AVROBUSTBENCH

Benchmarking the Robustness of Audio-Visual Recognition Models at Test-Time


Sarthak Kumar Maharana


Saksham Singh Kushwaha


Baoming Zhang


Adrian Rodriguez


Songtao Wei


Yapeng Tian

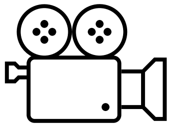

Yunhui Guo

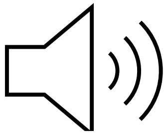
arXiv

Data shifts are unavoidable in the real-world

Data shifts are unavoidable in the real-world

Noisy audio & visual environments

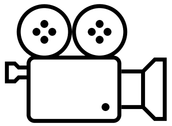


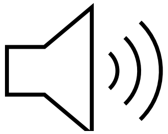


Limited recognition ?

Data shifts are unavoidable in the real-world

Noisy audio & visual environments

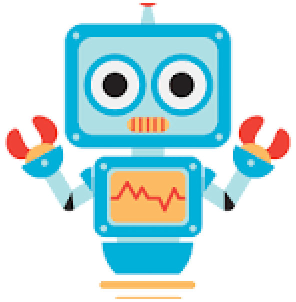




Limited recognition

Another critical example....

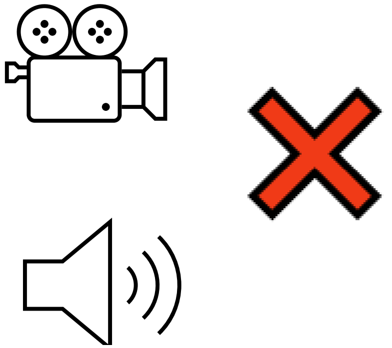
Robots deployed in crowded streets



Unavoidable bimodal shifts can hamper scene understanding

Data shifts are unavoidable in the real-world

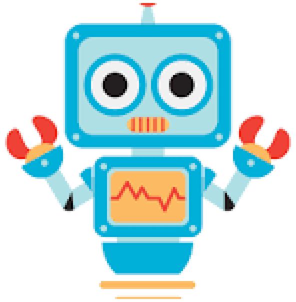
Noisy audio & visual environments



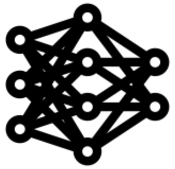
Limited recognition

Another critical example....

