SAM-based Audio-Visual Segmentation

with Spatio-Temporal, Bidirectional Audio-Visual Attention

Ju-hyeong Seon (선 주 형)

Advisor: Prof. Sung-Eui Yoon

School of Computing, KAIST

SGVR Lab
KAIST

Contents

Background

Segment Anything Model (SAM)

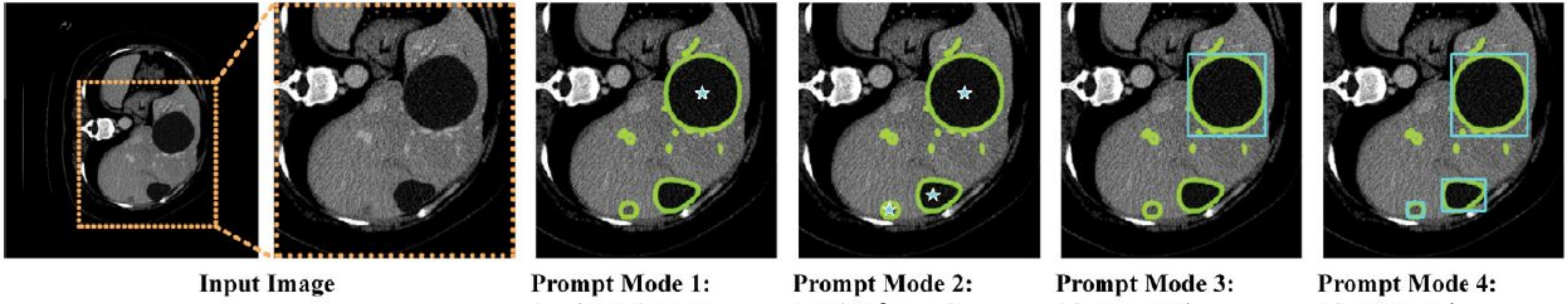
Segment Anything Model (SAM)

- A foundation model for image segmentation
- Strongly impacts on various dense prediction problems

Segment anything, ICCV 2023

Applications of SAM

- Applied in various dense prediction problems
- Medical Image Segmentation, Shadow Detection, 3D Segmentation, etc.

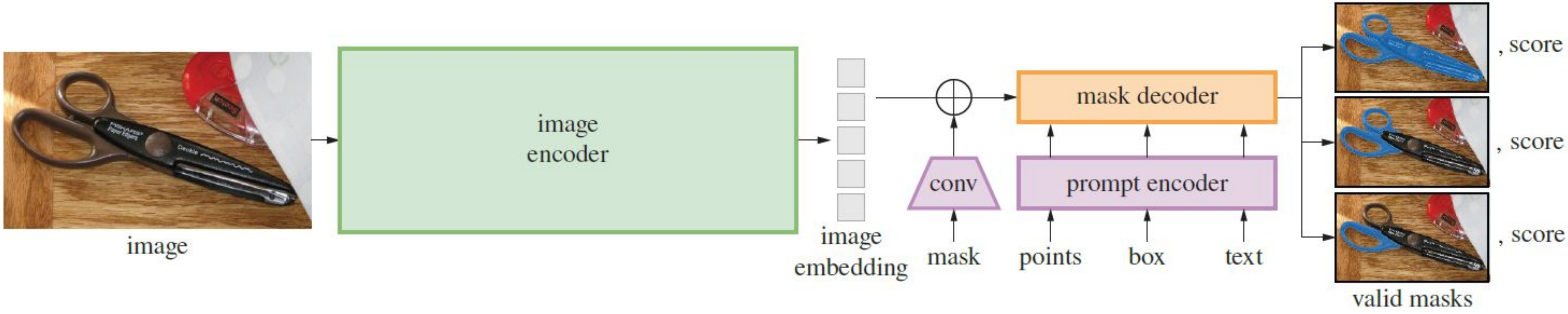


SGVR Lab KAIST

Segment anything model for medical image analysis: An experimental study, Medical Image Analysis 89 (2023)

Architecture of SAM

- Large image encoder (ViT-H) is important to generalization performance
- Mask decoder and prompt encoder works for promptable segmentation



