CVPR JUNE 3-7, 2026 DENVER COLORADO

AV-CASS

KAIST UNIST

Scan QR code to enjoy 30+ real-world results!

Cinematic Audio Source Separation Using Visual Cues

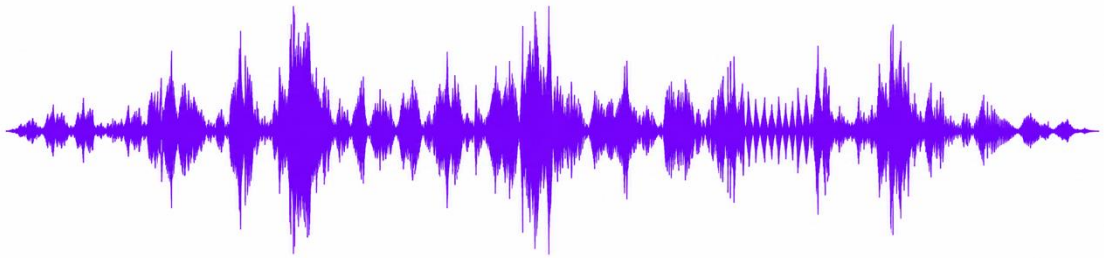
Kang Zhang[1*], Suyeon Lee[1*], Arda Senocak[2†], Joon Son Chung[1†]

* Equal contribution, † Corresponding authors

[1]School of Electrical Engineering, KAIST

[2]Graduate School of Artificial Intelligence, UNIST

What is the task?



Movie soundtrack

+

Movie frames

Decomposes a **movie soundtrack** into **dialogue**, **sound effects**, and **music**

Dialogue

Sound effects

Music

Challenges & Our Solutions

Audio-Only Limitations

Current CASS ignores the inherent audio-visual nature of film

Dual-Stream Video

Clearer source designation
- Speech ↔ Lip movements
- Sound effects ↔ Objects / actions / place

Audio-Visual Dataset Scarcity

Limited access to video-paired GT audio tracks in the real-world videos

Training Data Synthesis Pipeline

Utilize off-the-shelf & individually available audio-visual data for CASS

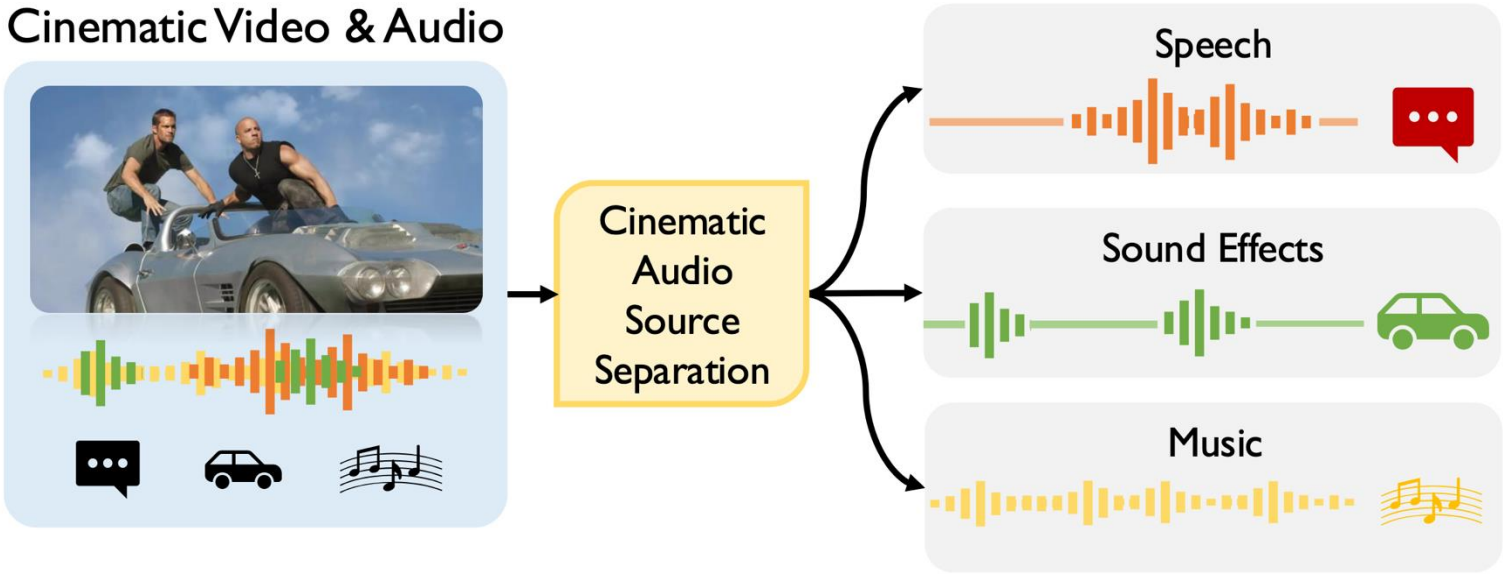
Artifacts due to Masking-based Separator

Discriminative, masking-based separator leaves artifacts (spectral holes) on the spectrogram

Generative Separator

Flow-matching-based separator generates natural sounds w/o artifacts

Why Is CASS Different?