Robots deployed in crowded streets



Unavoidable bimodal shifts can hamper scene understanding

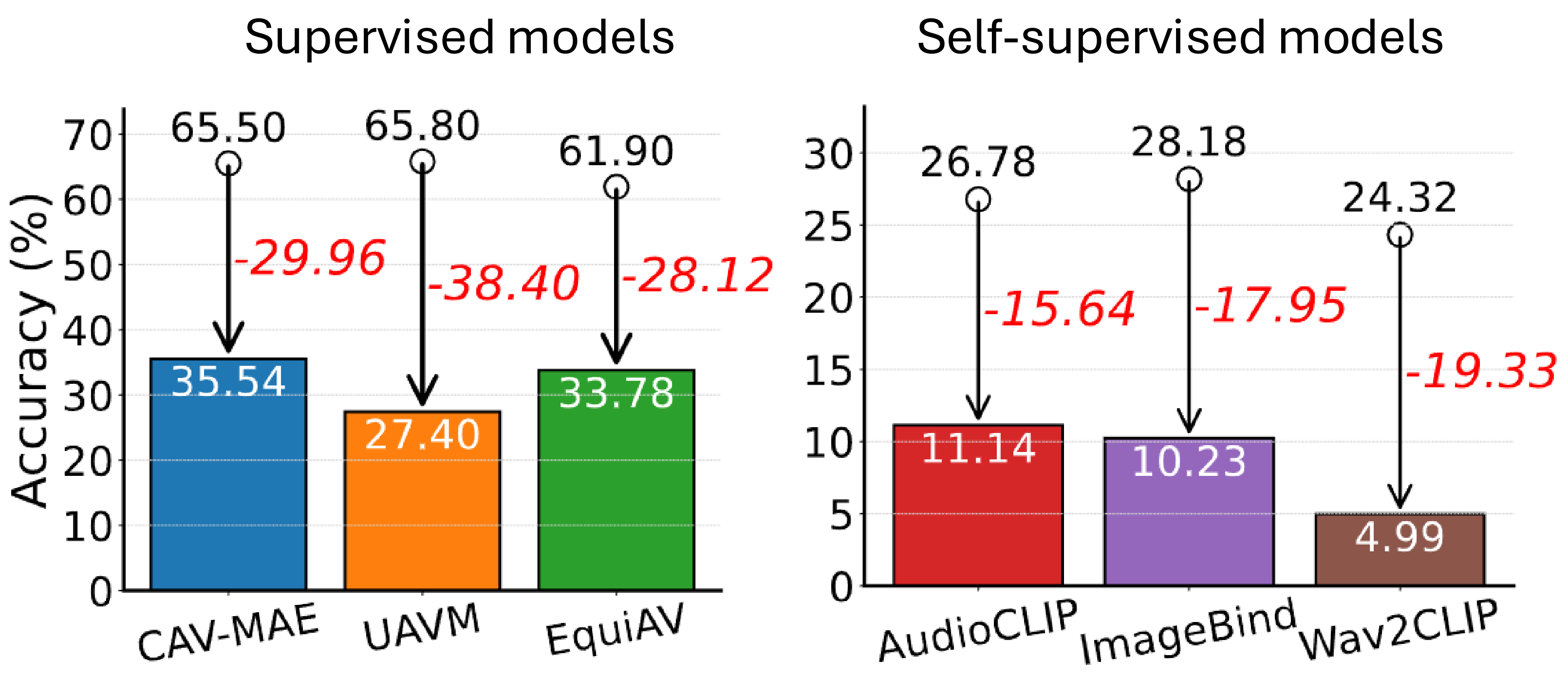


Testing data varies in many ways in the real-world.
Yet, our models are the same and fail to generalize!

Limited recognition of popular audio-visual recognition models

Popular supervised and self-supervised models struggle with bimodal corruptions at test-time!

