Par.^{FT}	$A_{in}/V_{in}{}^{PT}$	$A_{in}/V_{in}{}^{FT}$	A	V	A+V	A	V	A+V	A	V	A+V
Audio-only Models															
Aud-SlowFast	Res	-	-	27M	-/-	400×128/-	-	-	-	-	-	-	50.1	-	-
VGGSound	Res	-	-	26M	-/-	500×257/-	-	-	-	-	-	-	48.8	-	-
PANNs	Res	-	-	81M	-/-	1000×64/-	27.8	-	-	43.9	-	-	-	-	-
AST	ViT	IN-SL	87M	87M	-/224×224	1024×128/-	34.7	-	-	45.9	-	-	-	-	-
HTS-AT	ViT	IN-SL	31M	31M	-/224×224	1024×64/-	-	-	-	47.1	-	-	-	-	-
PaSST	ViT	IN-SL	87M	87M	-/224×224	1024×64/-	-	-	-	47.1	-	-	-	-	-
Data2vec	ViT	AS-SSL	87M	87M	1×160K/-	1×160K/-	34.5	-	-	-	-	-	-	-	-
SS-AST	ViT	AS-SSL	87M	87M	1024×128/-	1024×128/-	31.0	-	-	-	-	-	-	-	-
MAE-AST	ViT	AS-SSL	87M	87M	1024×128/-	1024×128/-	30.6	-	-	-	-	-	-	-	-
Aud-MAE	ViT	AS-SSL	87M	87M	1024×128/-	1024×128/-	37.0	-	-	47.3	-	-	-	-	-
Audio-Video Models															
G-Blend	Res	-	-	n/a	-/-	$100\times40/16\times224^2$	29.1	22.1	37.8	32.4	18.8	41.8	-	-	-
Perceiver	ViT	-	-	n/a	-/-	$480\times128/32\times224^2$	-	-	-	38.4	25.8	44.2	-	-	-
Attn AV	Res	IN-SL	60M	60M/60M	-/224^2	$1024\times64/64\times224^2$	-	-	-	38.4	25.7	44.2	-	-	-
CAV-MAE	ViT	IN-SSL, AS-SSL	191M	87M/87M	$1024\times128/1\times224^2$	$1024\times128/10\times224^2$	37.7	19.8	42.0	46.6	26.2	51.2	59.5	47.0	65.5
MBT*	ViT	IN21K-SL	88M	88M/88M	-/224^2	$1024\times128/8\times224^2$	31.3	27.7	43.9	41.5	31.3	49.6	52.3	**51.2**	64.1
MAViL	ViT	AS-SSL	192M	87M/87M	$1024\times128/8\times224^2$	$1024\times128/8\times224^2$	41.6	23.7	44.6	**48.7**	28.3	51.9	60.6	50.0	66.5
MAViL	ViT	AS-SSL,IN-SSL	192M	87M/87M	$1024\times128/8\times224^2$	$1024\times128/8\times224^2$	**41.8**	**24.8**	**44.9**	**48.7**	**30.3**	**53.3**	**60.8**	50.9	**67.1**

Table 1: **Comparison to prior work on AudioSet (AS-20K, AS-2M) and VGGSound** in the audio (A), video (V) and audio+video (A+V) classification tasks. PT: pre-training dataset and type; IN: ImageNet; SL: supervised learning; SSL: self-supervised learning; *:We de-emphasize the model using non-standard dataset splits. We bold the best-performing single model. Res: Res/ConvNet mixture. ViT: (Audio/vision) transformer. Par: number of model parameters for A/V modules. $A_{in}/V_{in}{}^{PT,FT}$: pretraining and finetuning input shape. We drop channel size for clarity.

	Modalities	Noun	Verb	Action
A-SlowFast	A	22.8	46.5	15.4
MBT	A	22.4	44.3	13.0
MAViL	A	26.8	53.1	20.2
SlowFast	V	50.0	65.6	38.5
MTV	V	60.5	67.8	46.7
MBT	V	56.4	62.0	40.7
MAViL	V	54.8	69.4	45.1
MBT	A+V	58.0	64.8	43.4
MAViL	A+V	56.6	71.8	46.0

Table 2: **Comparison to prior work on Epic-Kitchen**. Modalities are A: Audio, V: Video. Metric is top-1 accuracy.

	PT-Data	ESC-FT	ESC-Linear	AS-20K-FT	AS-20K-Linear
XDC	A+V,AS	-	84.8	-	-
AVID	A+V,AS	-	89.1	-	-
BraVe	A+V,AS	-	90.4	-	-
CrissCross	A+V,AS	-	90.5	-	-
SS-AST	A,AS	88.8	85.6	31.0	-
A-MAE	A,AS	94.1	89.5	37.0	-
MAViL	A+V,AS	94.4	90.8	41.8	30.0

Table 3: **Linear-probing and fine-tuning on ESC and AS-20K**. PT: Pre-training. FT: end-to-end fine-tuning. Linear: Linear-probing under standard protocol. AS: AudioSet. Metric: Top-1 accuracy for ESC-50 and mAP for AS-20K.

	PT-Data	#fms	U101-Linear	U101-FT	H51-Linear	H51-FT
XDC	A+V,IG65M	8	-	84.9	-	48.8
AVID	A+V,AS	8	-	88.6	-	57.6
CrissCross	A+V,AS	8	87.7	89.4	56.2	58.3
MAViL	A+V,AS	8	89.1	90.5	58.3	60.7
BraVe*	A+V,AS	32*	93.0	95.6	69.4	76.5

Table 4: **Video tasks (UCF101 & HMDB51)**. Metric: top-1 accuracy. *: Using longer input video frames.