SAM Audio: Segment Anything in Audio

A foundation model for general audio separation

Bowen Shi*, Andros Tjandra*, John Hoffman*, Helin Wang*, Yi-Chiao Wu*, Luya Gao*,
Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer,
Piotr Dollár, Wei-Ning Hsu, Ann Lee

Meta

Conversational AI Reading Group
March 2026

Outline

Background & Motivation

SAM Audio Model

Training Data

Evaluation Data & Task

SAM Audio Judge (SAJ)

Results

Conclusion

The Audio Separation Landscape

Audio source separation aims to decompose **a complex sound mixture** into **individual source** tracks corresponding to **distinct sound events**.

Speech Separation

Focuses on isolating human speech from background noise or separating multiple overlapping speakers.

Prior works

Conv-TasNet, SepFormer

Music Source Separation

Decomposes musical mixtures into a fixed ontology of stems (e.g., vocals, drums, bass, other).

Prior works

Demucs, Spleeter

Sound Event Separation

Extracts general audio events based on text descriptions, but often struggles with complex music or speech.

Prior works

AudioSep, CLAPSep

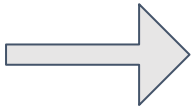
Limitations of Existing Models

- **Domain-Specific Architecture**
  Models are specialized for either speech, music, or specific sound effects, failing to generalize across domains.
- **Fixed Closed-Set Categories**
  Many systems rely on predefined taxonomies (e.g., vocals, drums, bass), struggling with open-domain mixtures.
- **Single Prompt Modality**
  Limited to a single prompting method, typically just text descriptions, lacking flexibility for complex scenarios.
- **No Unified Evaluation**
  Absence of a comprehensive benchmark to evaluate open-domain audio separation in real-world conditions.

The field lacks a **unified foundation model** capable of handling arbitrary sounds across all domains with flexible, multi-modal prompting.

Inspiration: SAM for Vision

Segment Anything Model (Vision)



SAM Audio (Our Goal)

- **Unified Task:** Replaced specialized segmentation models with a single foundation model.
- **Promptable Interface:** Allows users to specify targets using points, bounding boxes, or text.
- **Zero-Shot Generalization:** Capable of segmenting unfamiliar objects without additional training.

- **Unified Task:** Replace specialized speech, music, and sound models with one foundation model.
- **Promptable Interface:** Allow users to specify targets using text descriptions, visual masks, or temporal spans.
- **Open-Domain Generalization:** Capable of separating arbitrary sounds in complex, real-world mixtures.

The SAM Audio Goal

Creating a unified foundation model for all audio types

Open-Domain

A single model that handles speech, music, and general sound events seamlessly without domain-specific constraints.

Arbitrary Sounds

Extracts any target source described by the user, moving beyond fixed taxonomies and predefined categories.

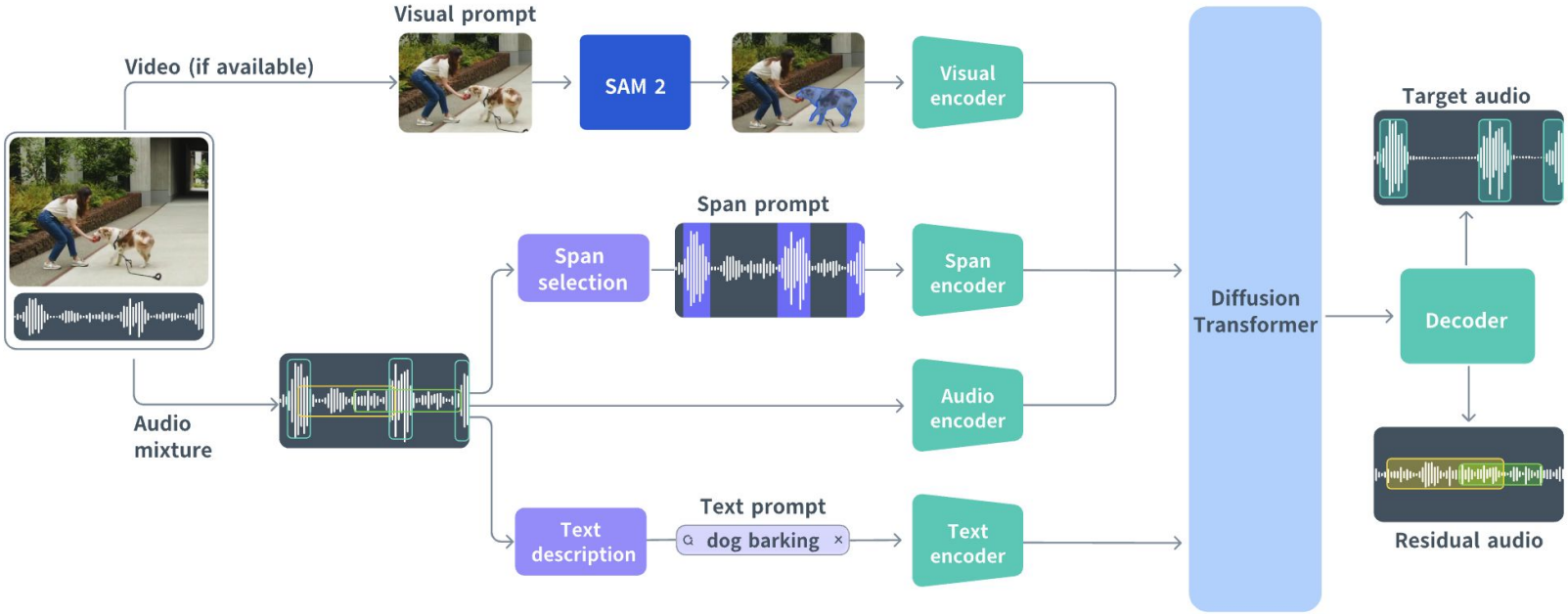
Multi-Modal Prompting

Supports diverse interaction methods including text descriptions, visual masks, and temporal spans.

Generative Separation

Utilizes flow-matching rather than traditional discriminative masking for more natural, high-fidelity outputs.

Model Overview



Three Prompt Modalities

Text Prompt

Describes the target sound in natural language (e.g., "dog barking", "piano playing").

ENCODER

T5 Text Encoder

Visual Prompt

Specifies the target source by providing a visual mask over the video frames.