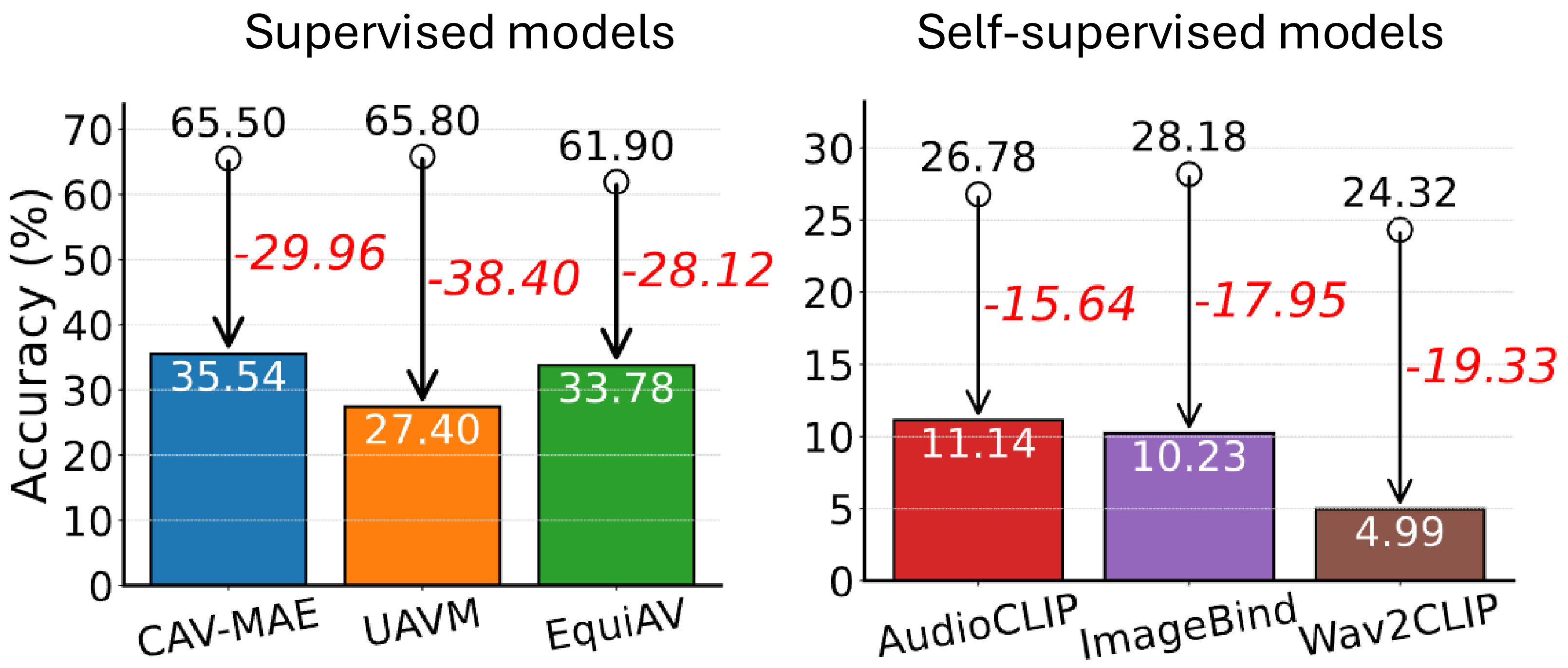
On our proposed **VGGSOUND-2C**



Limited recognition of popular audio-visual recognition models

Popular supervised and self-supervised models struggle with bimodal corruptions at test-time!

On our proposed **VGGSOUND-2C**



There is a need to establish rigorous robustness benchmarks for audio-visual learning problems!

