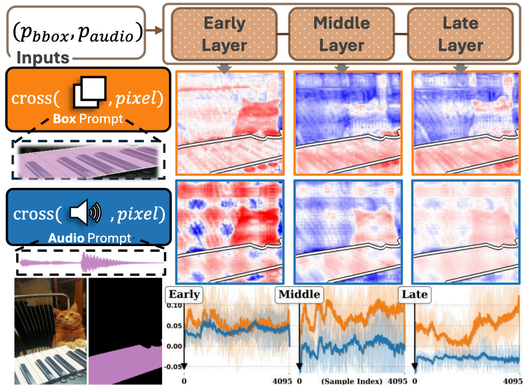
AuralSAM2: Enabling SAM2 Hear Through Pyramid Audio-Visual Feature Prompting

Yuyuan Liu[1], Yuanhong Chen[2], Chong Wang[3], Junlin Han[1], Junde Wu[1], Can Peng[1], Jingkun Chen[1], Yu Tian[4](✉), Gustavo Carneiro[5]

[1] University of Oxford [2] Adelaide University [3] Stanford University [5] University of Central Florida [5] University of Surrey

CVPR JUNE 3-7, 2026

Motivation



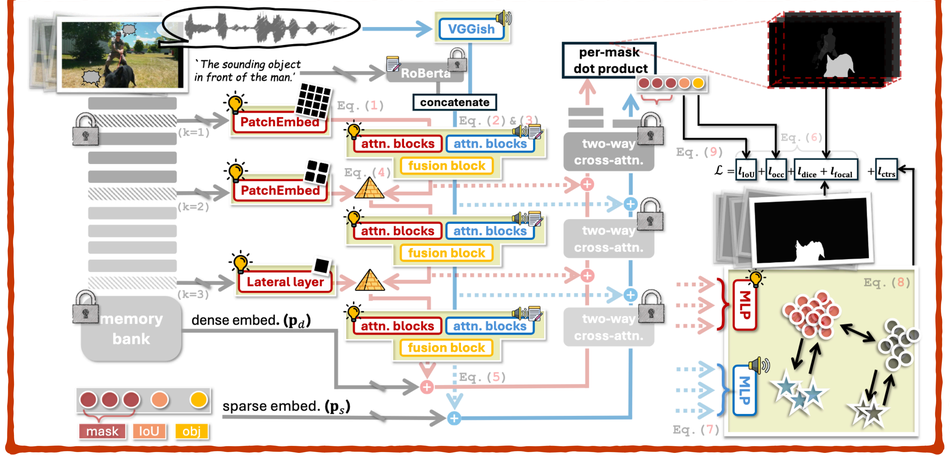
- Audio prompt in SAM2 suffers from dilution
  - I. Cross-modal signals weaken across network propagation.
  - II. Visual features dominate sparse audio embeddings.
- SAM-based AVS methods remain inefficient
  - I. Adapter-based fusion requires repeated inference.
  - II. Foundation models produce inaccurate audio prompts.

Contribution

- Propose **AuralFuser** for efficient audio-conditioned prompting in SAM2.
- Mitigate audio prompt dilution via hierarchical feature pyramid fusion.
- Introduce AudioCon for audio-visual alignment.

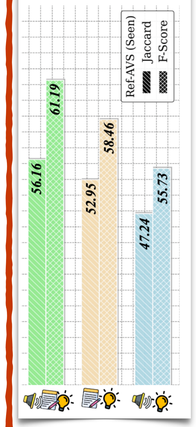
* Open to research opportunities in MLLM grounding, reasoning and RL post-training. Please check the QR code for more details.

Methodology



Ablation Studies

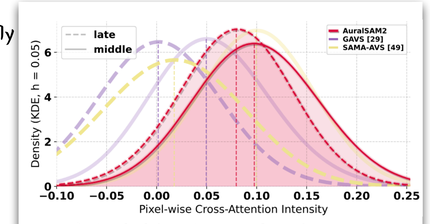
bar). Each Modalities matter in Ref-AVS.



Tab). Ablation Studies on AVSBench and Ref-AVS. The first row uses only visual inputs, while the remaining rows incorporate audio and optional language modalities.

Ablations	Pyramid	AVSBench [58, 59] V1 (single) $\mathcal{M}_\mathcal{J}\uparrow$	$\mathcal{M}_\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	V1 (multiple) $\mathcal{M}_\mathcal{J}\uparrow$	$\mathcal{M}_\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	Ref-AVS [51] Seen $\mathcal{M}_\mathcal{J}\uparrow$	$\mathcal{M}_\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	Unseen $\mathcal{M}_\mathcal{J}\uparrow$	$\mathcal{M}_\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$
	-	83.41	91.36	87.39	61.50	73.09	67.30	43.77	48.01	45.89	64.33	69.98	67.16
	-	84.55	92.08	88.32	71.52	79.57	75.55	53.36	57.49	55.43	66.94	72.18	69.56
	=2	85.96	92.97	89.47	73.42	81.94	77.68	54.67	58.91	56.79	67.81	72.86	70.34
	=3	86.33	93.27	89.80	74.43	82.76	78.60	55.32	60.69	58.00	67.74	73.92	70.83
AudioCon	=3	86.62	93.34	89.98	75.58	84.12	79.85	56.16	61.19	58.68	68.69	74.36	71.53

- Different modalities are effectively aligned and jointly contribute.
- Hierarchical feature prompting alleviates audio prompt dilution.
- Stronger cross-modal attention leads to better recognition.



