ETH zürich

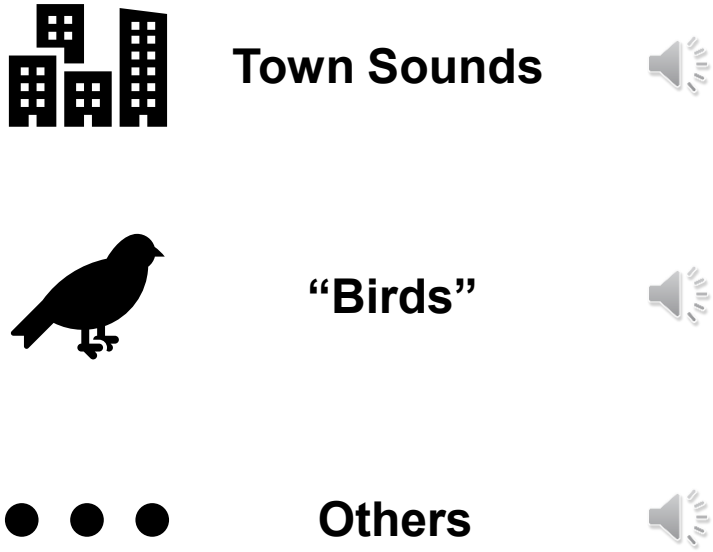
# SAM Audio: Segment Anything in Audio

**Mara Godeanu**

17 March 2026

# What is Audio Segmentation

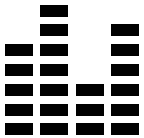
*…identifying different sound sources*



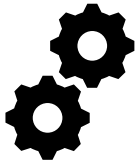

Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, Ann Lee. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

# Why Audio Segmentation matters

**Direct Applications**



**Preprocessing**



**Multimodal AI**

# What‘s holding us back

## Model

**domain-specific models**

...speaker / instrument separation

**limited sound taxonomy**

*...6 instruments, 2 speakers*

**single input modality**

*...text, video*

## Data

**disparate test sets**

... difficult comparison

**incomplete benchmarks**

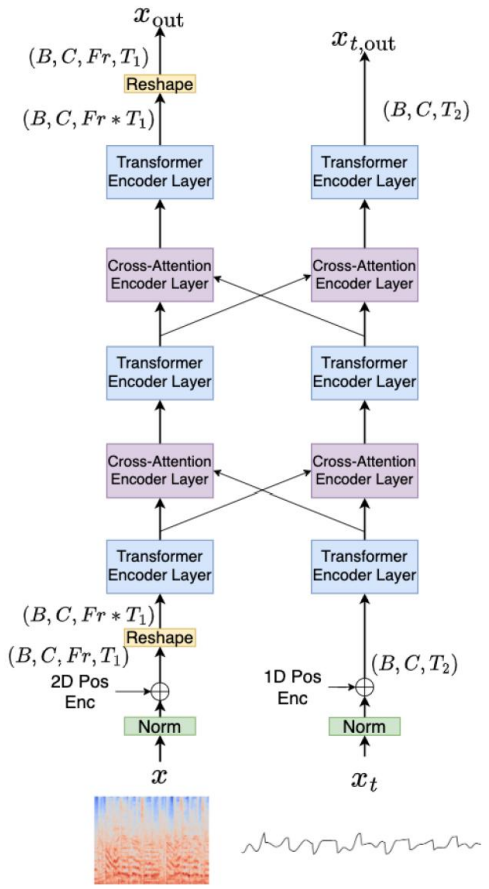
*...few sound types per benchmark*

**poor performance metrics**

*...signal distortion ratio, embedding similarity*

**» General purpose, general data model**

# Demucs v4 – 2022




(b) Cross-domain Transformer Encoder of depth 5

## Mixture

## Limitations

*e.g. Drums*

*specific application domain*

*e.g. Vocals*

*fixed stems*
*(drums, bass, vocals, other)*

*e.g. Other*

*single benchmark,*
*single metric (SDR)*

Simon Rouard, Francisco Massa, Alexandre Défossez. “Hybrid Transformers for Music Source Separation” in arXiv preprint arXiv:2211.08553 (2022).

# LASS (Language-queried Audio Source Separation) – 2022

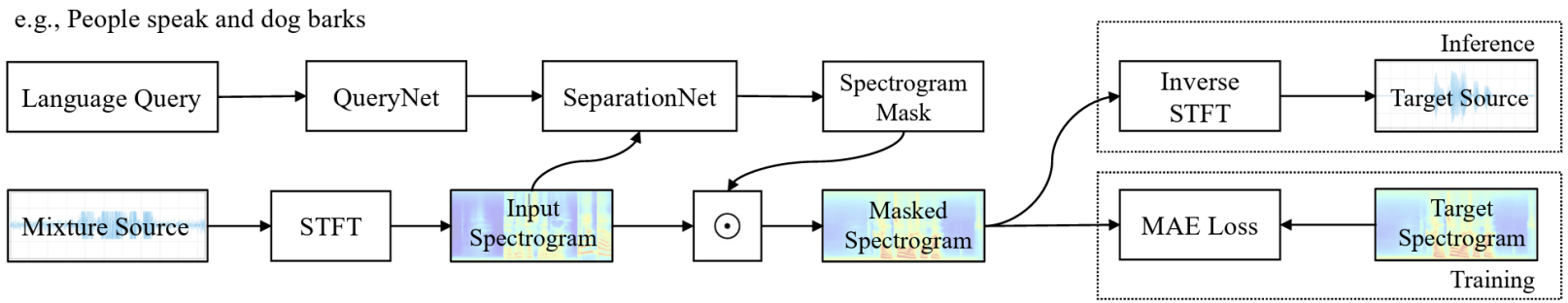



Figure 1: *Framework of our proposed LASS-Net.*

## Model

**QueryNet:**
*BERT-based Encoder*

**SeparationNet:**
*U-Net model with text-embedding conditioning*

## Limitations

*single prompt modality*

*analytic metrics (SDR, SIR etc.)*

*limited test data*

1. Xubo Liu, Haohe Liu, Qiuqiang Kong, Xinhao Mei, Jinzheng Zhao, Qiushi Huang, Mark D. Plumbley, Wenwu Wang. “Separate What You Describe: Language-Queried Audio Source Separation” arXiv preprint arXiv:2203.15147 (2022).
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pretraining of deep bidirectional Transformers for language understanding” arXiv preprint arXiv:1810.04805 (2018).
3. Qiuqiang Kong, Yin Cao, Haohe Liu, Keunwoo Choi, Yuxuan Wang. “Decoupling Magnitude and Phase Estimation with Deep ResUNet for Music Source Separation” arXiv preprint arXiv:2109.05418 (2021).