SGVR Lab KAIST

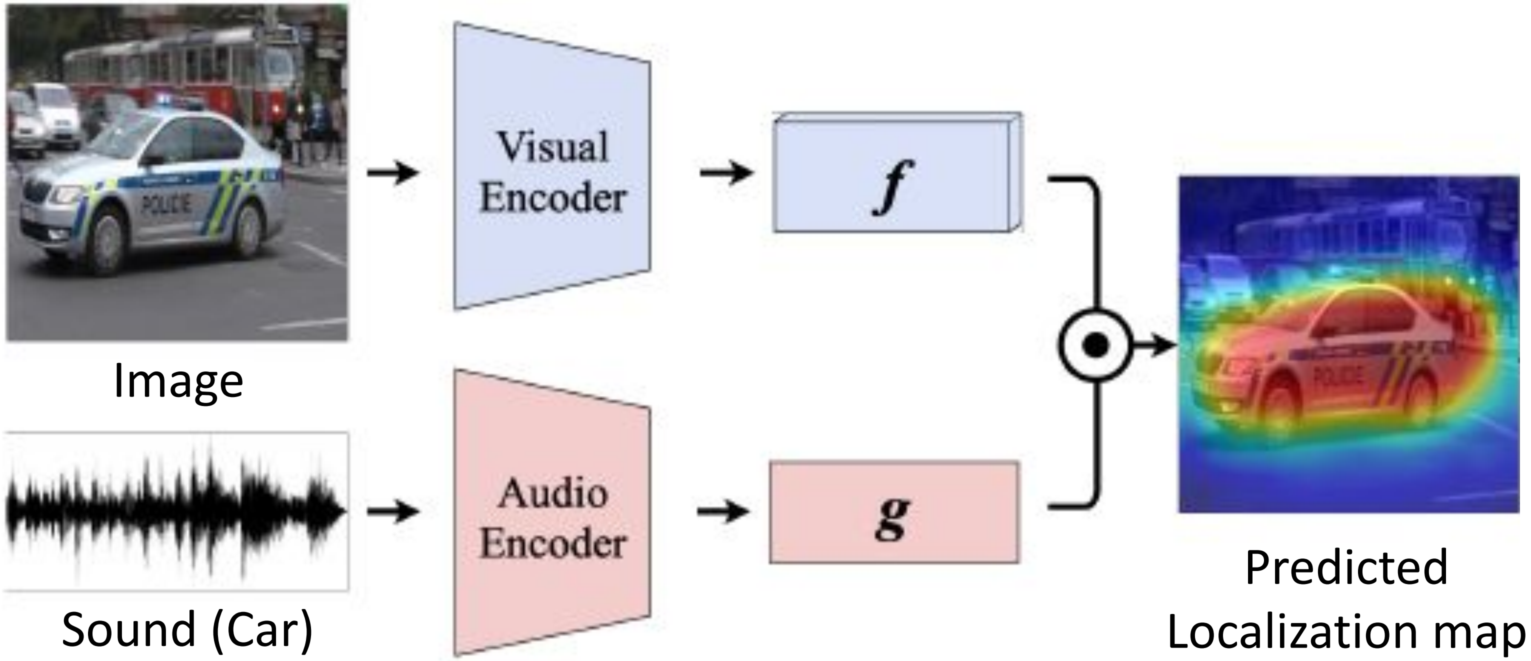
Background

Audio-Visual Segmentation

SGVR Lab KAIST

Sound Source Localization (SSL)

- A research field in audio-visual learning, using audio-visual correspondence
- Find the location of sound source on the image frame



Localizing Visual Sounds the Hardway, CVPR 2021

SGVR Lab KAIST

Audio-Visual Segmentation (AVS)

- Advanced task of sound source localization
- **Segment the sounding objects** in the sequence of frames