Multiple generation target needs corresponds to different part of the condition

- Dialogue is often tied to faces and lip motion.
- Sound effects are often tied to visible actions or objects.
- Music is usually less visually grounded and acts as background structure.

Data problem

Ideal supervised

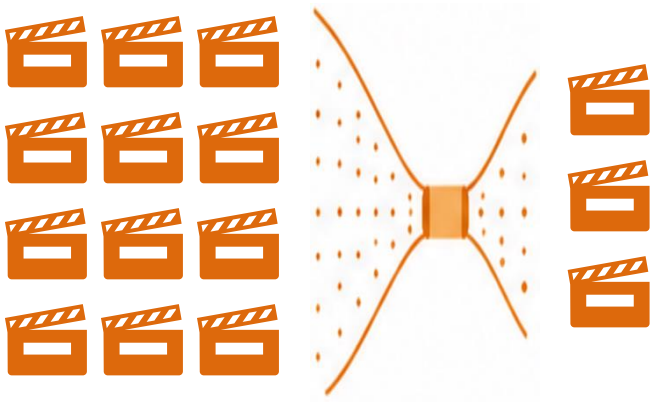
Dialogue

Sound effects

Music

Aligned video + clean stem

Data bottleneck



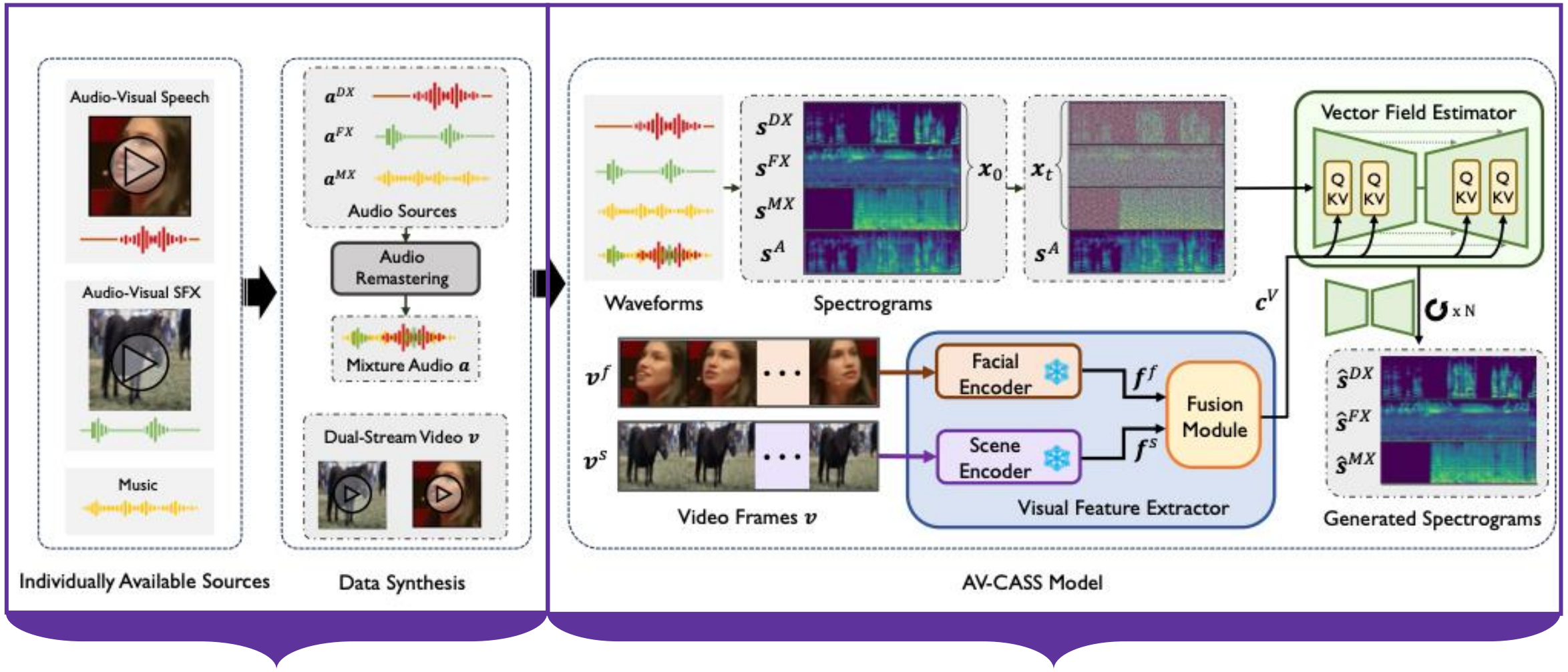
- **Clean stem are rare**
- **Alignment is hard**
- **Real mixtures are complex**