# Paper‘s Aim

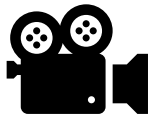
## Inputs

Music

Speech

\+

Video



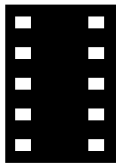
General Audio

## Prompt Modalities

Text

Time Span

Video Part

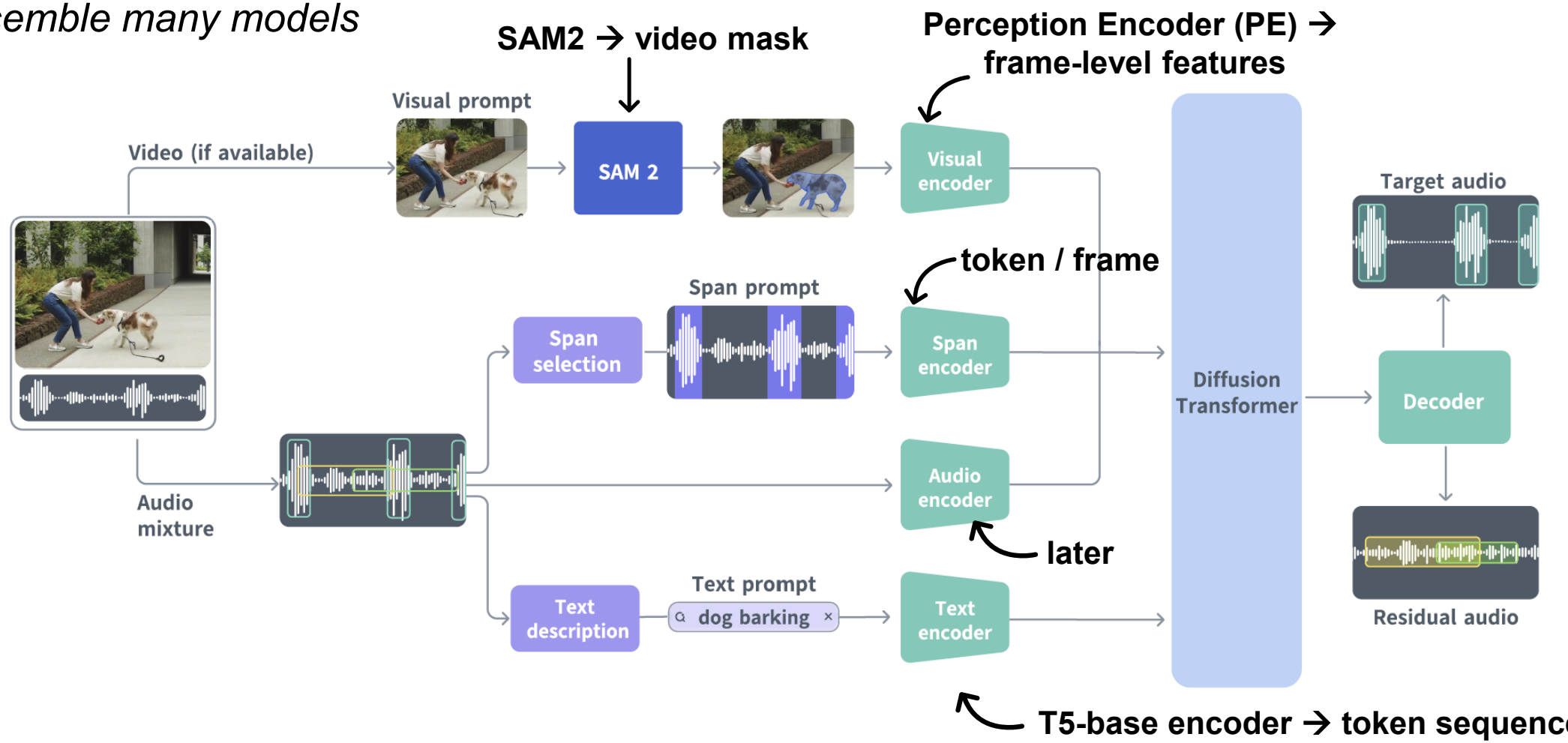


**SAM**

Target Audio + Residual Audio

# How?

*…assemble many models*

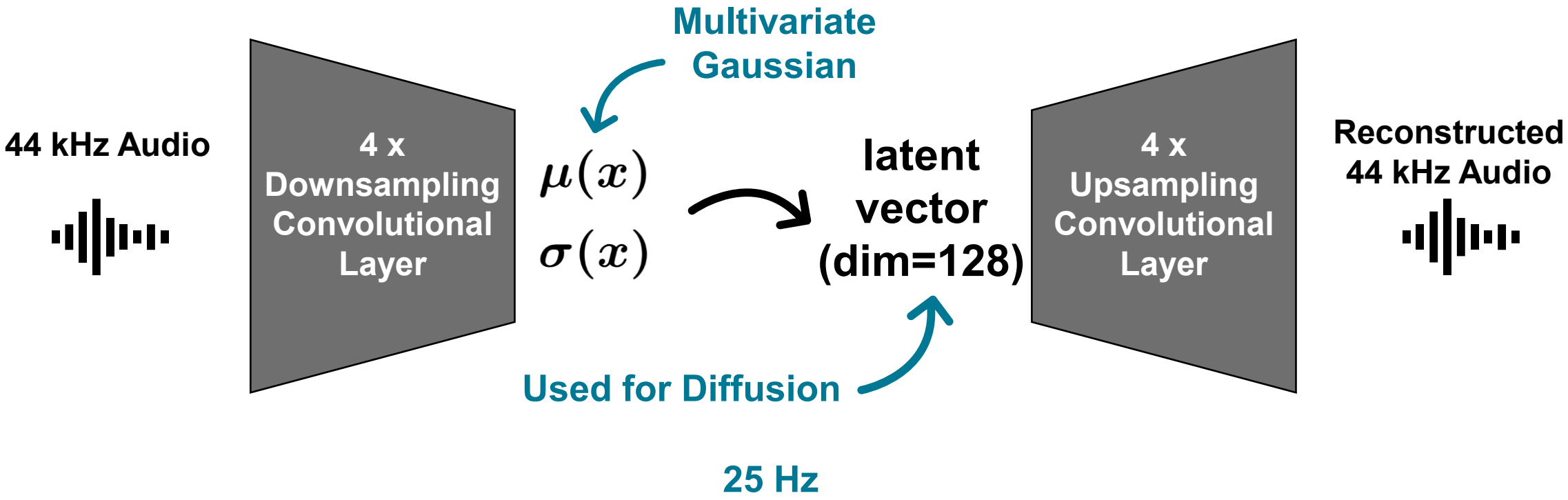




1. Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, Ann Lee. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).
2. Meta FAIR. “SAM2: Segment Anything in Images and Videos” arXiv preprint arXiv:2408.00714 (2024).
3. Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, Christoph Feichtenhofer. “Perception Encoder: The best visual embeddings are not at the output of the network” arXiv preprint arXiv:2504.13181 (2025).
4. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer” in Journal of Machine Learning Research 21 (2020).