Audio-visual segmentation, ECCV 2022

SGVR Lab
KAIST

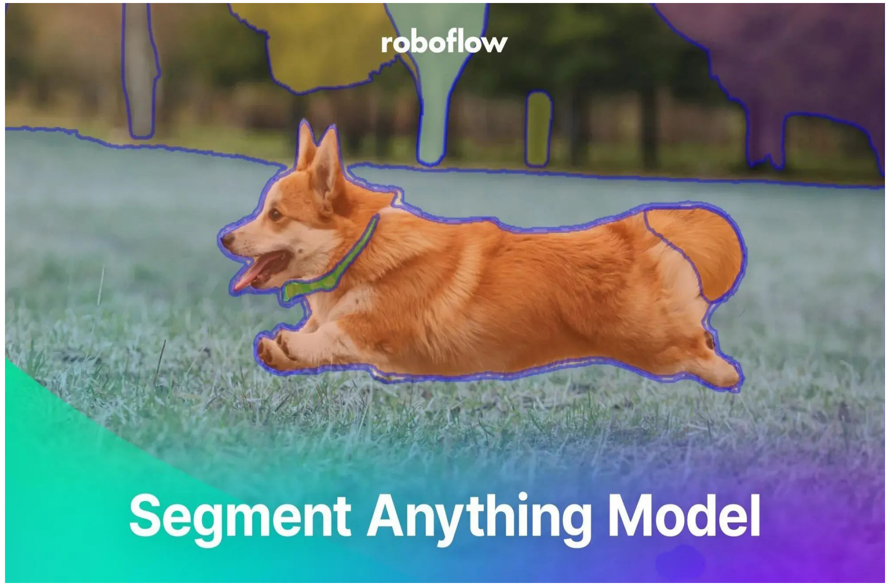
Introduction

SGVR Lab
KAIST

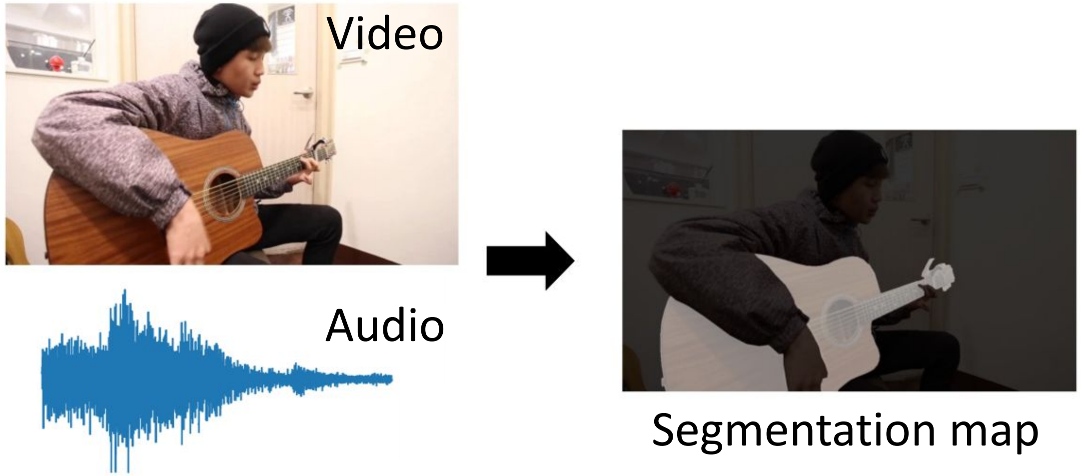
Research Goal

SAM for AVS

Segment Anything Model (SAM)