Implication

Purely supervised AV learning is not directly available

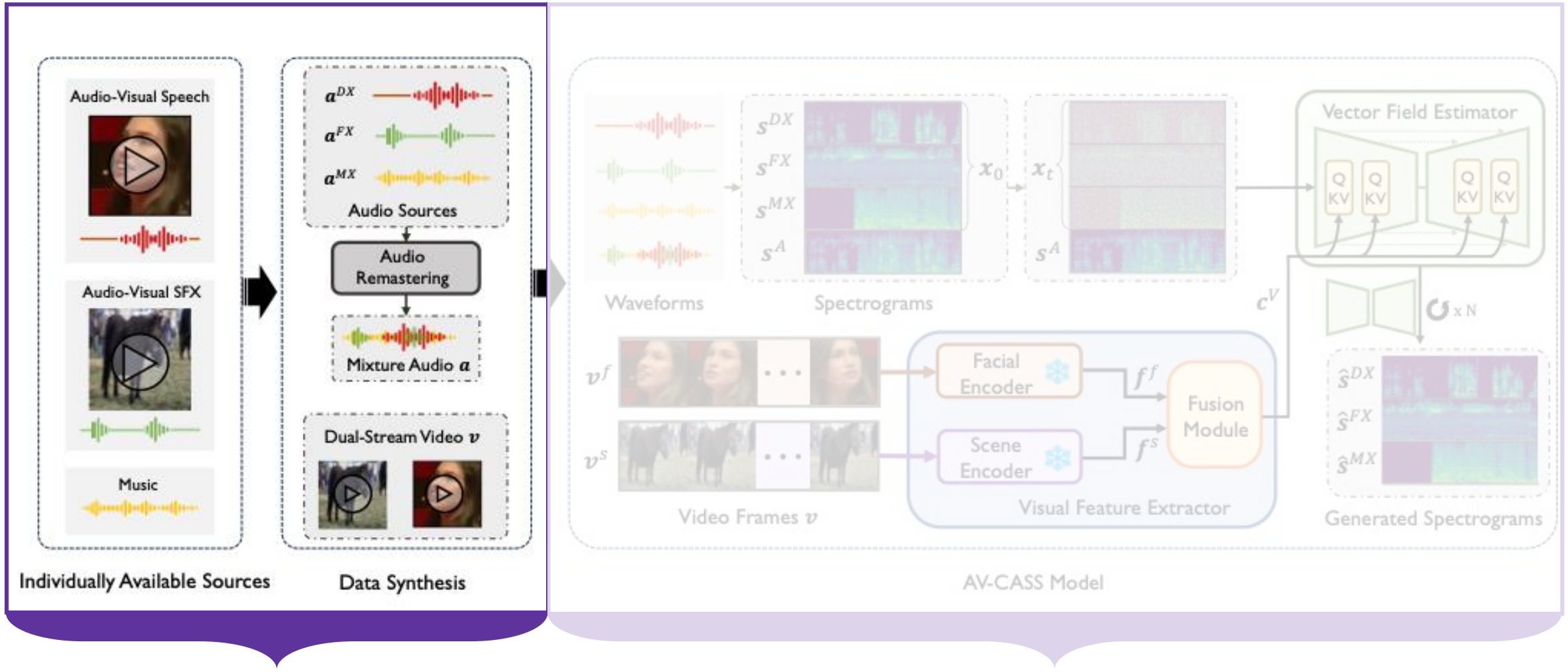
Dataset construction is needed

Main idea



Synthetic Data for Training **Training**

Synthetic Training Data



Synthetic Data for Training **Training**

Synthetic Training Data

① Source datasets

LRS3

speech + face video

VGGSound

event audio + scene video

FMA

music