# Audio Encoder – DAC-VAE

*…audio is too highly dimensional*

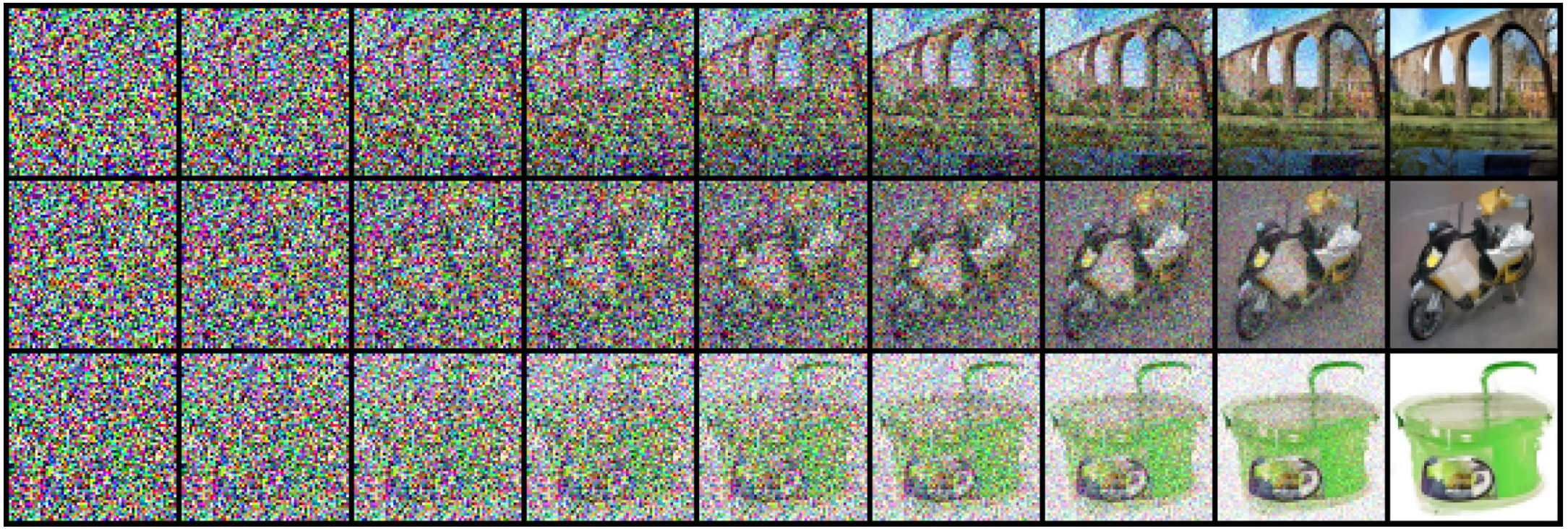




1. Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, Kundan Kumar. “High-Fidelity Audio Compression with Improved RVQGAN” arXiv preprint arXiv:2306.06546v2 (2023).
2. The Movie Gen team@Meta. “Movie Gen: A Cast of Media Foundation Models” arXiv preprint arXiv:2410.13720v2 (2025).

# Diffusion Models

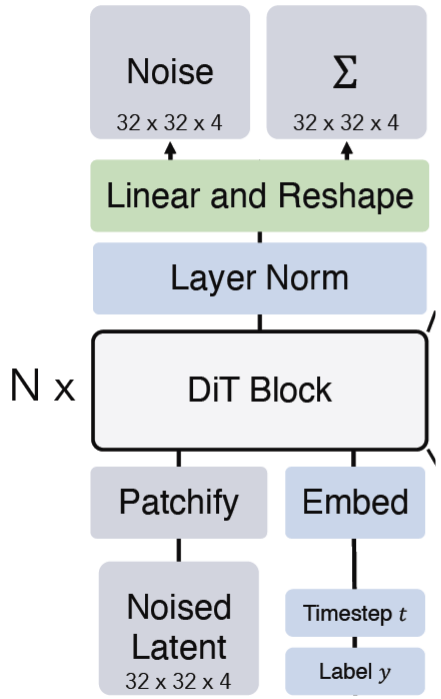
*…from noise to new data*



**Predict noise-to-be-removed iteratively**

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, Matt Le. “Flow Matching for Generative Modeling” arXiv preprint arXiv:2210.02747v2 (2023).

# Diffusion Transformer (DiT)




Latent Diffusion Transforme

Peebles, Saining Xie. “Scalable Diffusion Models with Transformers” arXiv preprint arXiv:2212.09748v2 (2023).