ENCODER

SAM 2 + Visual Encoder

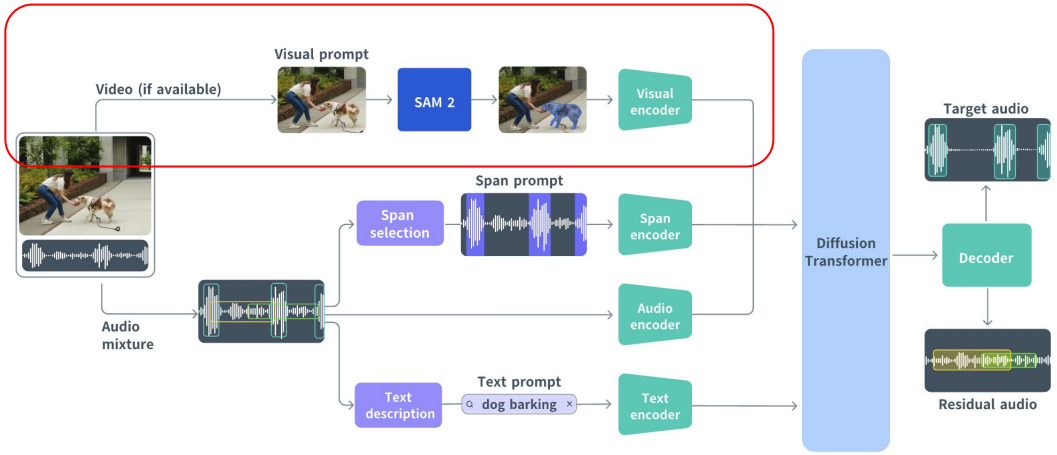
Span Prompt

Specifies the target source by providing a temporal window where the sound is active.

ENCODER

Span Selection + Encoder

Visual Prompt Path



Mask Generation

Users provide bounding boxes on video frames. SAM 2 converts these into precise visual masks for the target object.

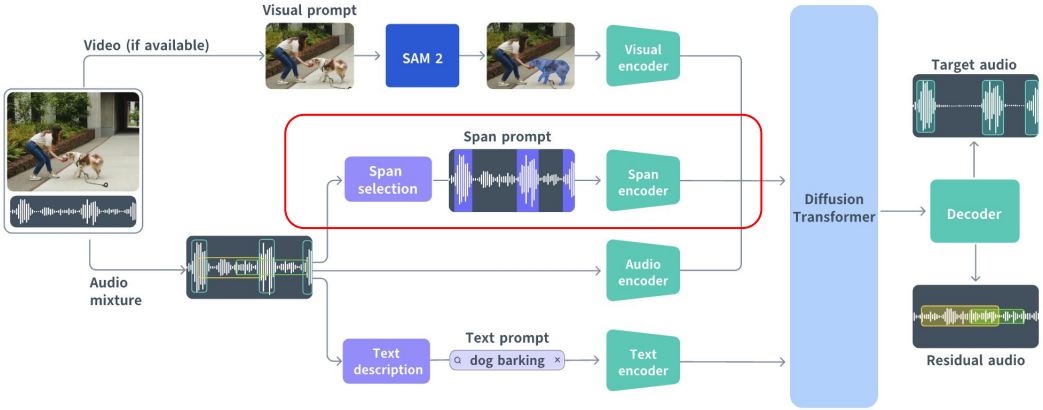
Perception Encoder

The masked video is processed by a Perception Encoder (PE) to extract frame-level visual features.

Feature Injection

These visual features are injected into the Diffusion Transformer to guide the separation process.

Span Prompt Path



Temporal Selection

Users provide start and end times for the target sound, allowing precise frame-level control over when the sound is active.

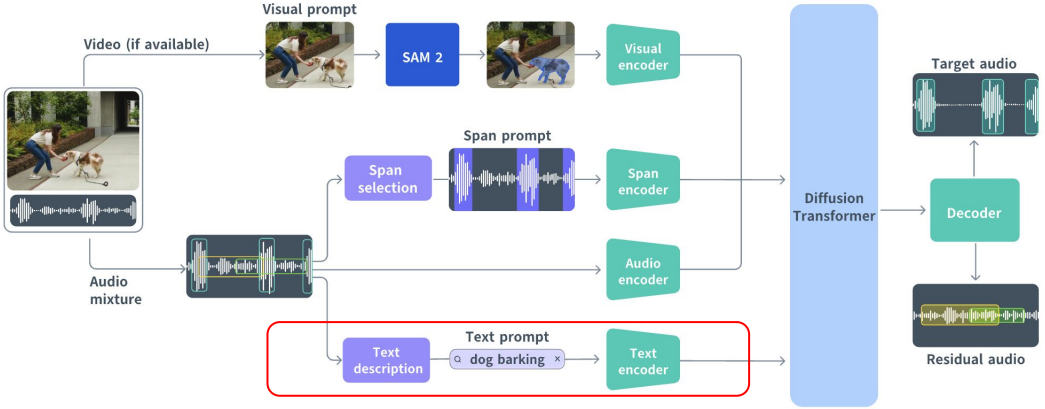
Token Sequence

The selected temporal span is converted into a sequence of tokens representing the active regions.

Span Encoder

A dedicated Span Encoder processes these tokens to guide the Diffusion Transformer during separation.

Text Prompt Path



Natural Language

Users describe the target sound using free-form text (e.g., "dog barking").

T5 Text Encoder

The text is processed by a pre-trained T5 encoder to extract rich semantic embeddings.

Cross-Attention

These embeddings are injected into the Diffusion Transformer via cross-attention layers.

Audio Encoder & Latent Space

DAC-VAE Latent Space

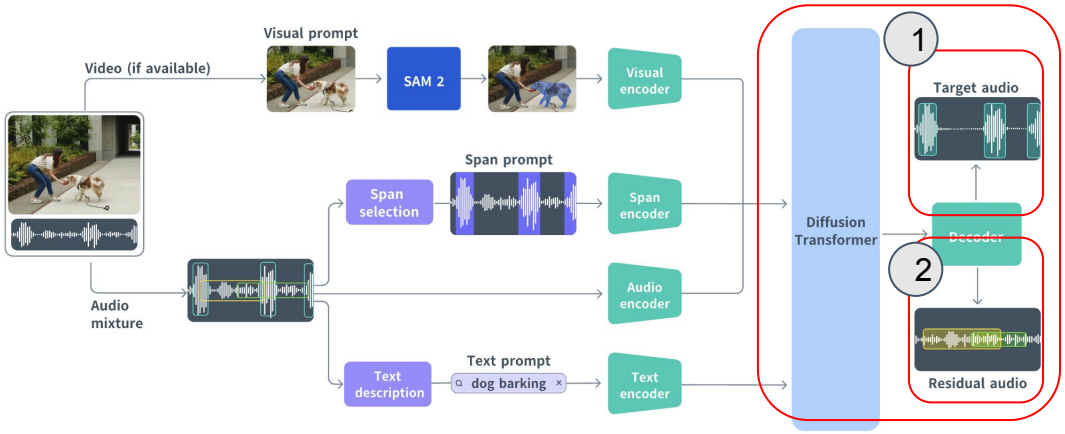
- **Continuous Representation:** Encodes raw audio waveforms into a continuous latent space rather than discrete tokens.
- **Dimensionality Reduction:** Significantly reduces the sequence length, making diffusion modeling computationally tractable.
- **High Fidelity:** Preserves fine-grained acoustic details necessary for high-quality audio reconstruction.

Audio Encoder

- **Mixture Processing:** Takes the latent representation of the complex audio mixture as input.
- **Feature Extraction:** Extracts rich acoustic features that capture the relationships between different sound sources.
- **Conditioning:** Provides the necessary context to the Diffusion Transformer for targeted separation.

https://github.com/facebookresearch/dacvae

Diffusion Transformer & Decoder



Flow-Matching

The DiT learns to map a simple noise distribution to the complex data distribution of the target audio.