②Synthetic source components

D_X

F_X

M_X

③ Mix

+ Mix

④ Synthetic mixture

Cinematic mixture

$D_X + F_X + M_X$

Learn correspondence

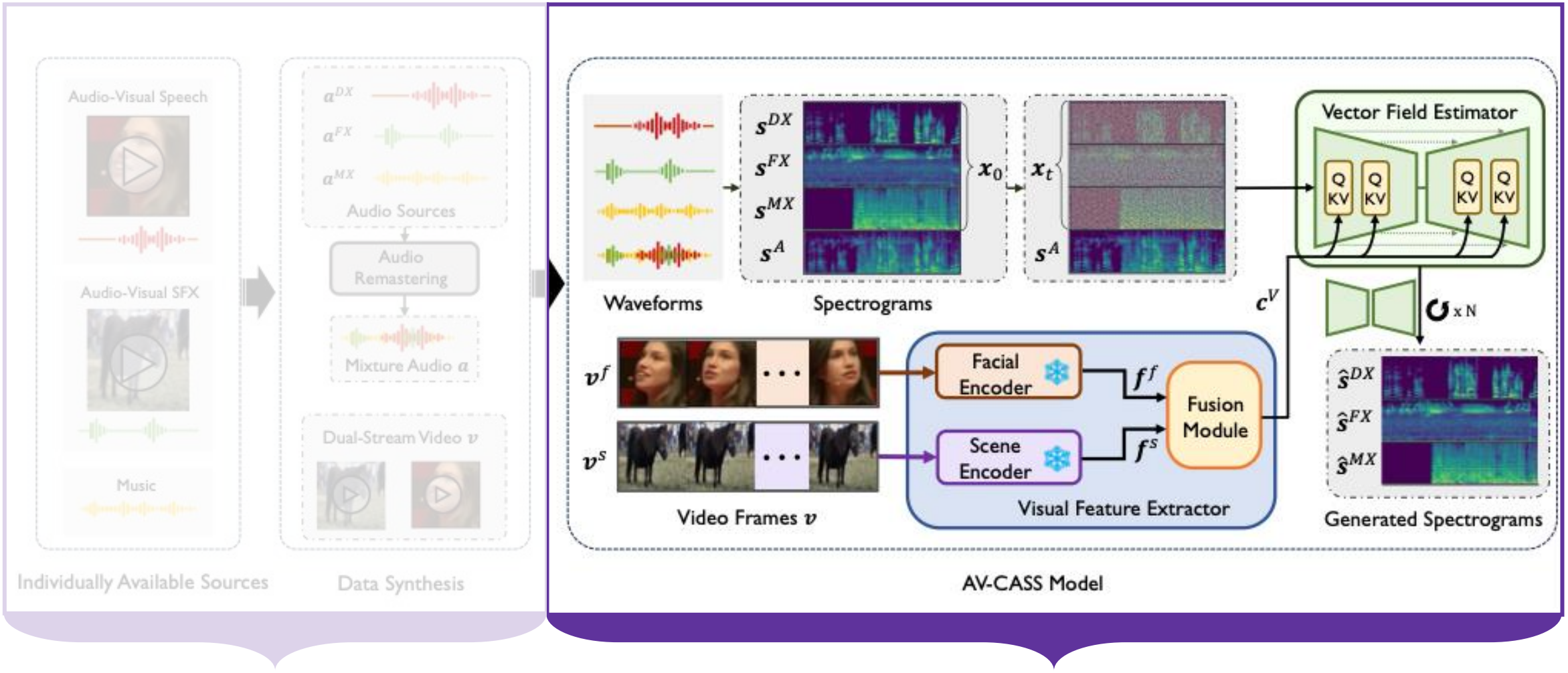
Dialogue (D_X)

Effects (F_X)

Music (M_X)

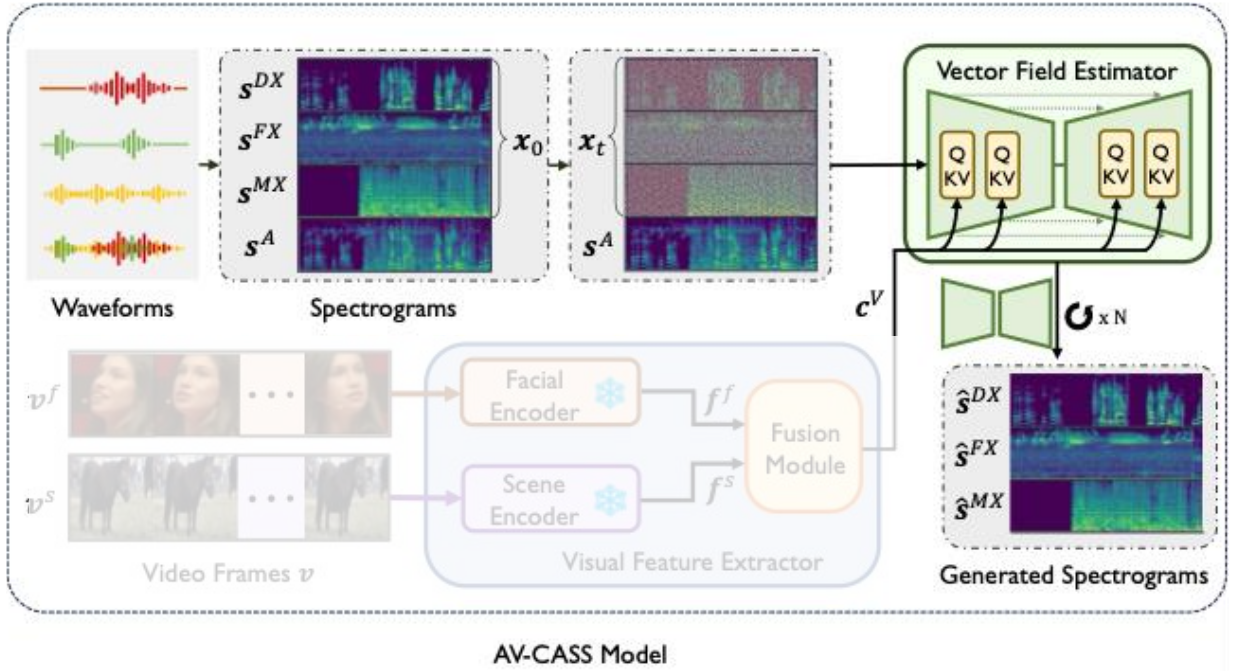
Synthetic training is not just data augmentation.
It teaches the model which visual cue corresponds to which source.