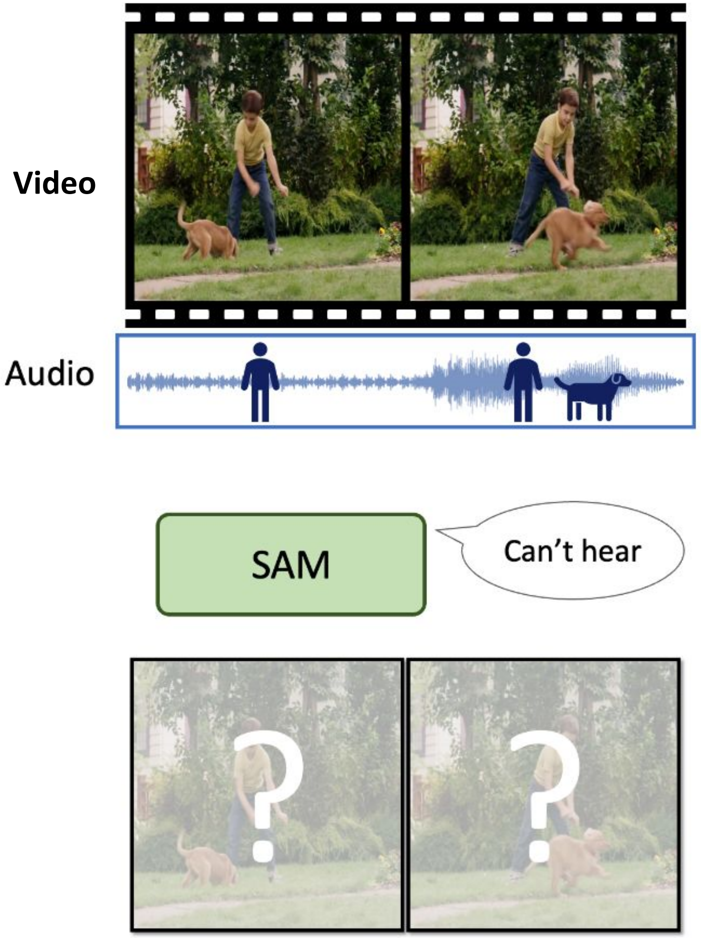


Audio-Visual Segmentation (AVS)



Images from https://blog.roboflow.com/how-to-use-segment-anything-model-sam/, AVSegformer, AAAI 2024