Joint Generation

It simultaneously predicts both the target stem and the residual stem in the latent space.

DAC Decoder

The generated latents are passed through the DAC-VAE decoder to reconstruct the final high-fidelity audio waveforms.

Joint Target & Residual Generation

- **Simultaneous Output**
  Unlike models that only output the target sound, SAM Audio generates both the target stem and the residual audio simultaneously.
- **Coherent Separation**
  Joint generation ensures that the target and residual are acoustically coherent and perfectly complement each other.
- **Flow-Matching Formulation**
  The model learns a vector field that maps noise to the joint distribution of (target, residual) pairs.

Mixture = Target + Residual

The fundamental principle of separation

KEY CAPABILITY UNLOCKED

Sound Removal By generating the residual, SAM Audio inherently supports sound removal as a byproduct. Users can specify a sound to remove (e.g., a dog bark), and the model outputs the clean residual audio.

Diffusion Transformer Core

- **Flow-Matching Architecture:** DiT-based framework for stable, scalable generation.
- **Joint Generation:** Trains to produce target and residual audio together.
- **Scalability:** Model family ranges from 500M to 3B parameters.

Input & output

$$x_0 \sim \mathcal{N}(0,1)$$
$$x_1 = (\text{target audio, residual audio})$$

Input conditioning

$$c = \{x_{\text{mix}}, c_{\text{text}}, c_{\text{vid}}, c_{\text{span}}\}$$

Flow matching objective

$$\mathcal{L}_{FM} = \|u(x_t, t, c; \theta) - (x_1 - (1 - \sigma_{min})x_0)\|$$

Auxiliary alignment loss

$$\mathcal{L}_{aux} = \mathbb{E}_t \left[1 - \text{sim}(\hat{a}_t, a_{tgt})\right]$$

$$a_{tgt} = \text{AED\_embed}(x_1)$$
$$\hat{a}_t = \phi(h_t)$$

Final loss

$$\mathcal{L} = \mathcal{L}_{FM} + \lambda * \mathcal{L}_{aux}$$

Model Configurations

Model	Total params	Layers	Attn dim	FFN dim
SAM Audio-Small	500M	12	1,536	6,144
SAM Audio-Base	1B	16	2,048	8,192
SAM Audio-Large	3B	22	2,816	11,264

SAM Audio scales from **500M to 3B parameters**. The models are trained using a two-stage recipe: large-scale pre-training on synthetic mixtures, followed by fine-tuning on curated high-quality data. The **3B model (SAM Audio-Large)** demonstrates the best performance and is used for all key comparisons.

Generative vs. Discriminative

Traditional: Discriminative

- ✕ **Mask-Based:** Applies time-frequency masks to the input spectrogram to filter out unwanted sounds.
- ✕ **Information Loss:** Struggles when the target sound is heavily masked or corrupted by noise.
- ✕ **Artifacts:** Often produces unnatural, "robotic" sounding artifacts due to imperfect masking.

SAM Audio: Generative

- ✓ **Flow-Matching:** Synthesizes the target audio from scratch conditioned on the mixture and prompt.
- ✓ **Reconstruction:** Capable of hallucinating and reconstructing missing frequencies for a complete sound.
- ✓ **High Fidelity:** Produces much more natural, high-quality audio outputs without masking artifacts.

Training Data Overview

Category	Input Audio Mixture x_{mix}	Target Audio x_{tgt}	Residual Audio x_{res}
Fully-real triplets	✓	✓	✓
Synthetic mixtures	✗	✓	✓
Pseudo-labeled stems	✓	✗	✗

Mixing strategy

1) Fully-real triplet

$$(x_{\text{mix}}, x_{\text{tgt}}, x_{\text{res}})$$

Use triplet directly

OR

Re-mixing SNR +/- 5

$$x_{\text{mix}} = \sum_{i=1}^{N} x_i$$

$$x_{\text{res}} = \sum_{j=1, j \neq i}^{N} x_j$$

2) Synthetic audio mixture

not exist

sampled randomly from dataset

$$(\square, x_{\text{tgt}}, x_{\text{res}})$$

Mixing SNR +/- 15

Mixing strategy (3) - Pseudo-Labeling Pipeline

Data Filtering Criteria

- **Quality Assurance:** To ensure high-quality training data, all pseudo-labeled samples undergo rigorous filtering.
- **Text-Audio Alignment:** A CLAP score threshold (≥ 0.35) ensures the generated text prompt accurately matches the separated audio.
- **Audio Quality:** An Audio Evaluation Score PC < 2.5 guarantees the overall perceptual quality of the extracted sound.
- **Domain-Specific Checks:** Voice Activity Detection (VAD) is applied specifically for speech samples to ensure speech presence.

Criterio	Threshold
Text–Audio Filtering (all must pass)	
CLAP(text, target audio)	> 0.35
CLAP(text, residual audio)	< 0.0
Aesthetic PC score (target audio)	< 2.5
Silence ratio (target audio)	< 95%
Additional Visual-Audio Filtering	
Mask coverage ratio (masked region)	> 0.02
ImageBind(target audio, masked region)	> 0.2

Visual Pseudo-Labeling

Object Tracking

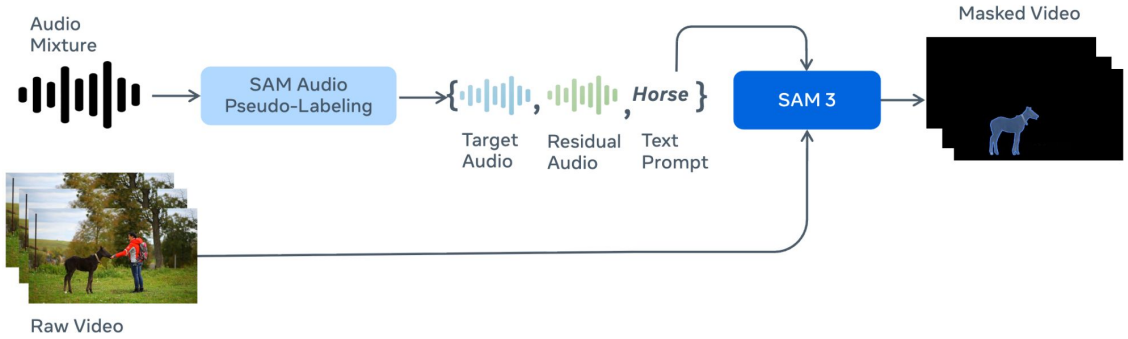
To generate visual-prompted training data, the pipeline uses SAM 2 to track objects in video.