# Audio-Visual Data

**required:** training tuples $(x_{mix},\ x_{tgt},\ x_{res})$

1. **Fully-real triplets:** *music, speech*

2. **Synthetic mixtures:** *mix 2 random audios for different domains*

*…noisy music, noisy speech*
*…general sound (combinations of in-the-wild clips with pro-sound clips)*

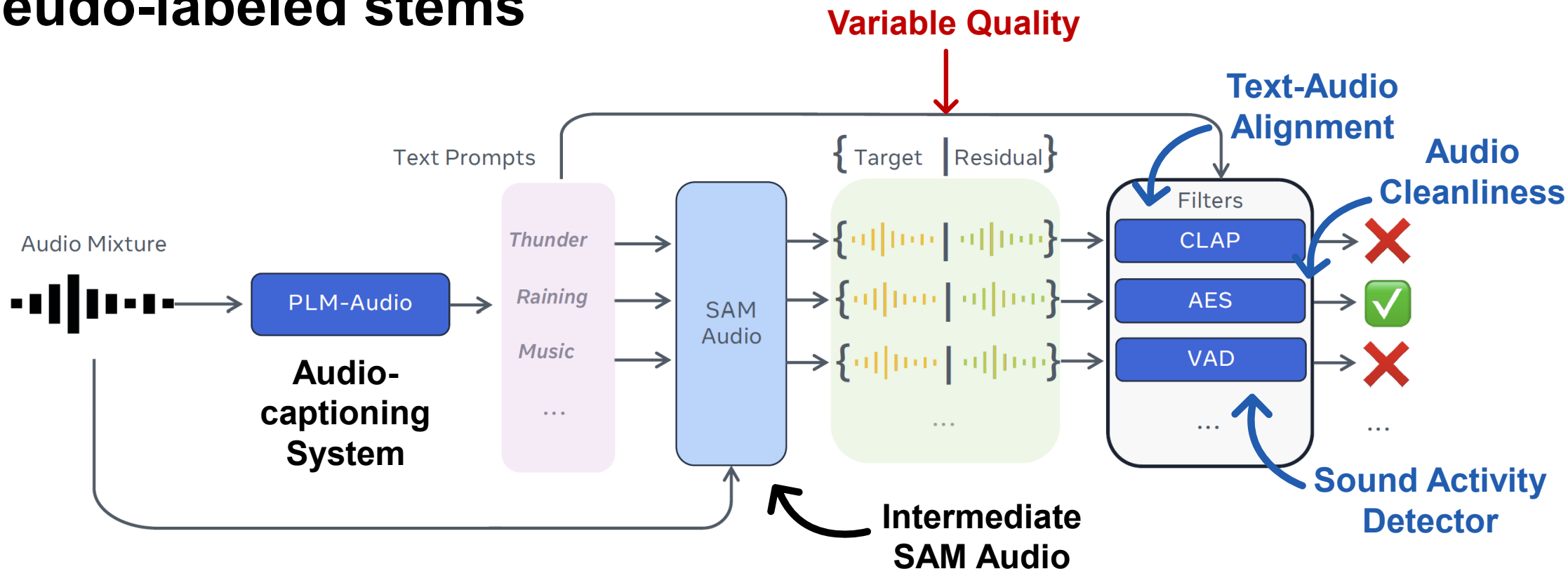
⋙ **unrealistic mixtures**

3. **Pseudo-labeled stems:** *real audio, model-based labeling*

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

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. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

# Pseudo-labeled stems




| Data Source | Modality | Synthetic Stems | Sound | Music | Speech | #Samples (M) | #Hours (K) |
|---|---|---|---|---|---|---|---|
| PL-Audio | audio | ✓ | ✓ | ✓ | ✓ | $\mathcal{O}(1)$ | $\mathcal{O}(1)$ |
| PL-Video | audio, video | ✓ | ✓ | ✓ | ✓ | $\mathcal{O}(0.1)$ | $\mathcal{O}(0.1)$ |


1. 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. "SAM Audio: Segment Anything in Audio" arXiv preprint arXiv:2512.18099v1 (2025).
2. Meta Superintelligence Labs. "Pushing the Frontier of Audiovisual Perception with Large-Scale Multimodal Correspondence Learning" in arXiv preprint arXiv:2512.19687 (2025).
3. Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, Shlomo Dubnov. "Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation" in arXiv preprint arXiv:2211.06687 (2022).

# Prompts

**we have:** training tuples $(x_{mix}, x_{tgt}, x_{res})$

**required:** prompts $c \in \{\text{text}, \text{span}, \text{video}\}$

## 1. Text Prompts

**initial caption** from PLM-Audio

↓

use LLM to merge the **initial caption** with **metadata**

↓

instruct LLM to extract only **NP/VP descriptions** (noun-verb phrase)

↓

**finally:** filter by text-audio similarity (CLAP)

Meta Superintelligence Labs. “Pushing the Frontier of Audiovisual Perception with Large-Scale Multimodal Correspondence Learning” in arXiv preprint arXiv:2512.19687 (2025).

# Prompts

**we have:** training tuples $(x_{mix}, x_{tgt}, x_{res})$

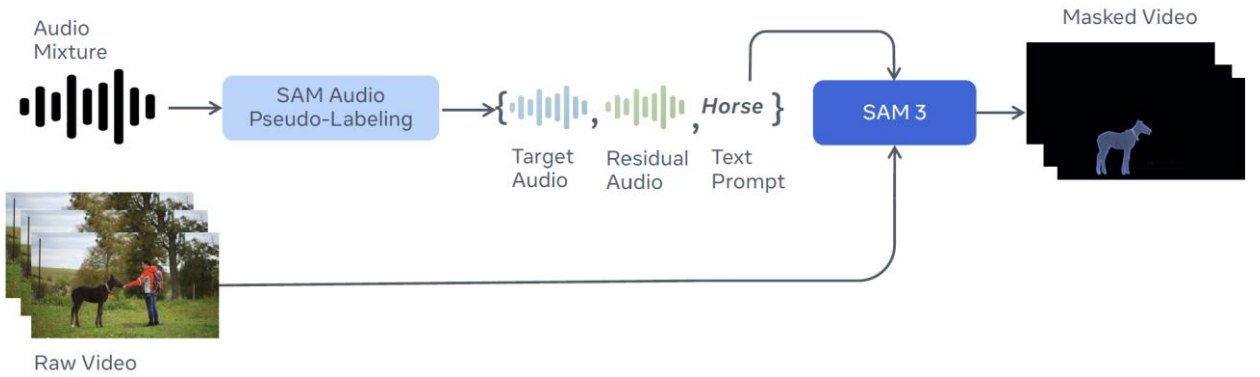
**required:** prompts $c \in \{\text{text}, \text{span}, \text{video}\}$