Training


Audio-Visual Speech
Audio-Visual SFX
Music
Individually Available Sources
Audio Sources
Audio Remastering
Mixture Audio a
Dual-Stream Video v
Data Synthesis
Waveforms
Spectrograms
Vector Field Estimator
Q
KV
x N
Facial Encoder
Scene Encoder
Fusion Module
Visual Feature Extractor
Video Frames v
Generated Spectrograms
AV-CASS Model
Synthetic Data for Training
Training

Training

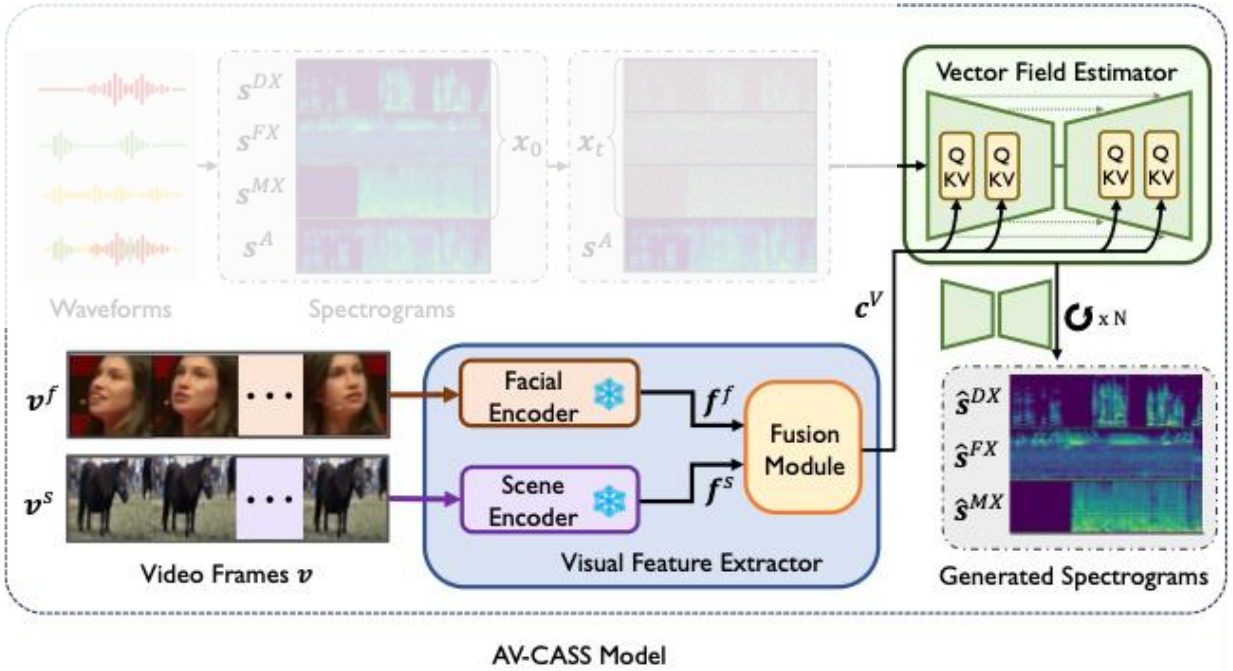
- **Input**: channel-wise stacked spectrograms
  + noise for source audios (for flow-matching training)
  - No noise added to mixture spectrogram s^A (serving as condition)