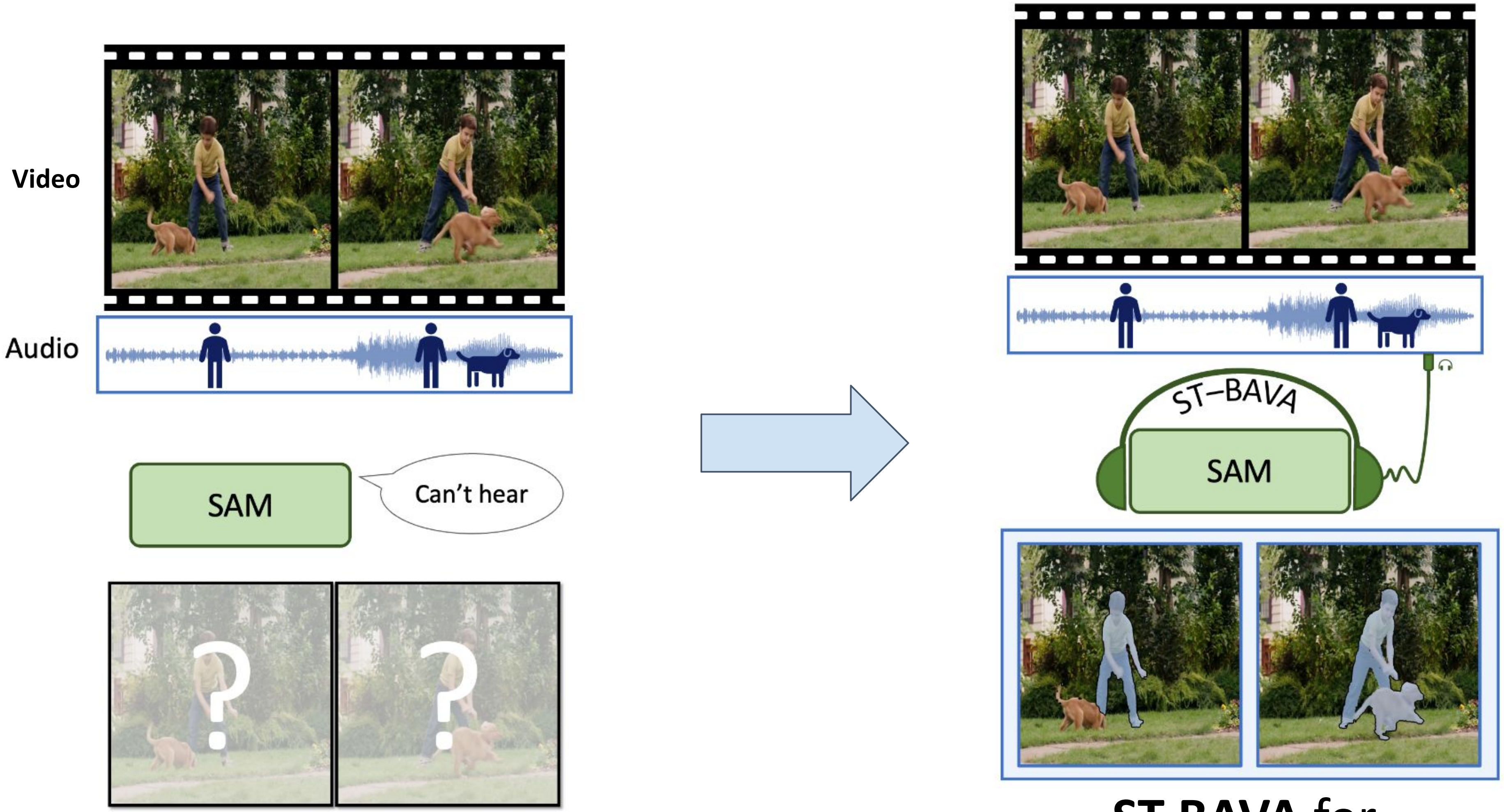
SAM for AVS



- **Original SAM**
  - Can't process audio and video inputs
  - Can't solve AVS

Research Goal

SAM for AVS



ST-BAVA for
Video and Audio

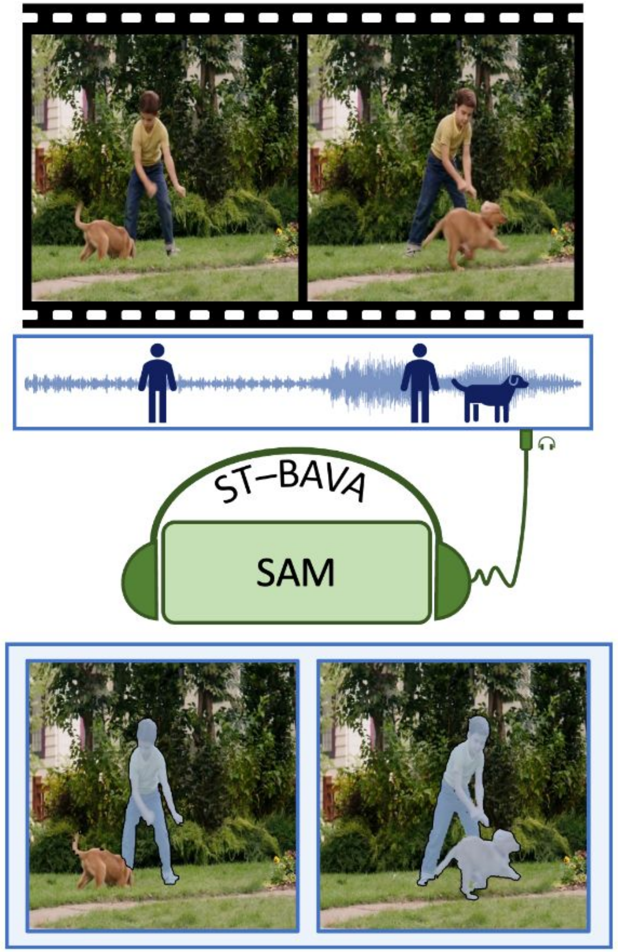
SGVR Lab
KAIST

Research Goal

SAM for AVS

- **ST-BAVA extends SAM into auditory and temporal dims**

- **ST-BAVA**
  - **S**patio-**T**emporal, **B**idirectional **A**udio-**V**isual **A**ttention
  - Exploits the **spatio-temporal and audio-visual** relationship via **cross-attention**