Curve). Probability density of pixel-wise cross-attention intensity for audio prompt dilution.

Experiments

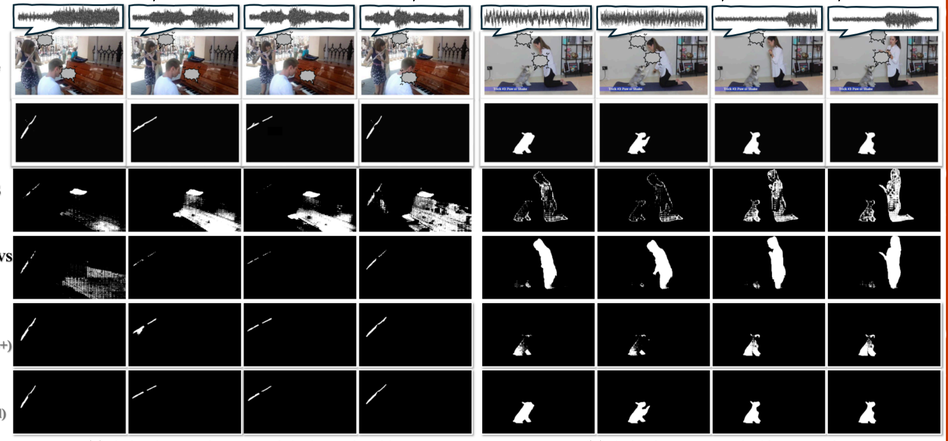
See paper for AVS results

Tab). Comparison with SOTA methods on the Ref-AVS dataset. Methods built upon SAM are highlighted in mauve, those based on SAM2 in yellow, while the remaining methods rely on task-specific architectures. Best results are marked in red.

Method	Backbone	Ref-AVS [51] Seen $\mathcal{M}_\mathcal{J}\uparrow$	$\mathcal{M}_\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	Unseen $\mathcal{M}_\mathcal{J}\uparrow$	$\mathcal{M}_\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	Mix $\mathcal{M}_\mathcal{J}\uparrow$	$\mathcal{M}_\mathcal{F}\uparrow$	$\mathcal{J}\&\mathcal{F}\uparrow$	Null $\mathcal{S}\downarrow$
TPAVI [58] [ECCV 2022]	PVT-v2	23.20	51.1	37.2	32.36	54.7	43.5	27.78	52.9	40.3	0.208
AVSegFormer [13] [AAAI 2024]	PVT-v2	33.47	47.0	40.2	36.05	50.1	43.1	34.76	48.6	41.7	0.171
EEMC [51] [ECCV 2024]	Swin-b	34.20	51.3	42.8	49.54	64.8	57.2	41.87	58.1	50.0	0.007
GAVS [50] [AAAI 2024]	ViT-h	28.9	49.8	39.35	29.8	49.7	39.8	29.4	49.8	39.6	0.190
SAMA-AVS [50] [WACV 2024]	ViT-h	39.2	56.2	47.7	47.5	56.6	52.1	43.4	56.4	49.9	0.130
TSAM [41] [CVPR 2025]	ViT-h	43.4	56.8	50.1	54.6	66.4	60.5	49.0	61.6	55.3	0.017
Ours SAM (w/ AuralFuser)	ViT-h	48.26	60.28	54.27	57.91	68.95	63.43	53.09	59.10	58.85	0.053
GroundedSAM2* [43] [arxiv 2024]	Hiera-b+	28.5	39.9	34.2	59.8	68.1	63.9	44.2	54.0	49.1	0.277
GAVS† [50] [AAAI 2024]	Hiera-b+	48.0	54.6	51.3	59.2	65.8	62.5	53.6	60.2	56.9	0.076
SAMA-AVS† [50] [WACV 2024]	Hiera-b+	49.5	56.7	53.1	60.6	66.4	63.5	55.1	61.5	58.3	0.103
SAM2-LOVE [52] [CVPR 2025]	Hiera-l	43.5	51.9	47.7	66.5	72.3	69.4	55.0	62.1	58.5	0.230
Ours AuralSAM2	Hiera-b+	53.16	58.83	56.00	63.45	70.44	66.95	58.31	64.64	61.48	0.129
	Hiera-l	56.16	61.19	58.68	68.69	74.36	71.53	62.43	67.78	65.11	0.065

Visualisation

Fig). Qualitative visualisations on the Ref-AVS dataset. Each column represents frames sampled from video clips, with the audio waveform overlaid in the speech bubble. Ground-truth and predictions from different methods are presented in separate rows.



(a). *'The object making a sound by being played by the woman'*

(b). *'The dog sitting on the blue carpet.'*