Mask Generation

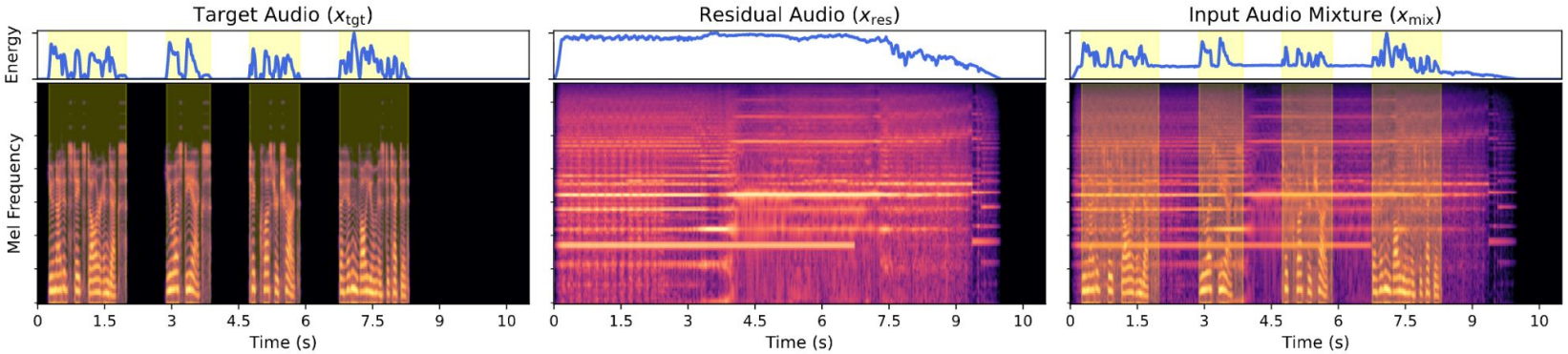
It extracts bounding boxes and generates precise visual masks for the target objects.

Audio-Visual Pairing

The audio is separated using the corresponding text prompt, creating a paired (video mask, audio stem) training sample.



Span Data Construction



Event Selection

Focuses on spiky, discrete sound events (e.g., dog bark, door slam) rather than continuous ambience, as they provide strong temporal cues.

VAD Processing

Applies Voice Activity Detection (VAD) to clean target audio with a -40 dBFS threshold and 250ms minimum duration to find active regions.

Interval Generation

Converts consecutive sounding segments into precise time intervals (yellow regions), which serve as the ground-truth span prompts.

Prompt Generation Strategy

Text Prompts

- **PLM-Audio+LLM:** Used to automatically caption audio mixtures for pseudo-labeling.
- **Templates:** Derived from instrument labels (e.g., "piano playing") for music datasets.
- **Classifiers:** Pretrained gender classifiers used to generate speech prompts (e.g., "female speaking").

Visual Prompts

- **SAM 3 Integration:** Masks are generated by prompting SAM 3 with the target sound's text caption.
- **ImageBind Filtering:** Ensures strong audio-visual correspondence by filtering out poorly matched pairs.
- **Whole-Video:** Also uses full video clips without explicit masks for weakly supervised learning.

Span Prompts

- **Temporal Windows:** Randomly sampled timeframes where the target sound is known to be active.
- **Grounding:** Teaches the model to associate specific acoustic events with their temporal occurrence.
- **Flexibility:** Allows users to simply highlight a section of audio to extract the dominant sound within it.

SAM Audio-Bench

Real-world test item

Based on real use case

Complex, in-the-wild audio mixtures

3

Prompt Modalities

Text, Visual, and Span prompts

100%

Human Annotated

Ensures natural, real-world prompt distributions

6 Comprehensive Tasks

Speech Extraction

Separate all speech from noise

Speaker Extraction

Isolate specific individuals

Music Extraction

Separate music from speech/SFX

Instrument (Wild)

Single stem from noisy mixtures

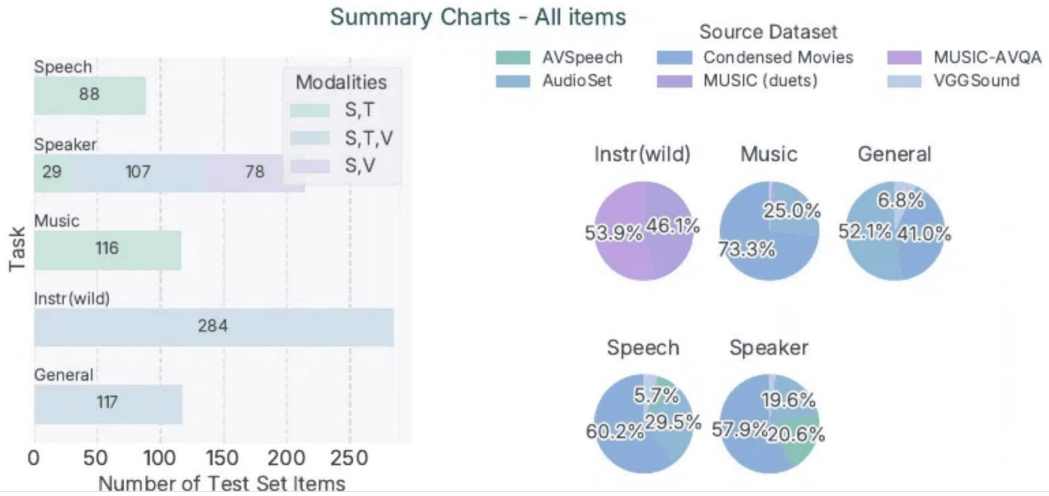
Instrument (Pro)

High-quality studio recordings

General Sound Events

Extract arbitrary target sounds

Benchmark Coverage



Human Annotation Process

Text Annotation

- **Natural Language:** Annotators write free-form descriptions of the target sound.
- **Contextual:** Descriptions are based on listening to the complex mixture, ensuring real-world relevance.
- **Diversity:** Captures the natural variance in how humans describe sounds.

Visual Annotation

- **Bounding Boxes:** Annotators draw boxes around the sound source in the video frame.
- **Mask Generation:** SAM 2 is used to convert these boxes into precise visual masks.
- **Verification:** Ensures the visual prompt accurately points to the sounding object.

Span Annotation

- **Temporal Marking:** Annotators mark the exact start and end times of the target event.
- **Precision:** Provides highly accurate temporal grounding for the model to learn from.
- **Challenging Cases:** Especially useful for overlapping sounds where text might be ambiguous.

Real-World Challenges

Difficulties present in the benchmark items

Overlapping Sounds

Multiple sound sources active simultaneously, requiring the model to disentangle complex acoustic mixtures.

Background Noise

High levels of ambient noise or reverberation that can mask the target sound and degrade separation quality.

Similar Timbre

Separating sources that sound alike, such as two female speakers or two similar instruments playing together.

Low Signal-to-Noise Ratio

The target sound is very quiet compared to the rest of the mixture, making it difficult to detect and extract.

SAM Audio Judge (SAJ)

In addition, we found that introducing a proxy task that predicts whether the output audio follows the text prompt (Wang et al., 2022b,a), significantly improves model performance. To this end, we pretrain the entire model on this text–audio alignment detection task using a large-scale simulated dataset that provides access to separated tracks within mixture audio. We alternate the output audio between the target sound and a random non-target sound from the same mixture to represent the presence or absence of the target sound. An additional linear layer is used to predict the presence or absence of the target sound described by the text prompt. After this pre-training stage, we finetune the SAJ model to predict the final SAJ scores.