## 2. Video Prompts

**whole-video:**

1. remove videos with unrelated sound (ImageBind similarity)
2. model learns weakly supervised audio-video associations

**visual mask:**




**filter** bad masked video (ImageBind score)


1. Meta Superintelligence Labs. “SAM3: Segment Anything with Concepts” in arXiv preprint arXiv:2511.16719 (2025).
2. Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra. “Imagebind: One embedding space to bind them all.” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023).
3. 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. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

# Prompts

**we have:** training tuples $(x_{mix},\ x_{tgt},\ x_{res})$

**required:** prompts $c \in \{\text{text}, \text{span}, \text{video}\}$

## 3. Span Prompts

apply VAD (sound activity) for a **frame-level mask**

↓

convert consecutive sounding segments into **intervals**

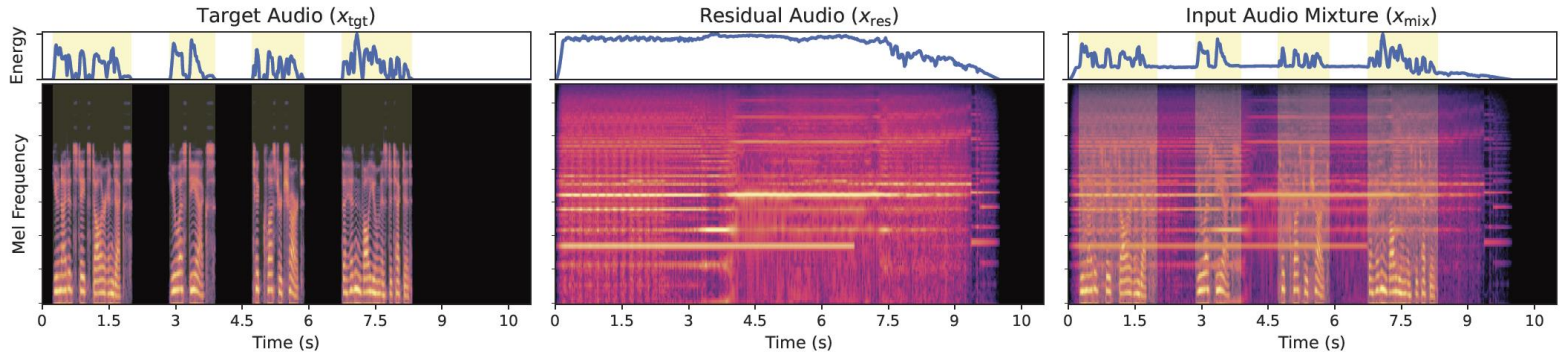



**Figure 4** Illustration of span generation. RMS energy (top) and Mel-spectrograms (bottom) for the target $x_{\text{tgt}}$, residual $x_{\text{res}}$, and mixture $x_{\text{mix}}$. Yellow intervals denote detected spans corresponding to active sound events.


1. 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. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).
2. pydub. pydub. https://github.com/jiaaro/pydub.

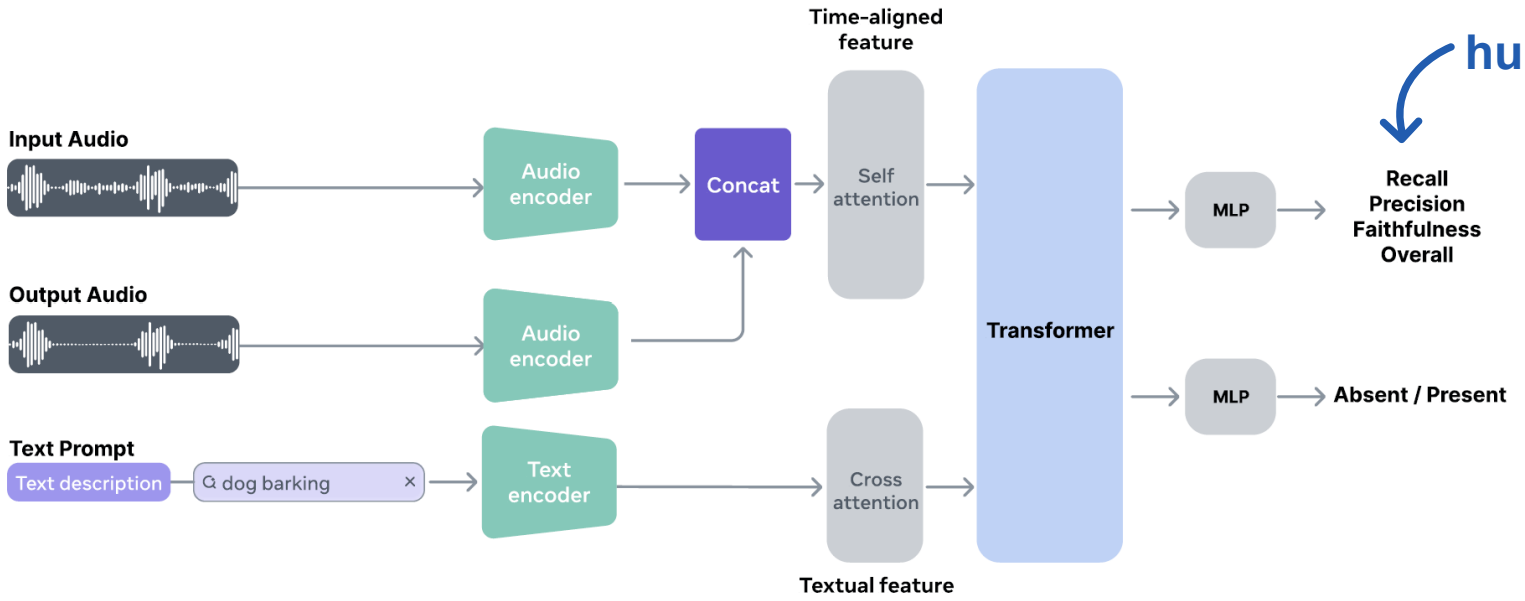
# SAJ: SAM Audio Judge

**we have:** training tuples $(x_{mix},\ x_{tgt},\ x_{res})$, prompts $c \in \{\text{text}, \text{span}, \text{video}\}$

**required: evaluation framework**

1. requires **no reference signal**
2. enables **fine-grained assessment** of segmented audio
3. exhibits **strong correlation with human listening judgement**