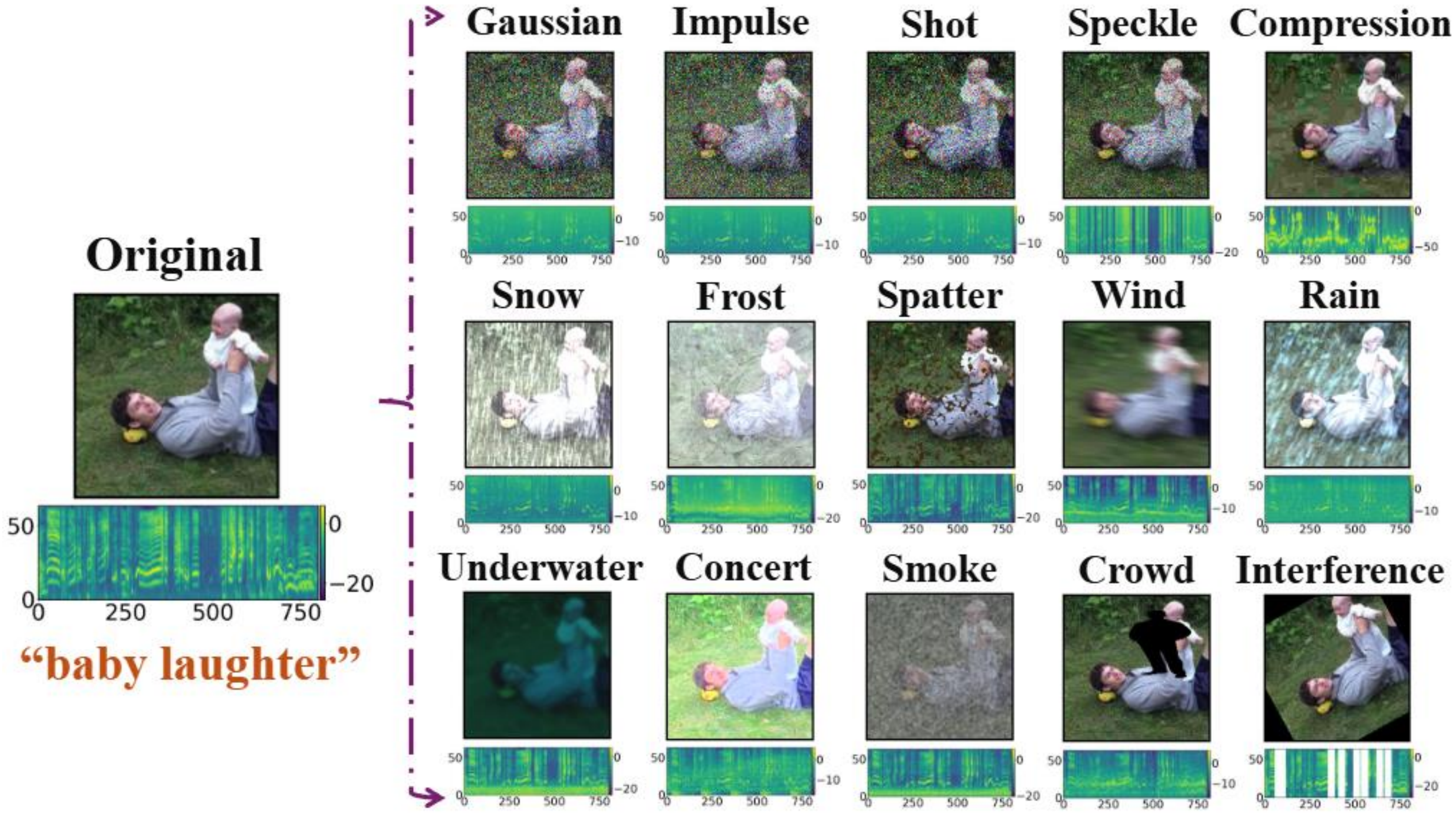
So, what's in **AVROBUSTBENCH**❓

We release ***realistic, co-occurring, and correlated corruptions*** to audio-visual modalities.

Benchmark	Modalities	Real-World Shifts?	Multimodal Corruptions?	Features	
				Co-occur?	Correlated?
ImageNet-C	{v}	✓	✗	✗	✗
MULTIBENCH	{a, v}	–	✗	✗	✗
YouCook2-P, MSRVTT-P	{l, v}	✓	✗	✗	✗
SRB	{s}	✓	✗	✗	✗
Chen et al.	{l, v}	✓	✗	✗	✗
Hong et al.	{s, v}	✓	✓	✓	✗
READ	{a, v}	✓	✗	✗	✗
AVROBUSTBENCH (Ours)	{a, v}	✓	✓	✓	✓

So, what's in **AVROBUSTBENCH**❓

We release ***realistic, co-occurring, and correlated corruptions*** to audio-visual modalities.



To emulate *real-world* shifts,

1. We introduce **75 corruptions** - 15 corruptions at 5 severity levels, each.
2. Both modalities are **jointly corrupted** at test-time.
3. Three major categories – *Digital*, *Environmental*, and *Human-Related*.
4. *Can be extended to any audio/speech-visual dataset to study robustness.*

So, what's in AVROBUSTBENCH❓

We also release four datasets – **AUDIOSET-2C, VGGSOUND-2C, KINETICS-2C,** **and** **EPICKITCHENS-2C**. Our proposed corruptions are *applied to the test sets* of the source datasets.