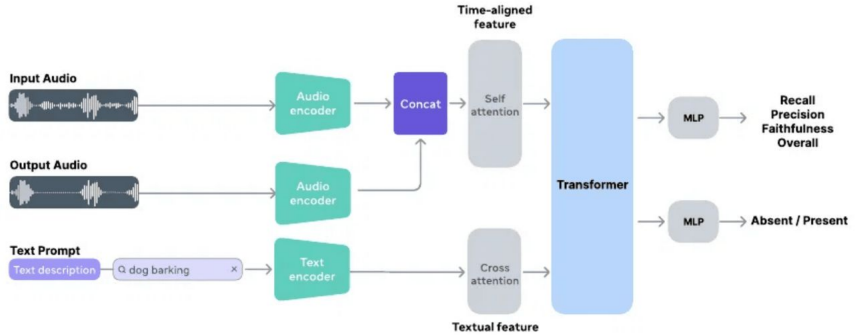


Figure 6 Diagram of SAM Audio Judge Model

See this paper for more information
Paper: SAM Audio Judge: A Unified Multimodal Framework for Perceptual Evaluation of Audio Separation
ArXiv: https://arxiv.org/abs/2601.19702

SAJ Performance Dimensions

Evaluating separation quality across 4 key metrics

Recall

Does the extracted audio contain all of the target sounds specified in the prompt?

Precision

How effectively does the model remove non-target sounds from the extracted audio?

Faithfulness

For target sounds present in the extracted audio, how similar do they sound to their counterparts in the original mixture?

Overall quality

What is the overall perceptual quality of the model's output?

SAJ Difficulty Dimensions

Counting
How many non-target sounds are present in the source audio?

Overlapping
To what extent do the target sounds overlap with non-target sounds?

Loudness
How loud are the target sounds relative to the non-target sounds?

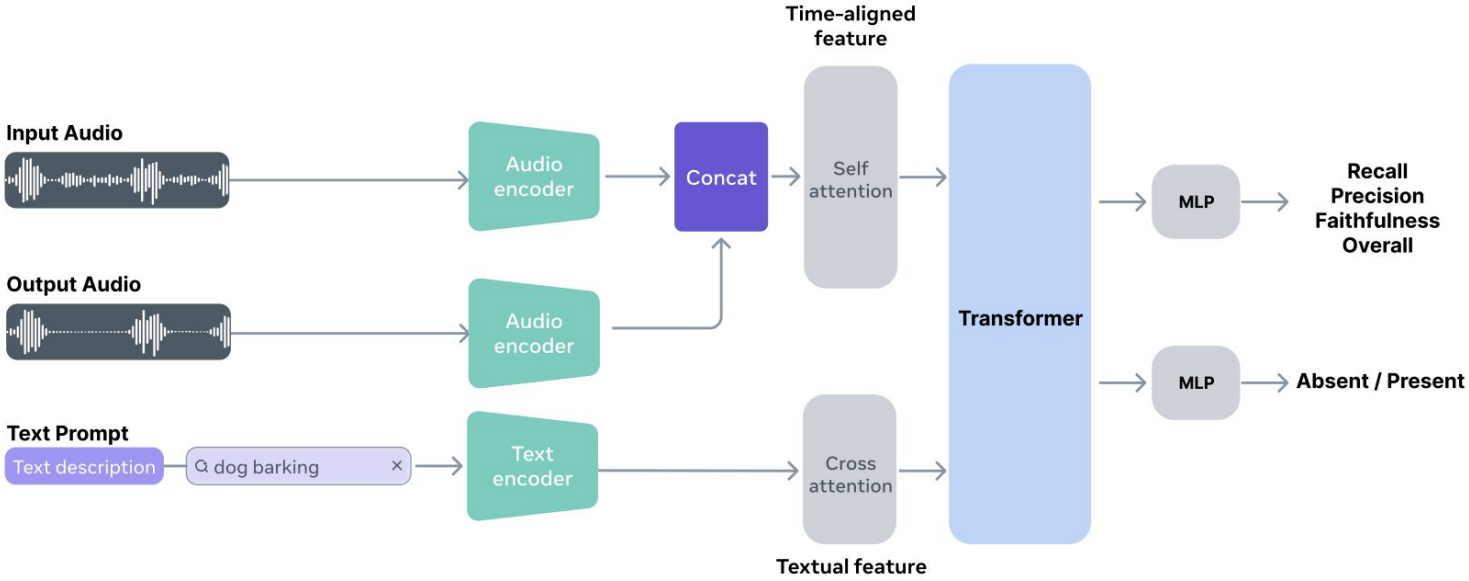
Confusion
How easily can the non-target sounds be mistaken for the target sounds?

Overall difficulty
Considering all the above factors, how difficult is it to extract the target sounds from the mixture?

Why Measure Difficulty?

By assessing the difficulty of each separation task, SAJ provides a nuanced understanding of model performance, revealing strengths and weaknesses across different acoustic conditions.

SAJ Model Overview



SAM Audio Judge

Model	Speech				Music				Sound			
	Overall	Recall	Precision	Faithfulness	Overall	Recall	Precision	Faithfulness	Overall	Recall	Precision	Faithfulness
Pearson Correlation Coefficient (PCC)												
CLAP	0.490	0.431	0.283	0.477	0.487	0.416	0.385	0.432	0.367	0.431	0.283	0.418
SDR Estimator	0.336	0.004	0.403	0.055	0.369	0.157	0.388	0.182	0.181	0.040	0.222	0.055
Gemini-2.5-pro	0.487	0.498	0.169	0.430	0.351	0.287	0.115	0.303	0.462	0.493	0.192	0.369
SAM Audio Judge	**0.883**	**0.943**	**0.841**	**0.891**	**0.815**	**0.858**	**0.766**	**0.791**	**0.815**	**0.837**	**0.775**	**0.818**
Spearman Rank Correlation Coefficient (SRCC)												
CLAP	0.380	0.291	0.325	0.273	0.285	0.293	0.199	0.296	0.493	0.376	0.388	0.406
SDR Estimator	0.338	0.000	0.395	0.079	0.390	0.203	0.375	0.210	0.173	0.053	0.220	0.073
Gemini-2.5-pro	0.495	0.361	-0.015	0.117	0.338	0.232	0.010	0.008	0.390	0.324	-0.006	0.180
SAM Audio Judge	**0.817**	**0.573**	**0.774**	**0.573**	**0.714**	**0.569**	**0.658**	**0.476**	**0.781**	**0.660**	**0.734**	**0.607**