Training

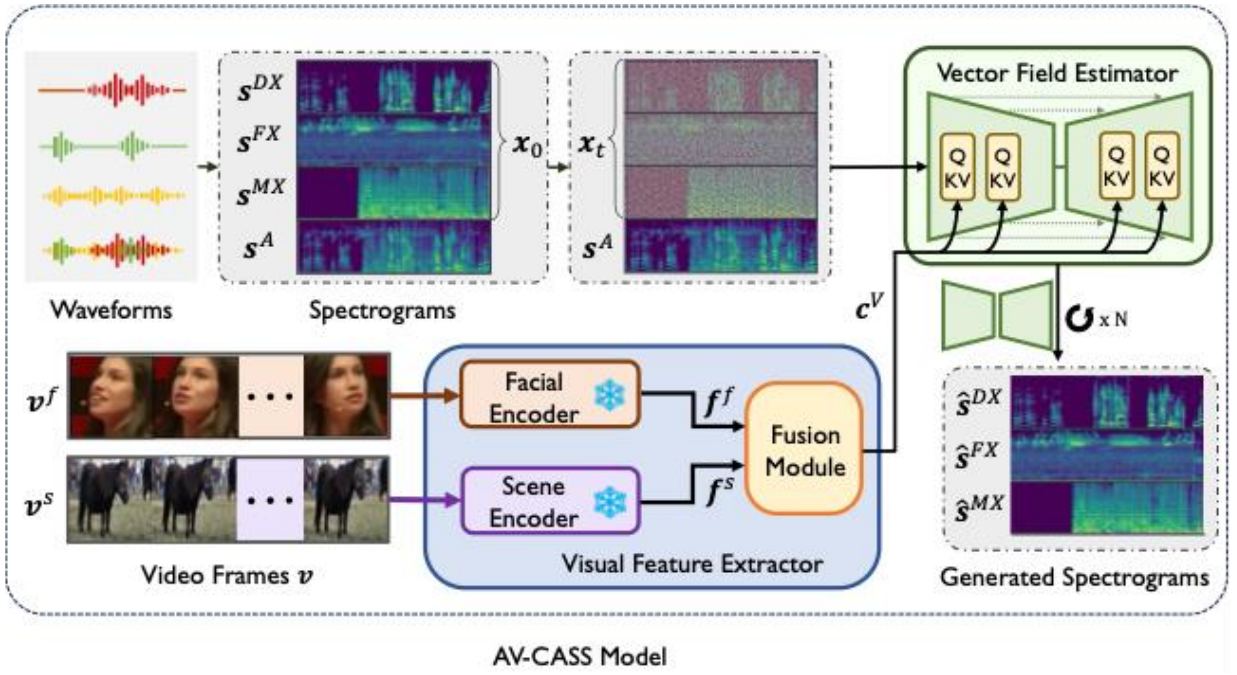
Condition:

- visual feature (c^V) from *Visual Feature Extractor*
  - Fused facial feature & scene feature
  - Facial encoder: pretrained encoder from AVSS model
  - Scene encoder: pretrained encoder from V2A model
  - Conditions the audio backbone via cross attention