**measured against human annotator scores**




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. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

# Training

| Data Source | Modality | w/ stem | Quality | Sound | Music | Speech | #Samples (M) | #Hours (K) |
|---|---|---|---|---|---|---|---|---|
| General Video | audio, video | ✗ | Medium | ✓ | ✓ | ✓ | $\mathcal{O}(100)$ | $\mathcal{O}(1000)$ |
| General Audio | audio | ✗ | Medium | ✓ | ✓ | ✓ | $\mathcal{O}(1)$ | $\mathcal{O}(1)$ |
| Speech Conversation | audio | ✓ | High | ✗ | ✗ | ✓ | $\mathcal{O}(10)$ | $\mathcal{O}(10)$ |
| HQ Music | audio | ✗ | High | ✗ | ✓ | ✗ | $\mathcal{O}(10)$ | $\mathcal{O}(10)$ |
| Multi-track Music | audio | ✓ | High | ✗ | ✓ | ✗ | $\mathcal{O}(0.1)$ | $\mathcal{O}(1)$ |
| HQ SFX | audio | ✗ | High | ✓ | ✓ | ✓ | $\mathcal{O}(10)$ | $\mathcal{O}(10)$ |
| HQ Video | audio, video | ✗ | High | ✓ | ✓ | ✓ | $\mathcal{O}(0.1)$ | $\mathcal{O}(0.1)$ |

**Pre-training:** synthetic audio mixtures from general video
**Fine-tuning:** high-quality and pseudo-labeled data

| 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 |

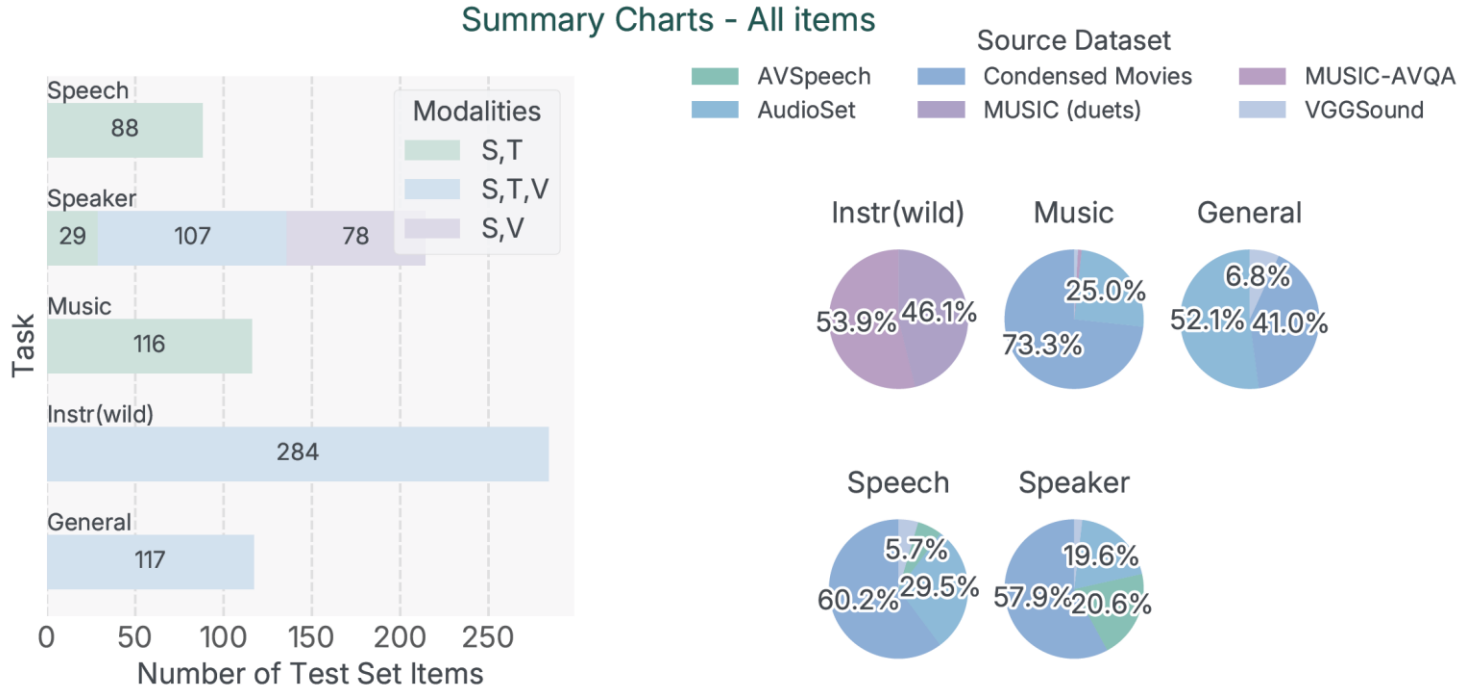
**Inference:** make 8 predictions, choose best one (SAJ, CLAP)

**Best model:** SAM Audio-Large

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. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

# SAM Audio-Bench

*… a real, in-the-wild, multi-modal separation benchmark*




**ecological validity:** in-the-wild audio/video or production-quality video

**multi-modal prompting:** human annotated masks, temporal spans, text descriptions

**taxonomic coverage:** speech cleaning, speaker separation, music cleaning/removal, instrument stems, general sounds with annotator-extendable stem taxonomy