Table 16 Comparison Between SAM Audio Judge Model and Baselines

Text-Prompted Results

Model	OSS	Promptable	General SFX			Speech			Speaker			Music			Instr(wild)			Instr(pro)		
			SAJ	CLAP	OVR	SAJ	CLAP	OVR	SAJ	CLAP	OVR	SAJ	CLAP	OVR	SAJ	CLAP	OVR	SAJ	CLAP	OVR
MossFormer2 (Zhao et al., 2024)	✓	✗	-	-	-	-	-	-	2.43	0.14	2.54	-	-	-	-	-	-	-	-	-
Tiger (Xu et al., 2024)	✓	✗	-	-	-	-	-	-	2.47	0.15	2.50	-	-	-	-	-	-	-	-	-
Fast-GeCo (Wang et al., 2024)	✓	✗	-	-	-	-	-	-	2.66	0.16	2.71	-	-	-	-	-	-	-	-	-
Demucs (Rouard et al., 2023)	✓	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-		4.48	0.15	4.26
Spleeter (Hennequin et al., 2020)	✓	✗	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	4.26	0.11	3.90
FlowSep (Yuan et al., 2025)	✓	✓	2.36	0.21	2.65	2.18	0.20	2.14	1.85	0.09	2.13	2.73	0.18	2.90	2.37	0.10	2.69	2.13	-0.01	2.02
AudioSep (Liu et al., 2022)	✓	✓	2.63	0.25	2.88	2.93	0.28	2.85	2.50	0.17	2.79	3.47	0.27	3.51	2.16	0.13	2.59	2.34	0.04	2.45
CLAPSep (Ma et al., 2024)	✓	✓	2.68	0.23	2.92	2.30	0.22	2.47	2.80	0.17	2.79	2.48	0.04	2.97	2.47	0.14	2.81	2.48	0.04	2.56
SoloAudio (Wang et al., 2025a)	✓	✓	3.29	0.25	2.97	3.45	0.30	3.32	2.26	0.19	2.45	2.68	0.21	2.47	2.92	0.13	2.71	2.65	0.01	2.30
AudioShake (Audioshake, 2025)	✗	✗	-	-	-	3.90	0.28	3.95	3.28	0.14	3.51	3.22	0.29	3.37	3.37	0.29	3.43	3.87	0.29	4.28
MoisesAI (Moises.AI, 2025)	✗	✗	-	-		-	-		-	-	-	3.79	0.27	3.90	3.03	0.29	3.12	3.78	0.28	4.22
FADR (FADR, 2025)	✗	✗	-	-	-	-	-	-	-	-	-	-	-	-	2.44	0.19	2.45	3.63	0.25	3.92
LalalAI (Lalal.AI, 2025)	✗	✗	-	-	-	3.77	0.33	3.92	-	-	-	-	-	-	3.07	0.25	3.03	3.83	0.27	4.18
Auphonic (Auphonic, 2025)	✗	✗	-	-	-	4.32	0.27	4.08	-	-	-	-	-	-	-	-	-	-	-	-
ElevenLabs (ElevenLabs, 2025)	✗	✗	-	-	-	3.79	0.25	3.72	-	-	-	-	-	-	-	-	-	-	-	-
SAM AUDIO	✓	✓	**4.35**	**0.31**	**3.59**	**4.67**	**0.35**	**4.29**	**4.51**	**0.18**	**4.15**	**4.45**	**0.26**	**4.05**	**4.32**	**0.31**	**4.00**	**4.82**	**0.28**	**4.45**

Table 10 Comparison against text-prompted baselines. –: not applicable. OVR: overall subjective score.

Text-Prompted Results

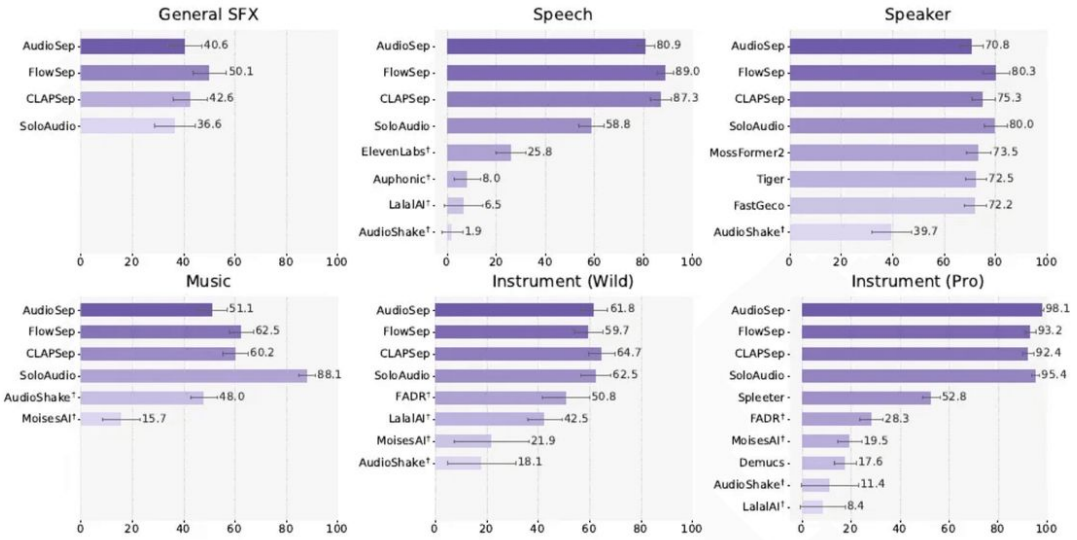


Figure 7 Net Win Rate (%) of SAM Audio against SoTA separation models in text-prompted tasks. †: proprietary models

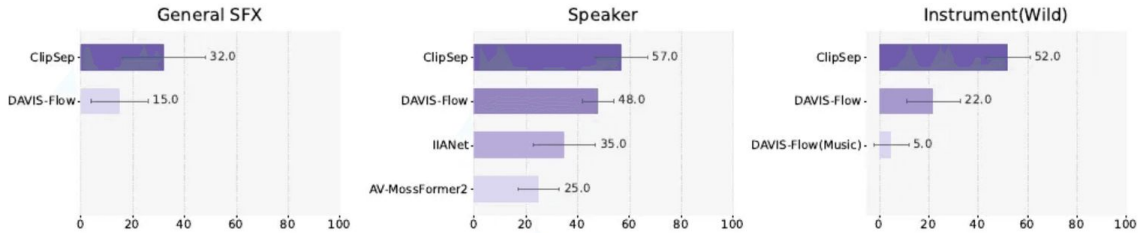
Visual-Prompted Results

Through unified training, SAM AUDIO generalize across domains and achieves SoTA performance.

7.2 Visual-prompted separation

Model	Generic	General SFX		Speaker		Instr (wild)	
		IB	OVR	IB	OVR	IB	OVR
AV-MossFormer2 (Zhao et al., 2025)	✗	-	-	0.20	2.62	-	-
IIANet (Li et al., 2024b)	✗	-	-	0.16	2.41	-	-
ClipSep (Dong et al., 2023)	✓	0.16	1.53	0.14	1.47	0.15	1.12
DAVIS-Flow (Huang et al., 2025)	✓	0.14	1.96	0.13	1.97	0.13	2.08
DAVIS-Flow (Music) (Huang et al., 2025)	✗	-	-	-	-	0.13	2.40
SAM AUDIO	✓	**0.25**	**2.61**	**0.24**	**3.07**	**0.24**	**2.56**

Table 11 Comparison against visual-prompted baselines. –: Not applicable. OVR: overall subjective score.