Dataset	# Samples	Classes	Avg. duration
AudioSet-2C	16,742	527	10 sec
VGGSound-2C	14,046	309	10 sec
Kinetics-2C	3,111	32	10 sec
EpicKitchens-2C	205	97 (Noun) 300 (Verb)	7.4 mins

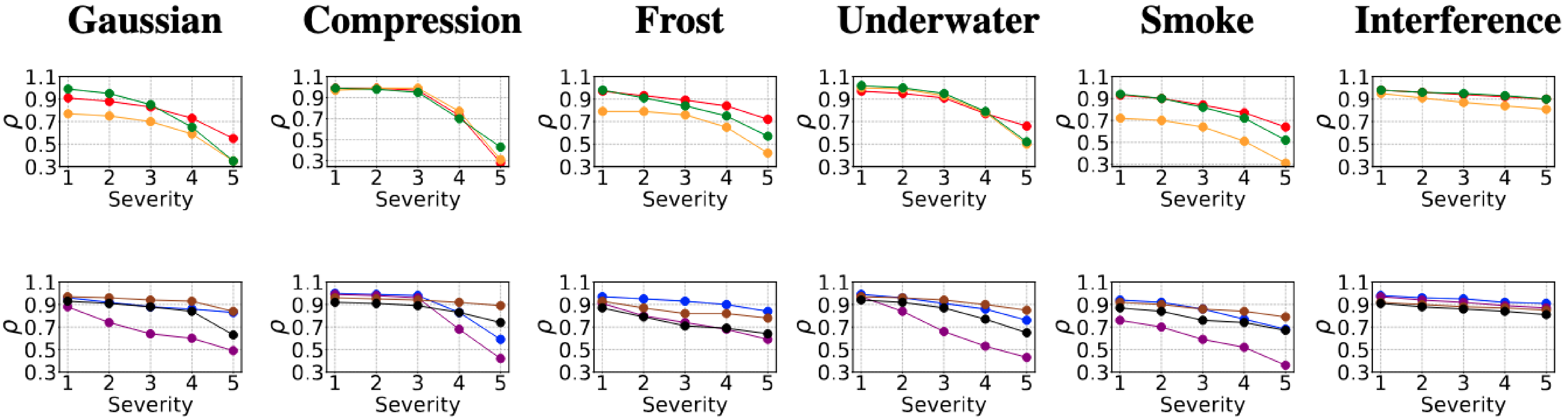
Check out our demo @ **https://www.youtube.com/watch?v=hYdcRO3BuIY&ab_channel=SarthakMaharana**

Results on robustness at test-time

Relative robustness $\rho = 1 - \frac{new\ test\ acc. - source\ test\ acc.}{source\ test\ acc.}$

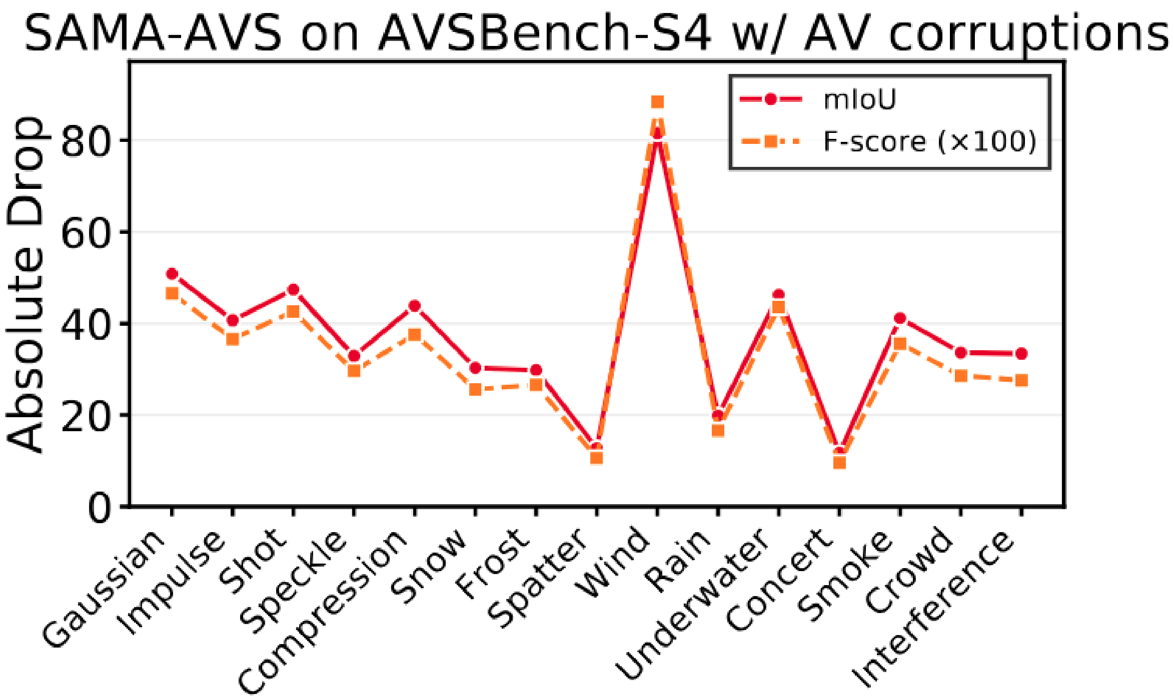
Corruption has a large effect on model robustness; increase in severity decreases robustness!