ST-BAVA for
Video and Audio

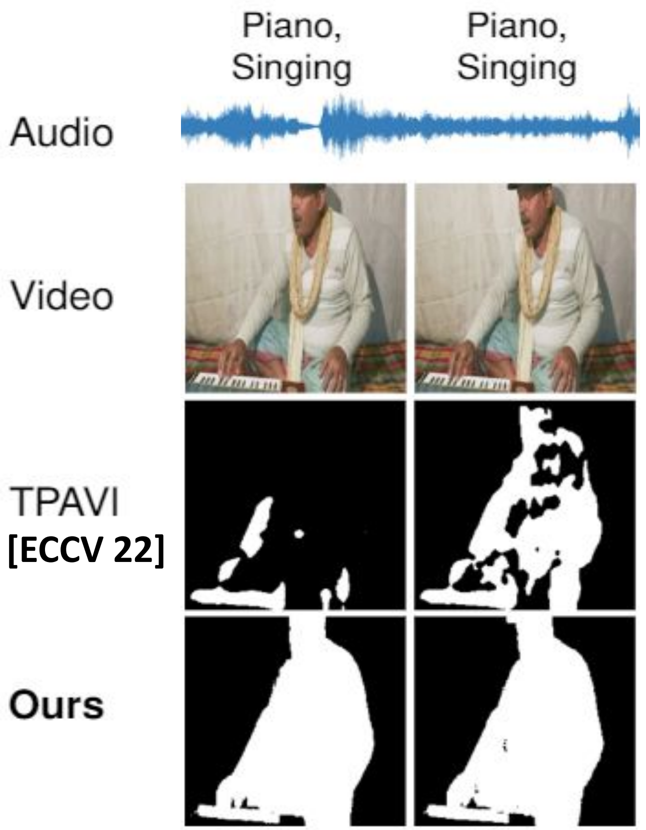
SGVR Lab KAIST

SAM for AVS

- Quantitative comparison with AVS methods on the AVS benchmark
- SAM shows **12.9%** mIoU improvement compared to SOTA model

	Methods	mIoU	F-score
	TPAVI [ECCV 22]	54.0	0.65
Prev. SOTA ➡	AQFormer [IJCAI 23]	61.1	0.72
Ours ➡	SAM + ST-BAVA (Ours)	**69.0**	**0.78**