Span-Prompted

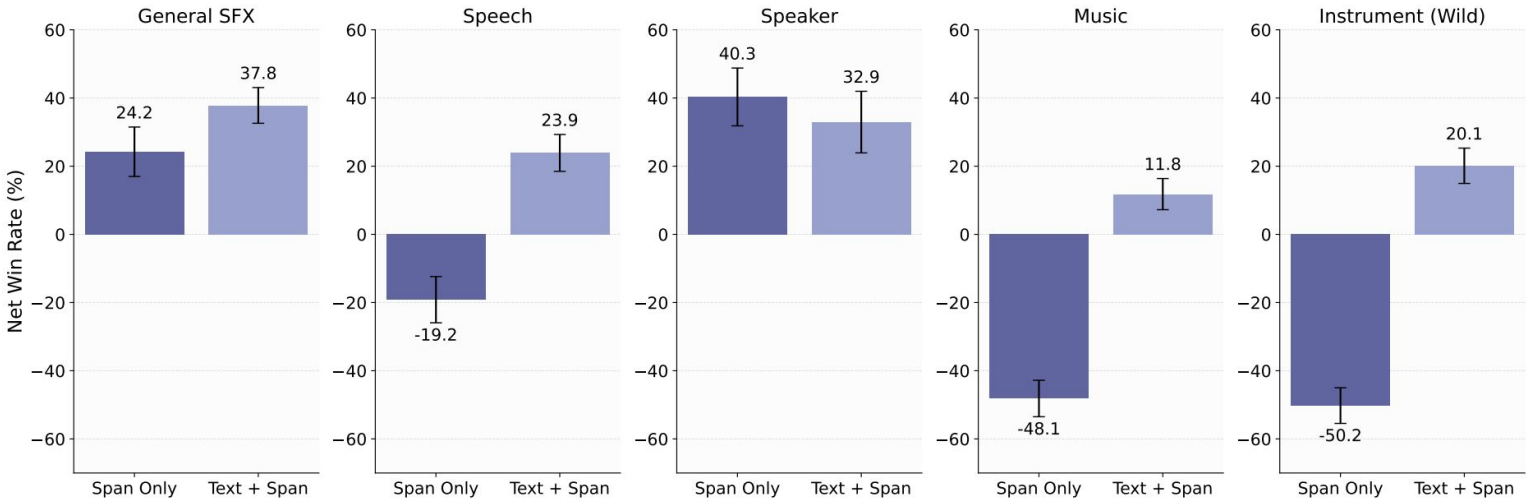


Figure 10 Net Win Rate (%) of SAM Audio with text & span / span as input against a text-only model

Span prediction

Task	w/ Pred Span	SAJ	CLAP	OVR
General SFX	✗	4.11	**0.31**	3.36
	✓	**4.35**	**0.31**	**3.89**
Speech	✗	4.59	0.33	4.17
	✓	**4.67**	**0.35**	**4.22**
Speaker	✗	4.08	0.17	3.62
	✓	**4.51**	**0.18**	**4.01**
Music	✗	4.30	**0.28**	**4.16**
	✓	**4.45**	0.26	4.12
Instr(wild)	✗	**4.45**	0.30	3.70
	✓	4.32	**0.31**	**3.88**
Instr(pro)	✗	**4.83**	**0.28**	**4.16**
	✓	4.82	**0.28**	4.12

Table 13 Using vs. not using predicted span for text-prompting. OVR: overall subjective score. For all the metrics below, higher is better.

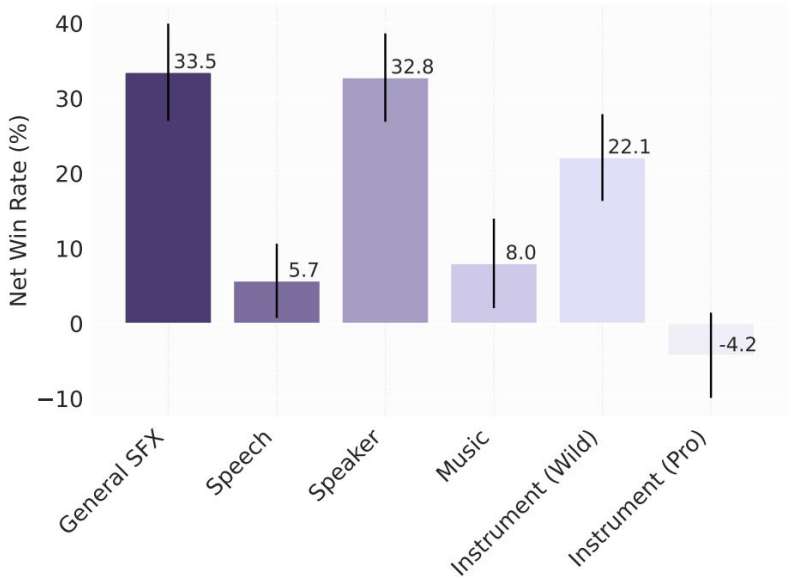


Figure 11 Net Win Rate (%) of SAM AUDIO using predicted span against not using predicted span for text prompting.

Sound Removal

Model	OVR
AudioShake (Audioshake, 2025)	3.75
MoisesAI (Moises.AI, 2025)	4.00
SAM AUDIO	**4.05**

Table 14 Comparison against baselines for music removal

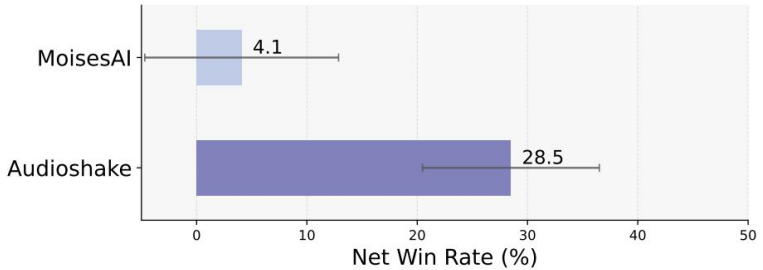


Figure 12 Net Win Rate (%) of SAM AUDIO over baselines in music removal

Latency

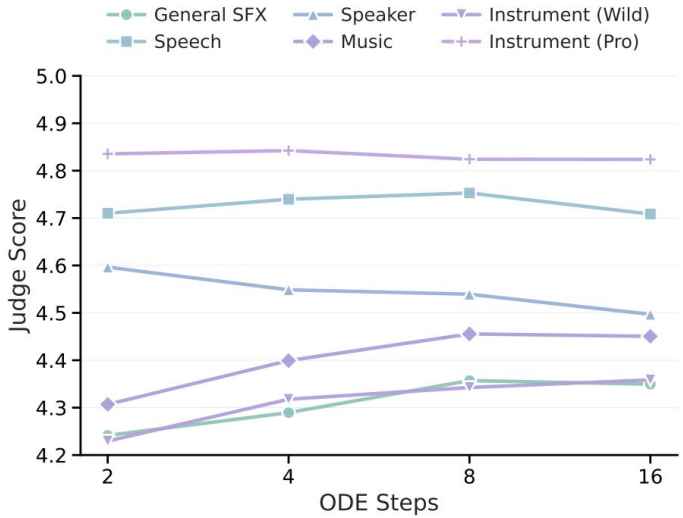


Figure 13 Effect of varying ODE steps under the midpoint solver. Fewer steps reduce computation at a modest cost in quality.