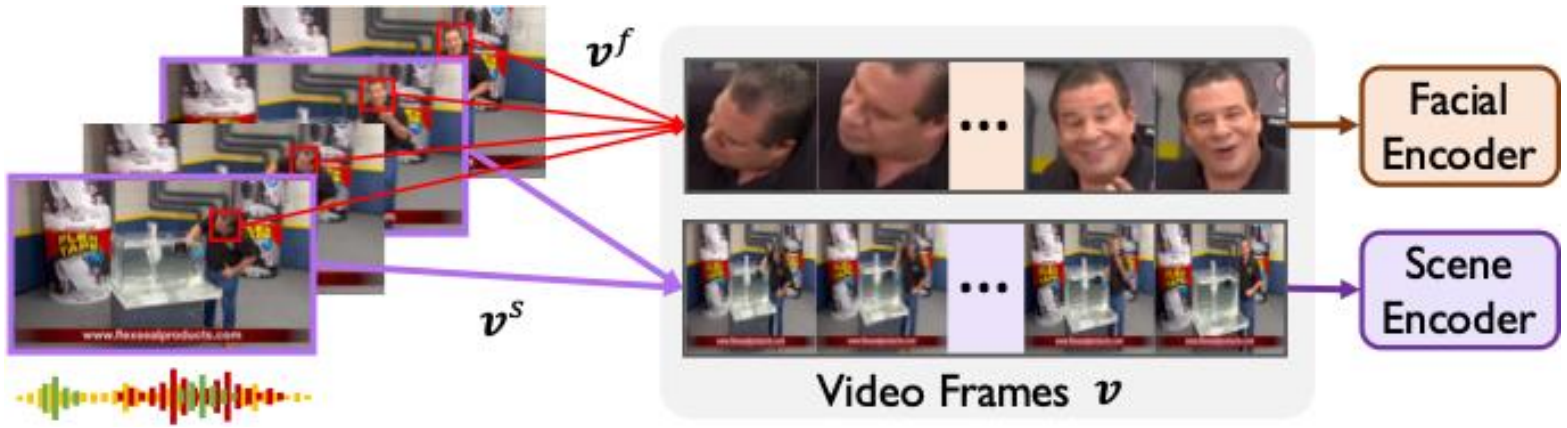


Inference

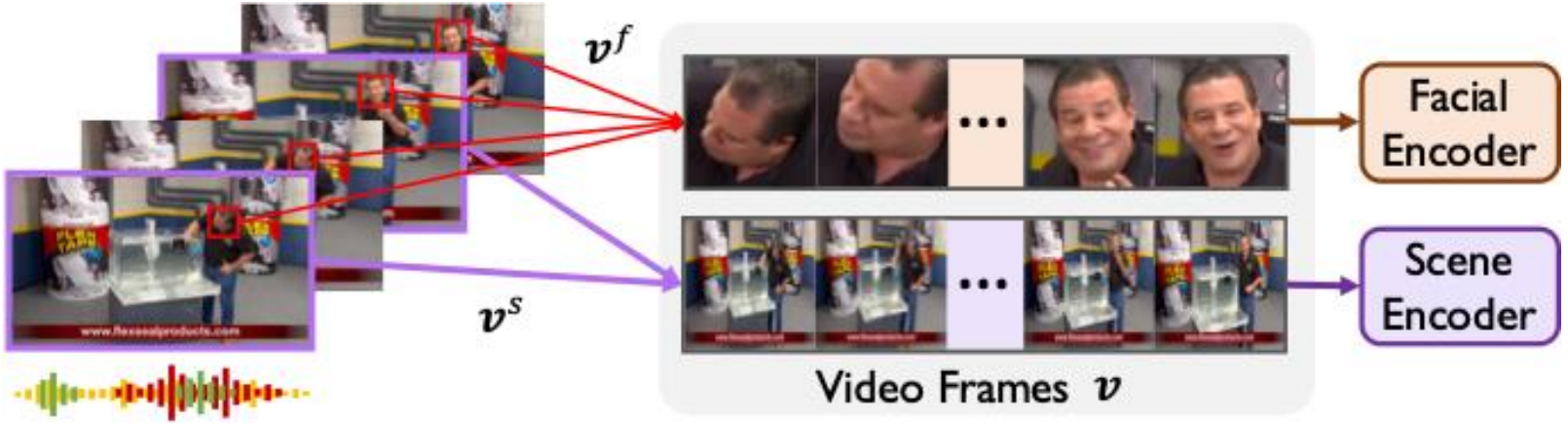
- **Inference input:**
  3 Gaussian noise & mixture spectrogram
- **Output**:
  channel-wise concatenated spectrograms of separated audios



- For real world samples:



Results on real world data



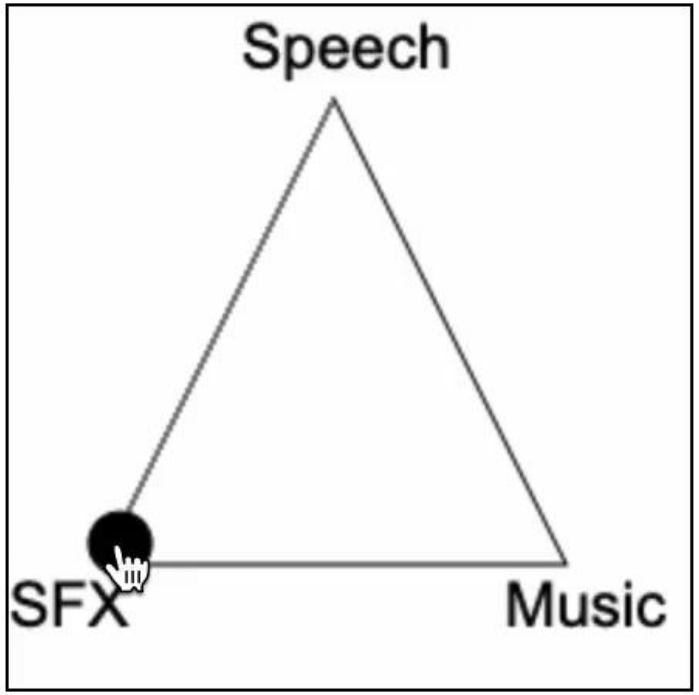
More Demo:

Interactive separation results & real-world separation results

https://cass-flowmatching.github.io

Interactive CASS

This video presents the remixed result of separated tracks from AV-CASS.



Click on the triangle to start the video and audio.
Move the cursor inside the triangle to mix audio stems.

Results on real world data

Inception (2010)