Top – **AUDIOSET-2C**, Bottom – **EPICKITCHENS-2C** for difference supervised and self-supervised models

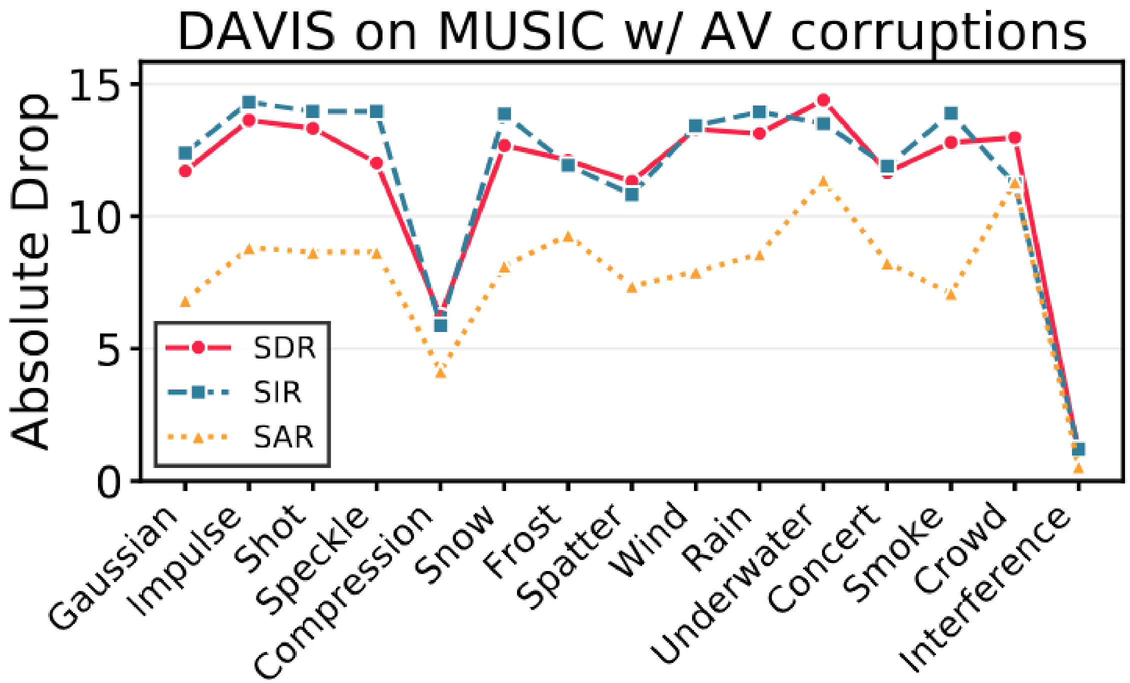


Beyond recognition tasks 🤔

Audio-Visual Segmentation



Sound Source Separation



AVROBUSTBENCH

Benchmarking the Robustness of Audio-Visual Recognition Models at Test-Time







Sarthak Kumar Maharana
Saksham Singh Kushwaha
Baoming Zhang
Adrian Rodriguez
Songtao Wei
Yapeng Tian
Yunhui Guo