Results on AVS Benchmark



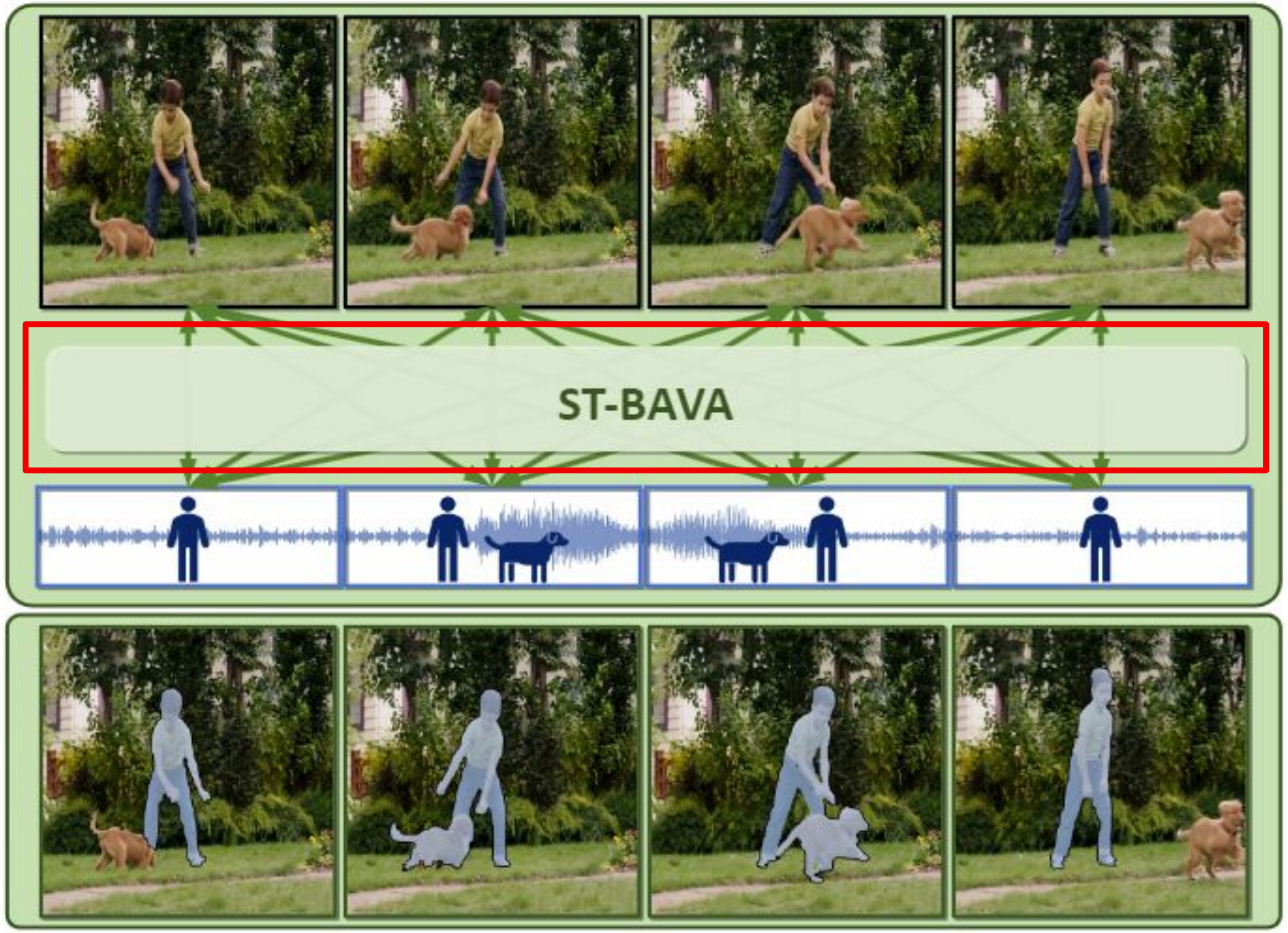
Methods

Problem definition (AVS)

- Input: ***T*** seconds video $\boldsymbol{V_i}$ into ***T*** images & audio streams
- Output: ***T*** binary masks $\in \{0,1\}^{H_i \times W_i}$ representing that the pixel sounds or not

Overview

- Enable SAM to handle the **consecutive video frames with corresponding audio**
- We insert audio-visual feature interaction module: **ST-BAVA**