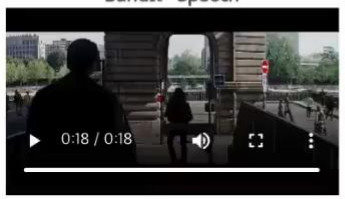
Input Video

Ours - Speech

Ours - SFX

Ours - Music

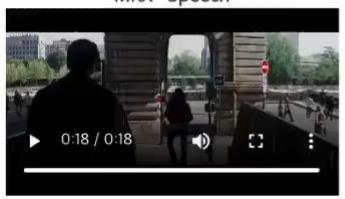
BandIt - Speech



BandIt - SFX

BandIt - Music

MRX - Speech

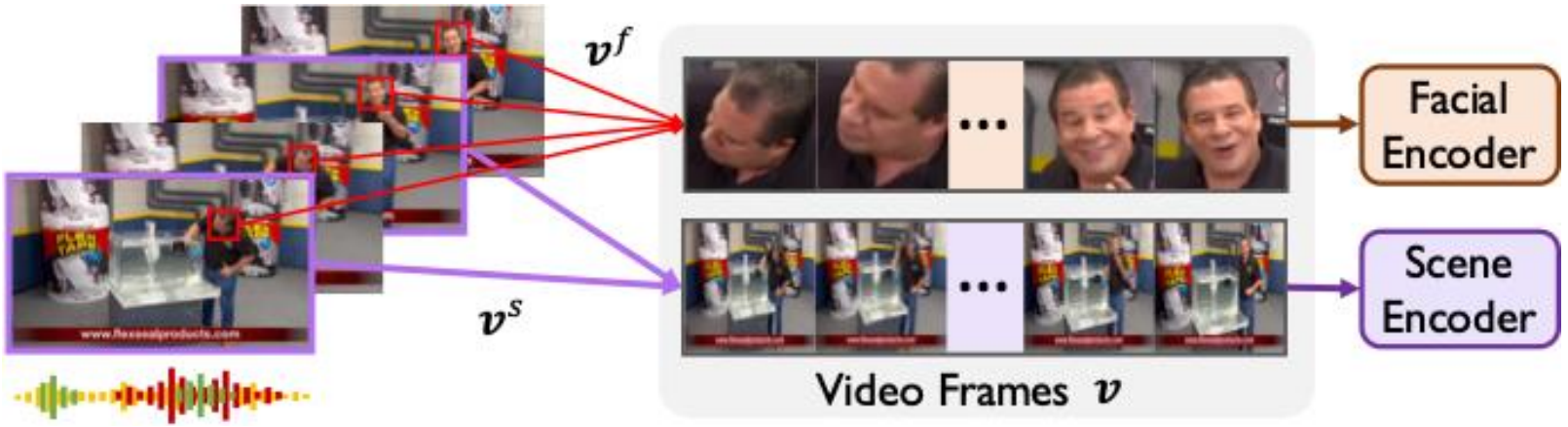


MRX - SFX

MRX - Music

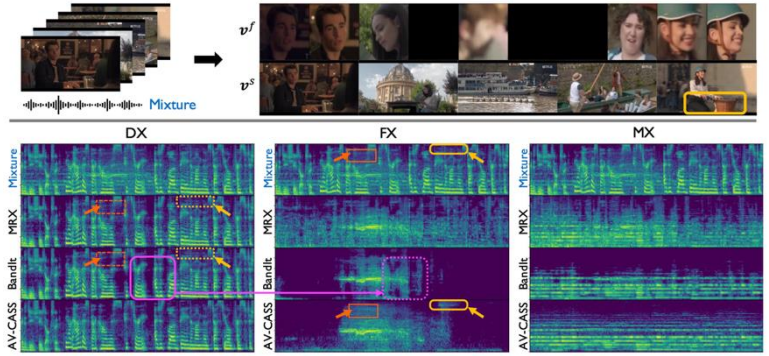
Due to time constraints, comparison audio samples from other methods are available on our project page.

Results on real world data



Method	MRX [48]	BandIt [59]	**AV-CASS (Ours)**
MOS (↑)	2.55 ± 0.10	3.78 ± 0.10	$\mathbf{4.13 \pm 0.09}$

Table 1. **MOS results of CASS models on real-world samples.** The scores are computed based on 95% confidence intervals.



Results on synthetic data

- **Tab. 1. & 2.** AV-CASS achieves the best MOS, FAD, KL, PESQ, and WPR among the compared methods
- **Tab. 3.** - Using both visual streams yields the best performance

Method	A-V	FAD (↓)	KL (↓)	SI-SDRi (↑)	PESQ (↑)	WPR (↓)
Predictive Model						
Hybrid Demucs [14]	✗	2.05	1.03	13.57	2.16	5.24
HT Demucs [52]	✗	2.08	1.06	13.41	2.06	9.23
MRX [48]	✗	3.47	1.67	10.60	1.89	14.91
BandIt [59]	✗	2.15	1.14	**14.40**	2.15	4.65
Generative Model						
MSDM [44]	✗	2.90	2.90	11.63	2.12	5.65
DAVIS-Flow [29]	✓	5.94	1.64	9.25	1.96	12.14
AV-CASS (Ours)	✓	**0.84**	**0.93**	12.32	**2.26**	**1.84**
Ours (Audio-only)	✗	1.63	1.15	12.23	2.08	2.01

Tab. 1. Results on AVDnR (synthesized) testset

Method	MRX [48]	BandIt [59]	AV-CASS (Ours)
MOS (↑)	2.55 ± 0.10	3.78 ± 0.10	**4.13** $\pm$ **0.09**

Tab. 2. MOS result on real-world movie samples

Method	v^f	v^s	FAD (↓)	KL (↓)	SI-SDRi (↑)	PESQ (↑)
Audio-only	✗	✗	1.63	1.15	12.23	2.08
+ Facial stream	✓	✗	0.91	1.00	12.13	2.21
+ Scene stream	✗	✓	0.87	1.00	12.27	2.24
+ Both (Ours)	✓	✓	**0.84**	**0.93**	**12.32**	**2.26**

Tab. 3. Ablation on visual streams