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. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

# Results: Span-prompted separation

| Prompt Modality | General | | | Speech | | | Speaker | | | Music | | | Instr(wild) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | SAJ | CLAP | OVR | SAJ | CLAP | OVR | SAJ | CLAP | OVR | SAJ | CLAP | OVR | SAJ | CLAP | OVR |
| text | 4.11 | 0.31 | 3.32 | 4.59 | 0.33 | 4.18 | 4.08 | 0.17 | 3.63 | 4.30 | 0.28 | 4.09 | **4.45** | 0.30 | 3.78 |
| span | 3.27 | 0.30 | 3.54 | 3.37 | 0.28 | 3.65 | 3.26 | 0.26 | 4.08 | 2.18 | -0.04 | 2.57 | 1.81 | 0.01 | 2.20 |
| text + span | **4.25** | **0.31** | **4.04** | **4.66** | **0.35** | **4.33** | **4.51** | **0.18** | **4.22** | **4.38** | **0.27** | **4.19** | 4.42 | **0.32** | **3.88** |

**Table 12** Comparison between span prompting and text prompting for SAM AUDIO. OVR: overall subjective score. For all the metrics below, higher is better.

| 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.

**OVR (overall subjective score):**
overall preference scores across all pairwise comparisons involving SAM Audio

Always use **predicted span** for text-prompting

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. "SAM Audio: Segment Anything in Audio" arXiv preprint arXiv:2512.18099v1 (2025).

# Results: Text-prompted separation

| 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** |

**OVR (overall subjective score):**
overall preference scores across all pairwise comparisons involving SAM Audio

**SAM** outperforms both *general* and *specialized* models

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. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

# Results: 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** |

**OVR (overall subjective score):**
overall preference scores across all pairwise comparisons involving SAM Audio

**IB:** ImageBind score

\+ **SAM** again outperforms all other *general* and *specialized* models

− **poorer performance** than in text-prompted 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. "SAM Audio: Segment Anything in Audio" arXiv preprint arXiv:2512.18099v1 (2025).

# Results: 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** |

**SAM Audio Judge** outperforms baselines

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. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

# Thoughts

**the good:** genuinely impressive scope and effort

reusable tools (SAJ, SAM Audio Benchmark)

**the bad:** same struggles (video separation, general SFX)

limited benchmarking *general SFX: e.g. dog barking*

*speaker extraction: primarily gender-based*

*instrument separation: single instrument*

biased benchmark (e.g. speaker separation)

unclear data pipeline

# Thank you for your attention!

# Extra Slides – SAM Audio Models

| 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** |

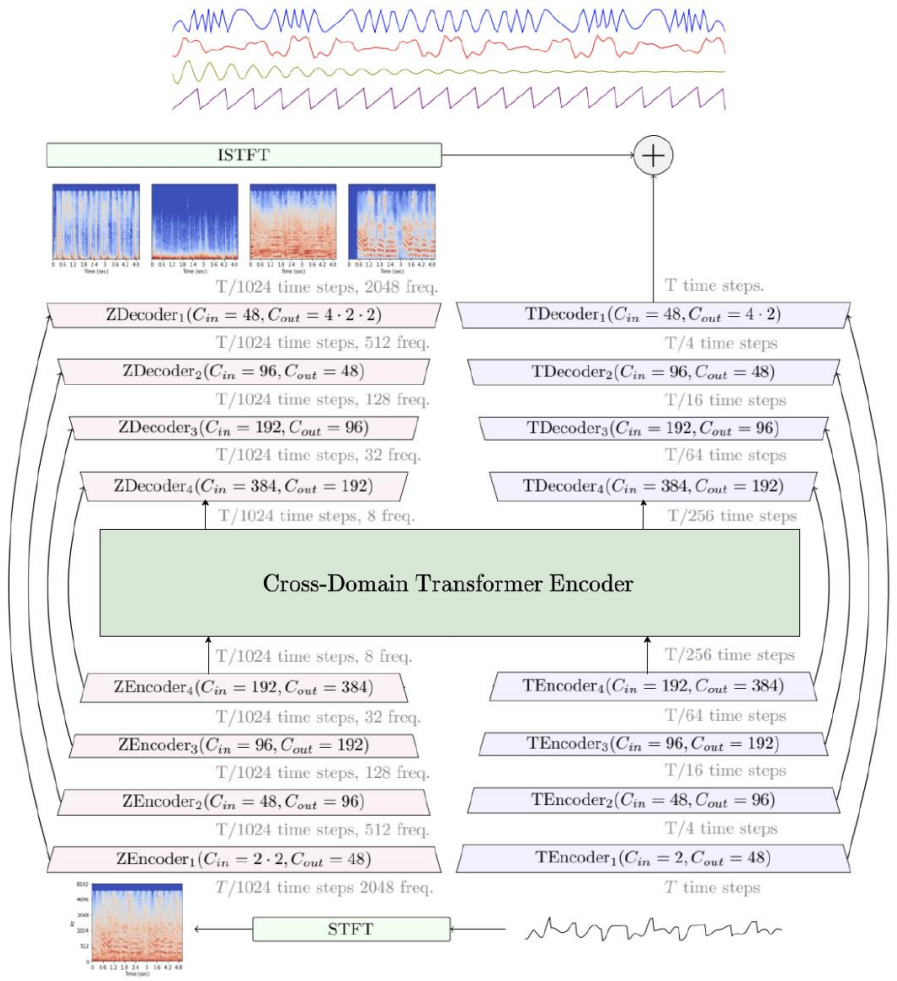
**Table 19** Comparison of SAM Audio of different scales in text prompting. –: not applicable. OVR: overall subjective score.

| 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.

Wei-Ning Hsu, Ann Lee. “SAM Audio: Segment Anything in Audio” arXiv preprint arXiv:2512.18099v1 (2025).