Model scale ablation

Model	General SFX			Speech			Speaker			Music			Instr(wild)			Instr(pro)		
	SAJ	CLAP	OVR	SAJ	CLAP	OVR	SAJ	CLAP	OVR	SAJ	CLAP	OVR	SAJ	CLAP	OVR	SAJ	CLAP	OVR
SAM Audio-Small	**4.25**	0.30	**3.62**	4.55	**0.35**	3.99	3.89	**0.17**	3.12	4.32	**0.28**	4.11	4.27	0.27	3.56	4.78	**0.30**	4.24
SAM Audio-Base	4.23	0.28	3.28	**4.61**	0.33	**4.25**	3.94	0.15	3.57	4.26	0.27	3.87	4.33	0.29	**3.66**	4.78	**0.30**	4.27
SAM Audio-Large	4.11	**0.31**	3.50	4.59	0.33	4.03	**4.08**	**0.17**	**3.60**	**4.30**	**0.28**	**4.22**	**4.45**	**0.30**	**3.66**	**4.83**	0.28	**4.49**

Table 19 Comparison of SAM Audio of different scales in text prompting. –: not applicable. OVR: overall subjective score.

Model	General SFX		Speaker		Instr (wild)	
	IB	OVR	IB	OVR	IB	OVR
SAM Audio-Small	0.24	2.62	0.23	2.79	0.21	2.25
SAM Audio-Base	**0.25**	**2.63**	0.24	**3.25**	0.22	**2.76**
SAM Audio-Large	**0.25**	2.61	**0.24**	2.95	**0.24**	2.58

Table 20 Comparison of SAM Audio of different scales in visual prompting. –: not applicable. OVR: overall subjective score.

Model scale ablation

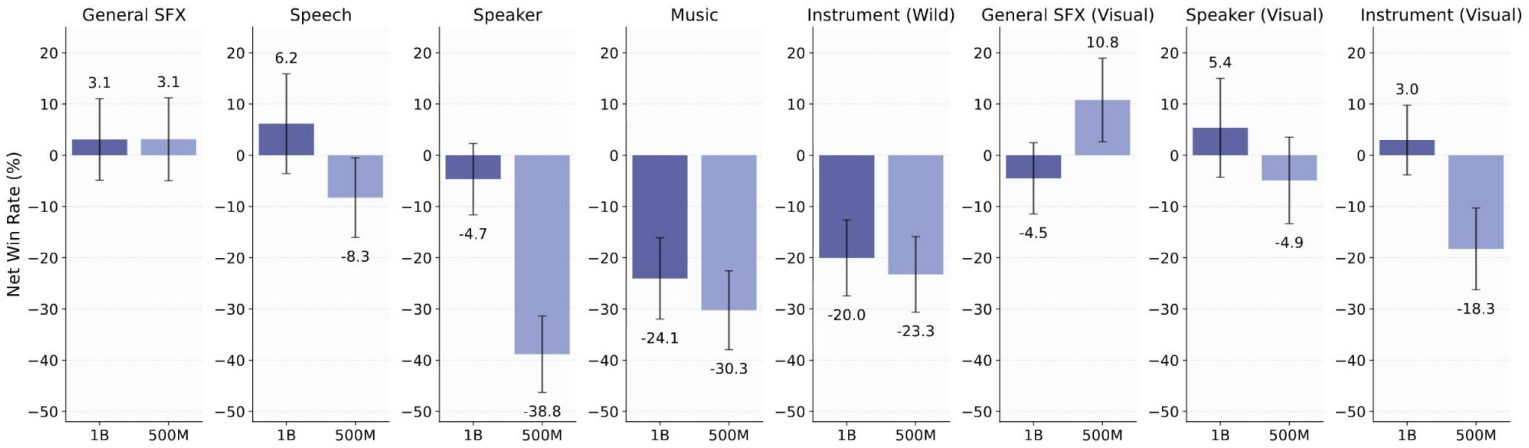


Figure 17 Net Win Rate (%) of SAM Audio-Base and SAM Audio-Small against SAM Audio-Large

Effect of Auxiliary loss

	General SFX (Text)		General SFX (Visual)
Auxiliary target	SAJ	CLAP	IB
None	2.48	0.24	0.17
Target AED	**3.18**	**0.29**	**0.18**

Effect of fine-tuning

		Text						Visual		
Separation Task	**Stage**	General SFX	Speech	Speaker	Music	Instr(wild)	Instr(pro)	General SFX	Speaker	Instr(wild)
SAJ (↑)	PT	3.93	2.90	3.28	4.14	3.17	3.36	-	-	-
	FT	**4.14**	**4.55**	**4.07**	**4.38**	**4.48**	**4.82**	-	-	-
CLAP (↑)	PT	**0.31**	0.28	**0.23**	**0.31**	0.21	0.15	-	-	-
	FT	0.30	**0.36**	0.21	**0.31**	**0.30**	**0.29**	-	-	-
IB (↑)	PT	-	-	-	-	-	-	**0.24**	0.22	0.20
	FT	-	-	-	-	-	-	**0.24**	**0.24**	**0.22**
AES-PC (↓)	PT	2.57	2.91	2.98	4.53	3.54	4.72	3.09	3.49	3.69
	FT	**2.08**	**1.89**	**1.94**	**4.44**	**3.16**	**3.24**	**2.22**	**2.20**	**3.26**

Table 22 Comparison of pre-trained vs fine-tuned results across audio separation tasks. –: not applicable.

Effect of using pseudo-label audio

		Text						Visual		
Separation Task	**Setting**	General SFX	Speech	Speaker	Music	Instr(wild)	Instr(pro)	General SFX	Speaker	Instr(wild)
SAJ (↑)	w/o PL	4.02	4.50	4.06	4.34	4.34	4.80	-	-	-
	PL	**4.14**	**4.55**	**4.07**	**4.38**	**4.48**	**4.82**	-	-	-
CLAP (↑)	w/o PL	**0.30**	0.34	**0.21**	**0.31**	0.29	0.28	-	-	-
	PL	**0.30**	**0.36**	**0.21**	**0.31**	**0.30**	**0.29**	-	-	-
IB (↑)	w/o PL	-	-	-	-	-	-	0.23	0.21	**0.22**
	PL	-	-	-	-	-	-	**0.24**	**0.24**	**0.22**
AES-PC (↓)	w/o PL	2.36	1.95	2.02	**4.38**	3.19	**3.24**	2.28	2.29	**3.25**
	PL	**2.08**	**1.89**	**1.94**	4.44	**3.16**	**3.24**	**2.22**	**2.20**	3.26

Table 23 Effect of using pseudo-labeled audio stem data

Conclusion & Future Work

Key Contributions

- **Unified Foundation Model:** A single model for open-domain audio separation across speech, music, and sound events.
- **Multi-Modal Prompting:** Flexible interaction via text descriptions, visual masks, and temporal spans.
- **Comprehensive Evaluation:** Introduction of SAM Audio-Bench and the SAJ perceptual evaluation model.

Future Directions

- **Real-Time Processing:** Optimizing the flow-matching architecture for low-latency, real-time applications.
- **Complex Interactions:** Extending prompting capabilities to handle more complex, multi-conditional queries.
- **Extreme Conditions:** Improving separation robustness in environments with extreme noise or highly overlapping similar sources.