

Table 1: **Audio-visual classification** on VGGSound-Music, VGGSound-All, and AudioSet datasets.

Method	VGGSound-Music		VGGSound-All		AudioSet	
	Linear (%)	Finetune (%)	Linear (%)	Finetune (%)	Linear (%)	Finetune (%)
CAV-MAE	60.53	67.26	55.27	65.53	40.56	51.29
CAV-MAE + Agentic Alignment	65.69	72.38	62.25	70.75	49.53	57.81

Table 2: **Sound source localization and segmentation.** Quantitative results on Flickr-SoundNet and AVSBench.

Method	Flickr-SoundNet			AVSBench	
	Precision	AP	F1	mIoU	F1
SLAVC	55.65	58.12	62.39	31.36	45.02
SLAVC+Agentic Alignment	59.37	61.36	67.21	38.28	50.76

Table 3: **Sound source separation.** Quantitative results on MUSIC and VGGSound datasets.

Method	MUSIC		VGGSound-Instruments		VGGSound-Music	
	SDR	SAR	SDR	SAR	SDR	SAR
OneAVM	7.38	7.48	5.36	5.52	2.51	2.61
OneAVM+Agentic Alignment	10.65	12.35	7.23	7.38	5.93	5.96

Table 4: **Datasets settings.** Details settings on the dataset, number of train/val/test splits, and downstream tasks.

Dataset	# Train/Val/Test splits	Downstream Tasks
VGGSound-Music	39900 / 1008 / 1201	source separation, audio-visual classification
VGGSound-Instruments	32000 / 503 / 500	source separation
MUSIC	300 / 58 / 90	source separation
Flickr-SoundNet	4000 / 500 / 500	source localization
AVSBench-Semantic	8498 / 1304 / 1554	source segmentation
VGGSound-All	168471 / 15500 / 15496	audio-visual classification
AudioSet	1667456 / 208432 / 208432	audio-visual classification

Table 5: **Comparison to all actions.** Quantitative results on VGGSound-All, Flickr-SoundNet, AVSBench, and VGGSound-Music datasets.

Method	VGGSound-All		Flickr-SoundNet			AVSBench		VGGSound-Music	
	Linear (%)	Finetune (%)	Precision	AP	F1	mIoU	F1	SDR	SAR
All Actions	46.78	52.13	40.75	42.35	43.23	15.76	26.38	0.36	0.37
AVAGENT (ours)	61.56	69.24	58.23	60.76	65.03	36.37	49.85	5.26	5.28

Table 6: **Audio-visual video parsing** on Look, Listen, and Parse (LLP) dataset.

Method	Segment-level					Event-level				
	A	V	AV	Type	Event	A	V	AV	Type	Event
VALOR	61.8	65.9	58.4	62.0	61.5	55.4	62.6	52.2	56.7	54.2
VALOR + Agentic Alignment	63.2	66.2	59.5	62.9	62.1	56.8	63.4	54.5	58.2	55.3