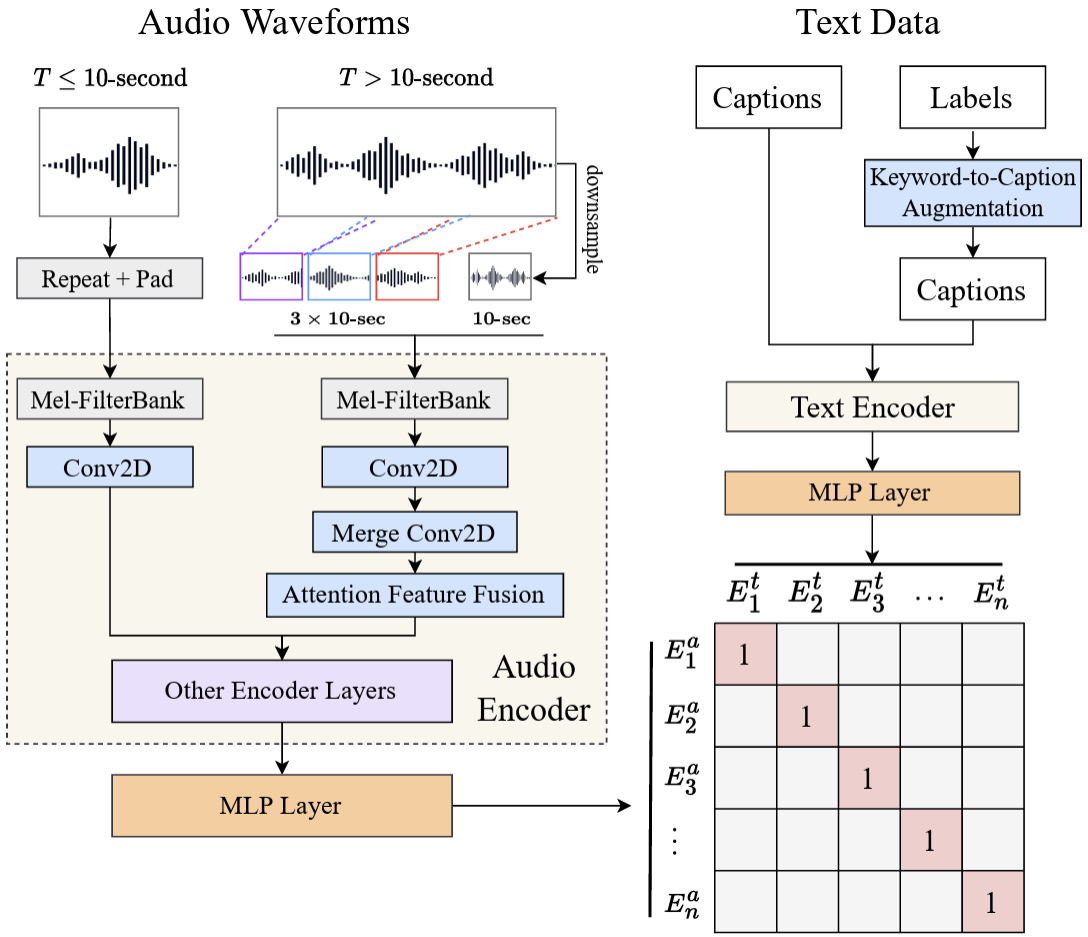
# Extra Slides - Demucs




(c) Hybrid Transformer Demucs

Simon Rouard, Francisco Massa, Alexandre Défossez. “Hybrid Transformers for Music Source Separation” in arXiv preprint arXiv:2211.08553 (2022).

# Extra Slides - CLAP




Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, Shlomo Dubnov. “Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation” in arXiv preprint arXiv:2211.06687 (2022).

# Extra Slides – SAM3

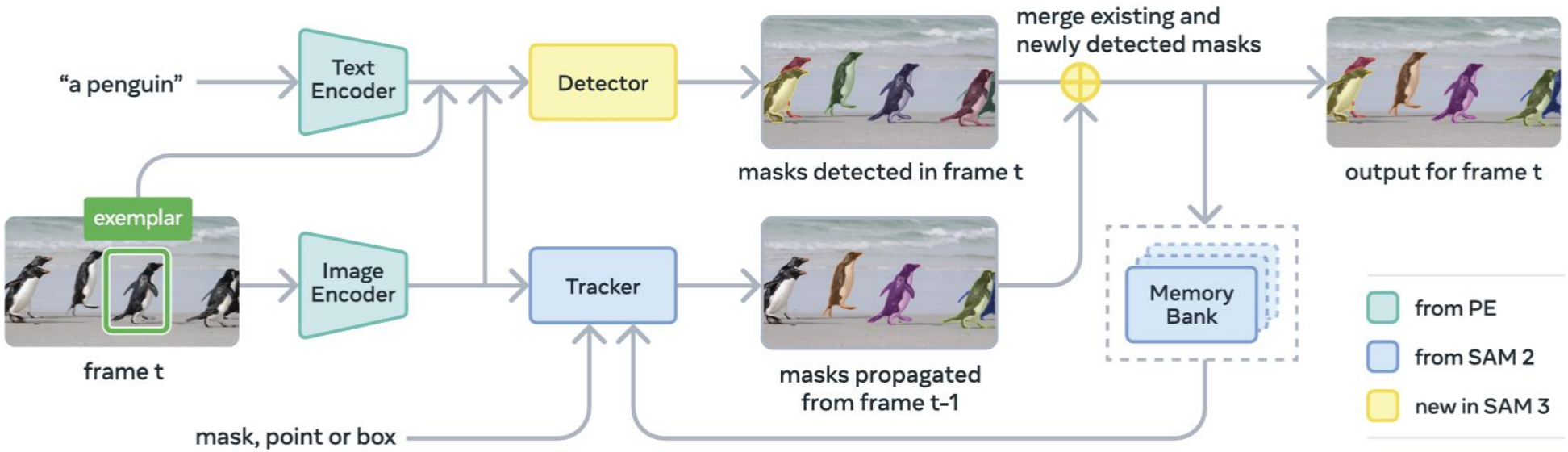



**Figure 4** SAM 3 architecture overview. See Fig. 10 for a more detailed diagram.

Meta Superintelligence Labs. "SAM3: Segment Anything with Concepts" in arXiv preprint arXiv:2511.16719 (2025).