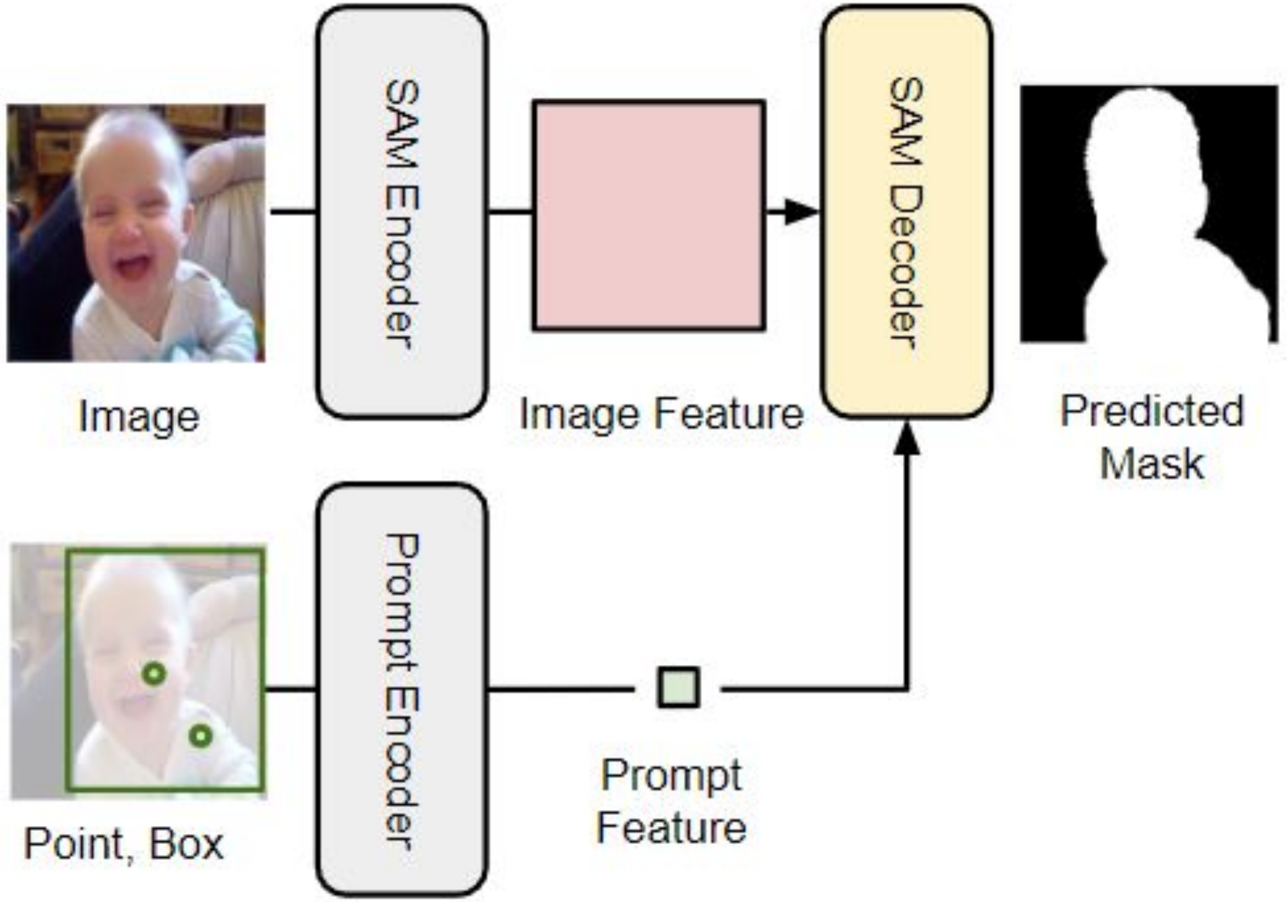


SAM with ST-BAVA (Ours)

SGVR Lab KAIST

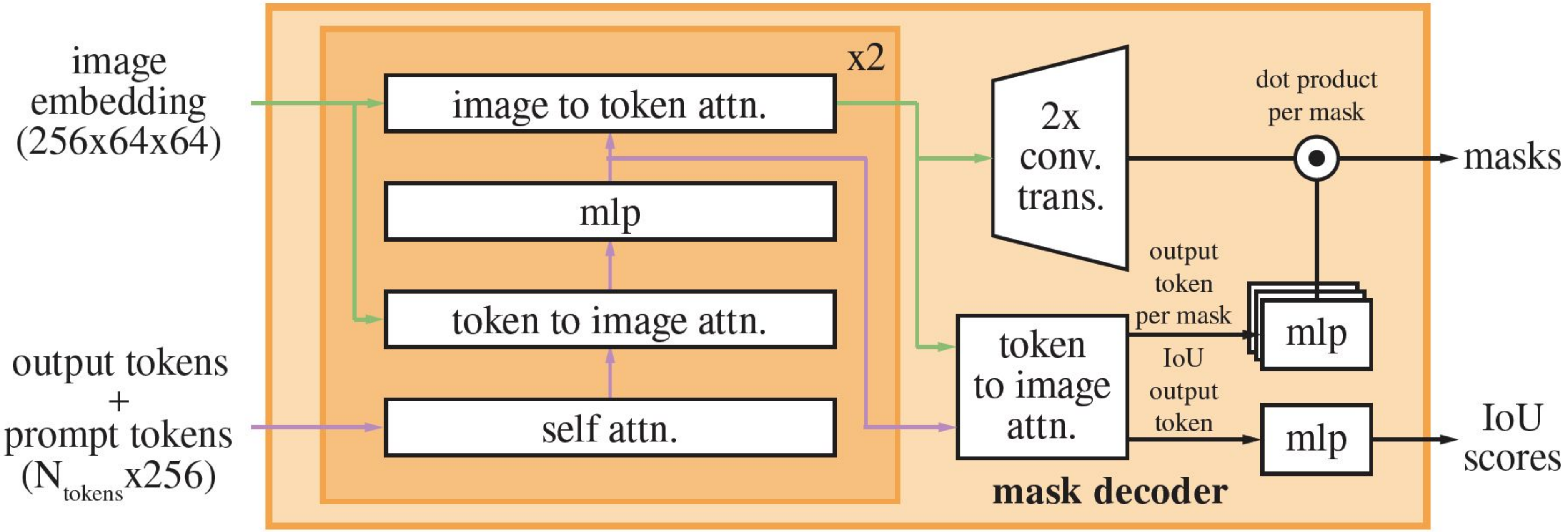
SAM pipeline

- In SAM, prompts guide where to segment in the mask decoder



SAM Decoder

- In SAM, prompts guide where to segment in the mask decoder

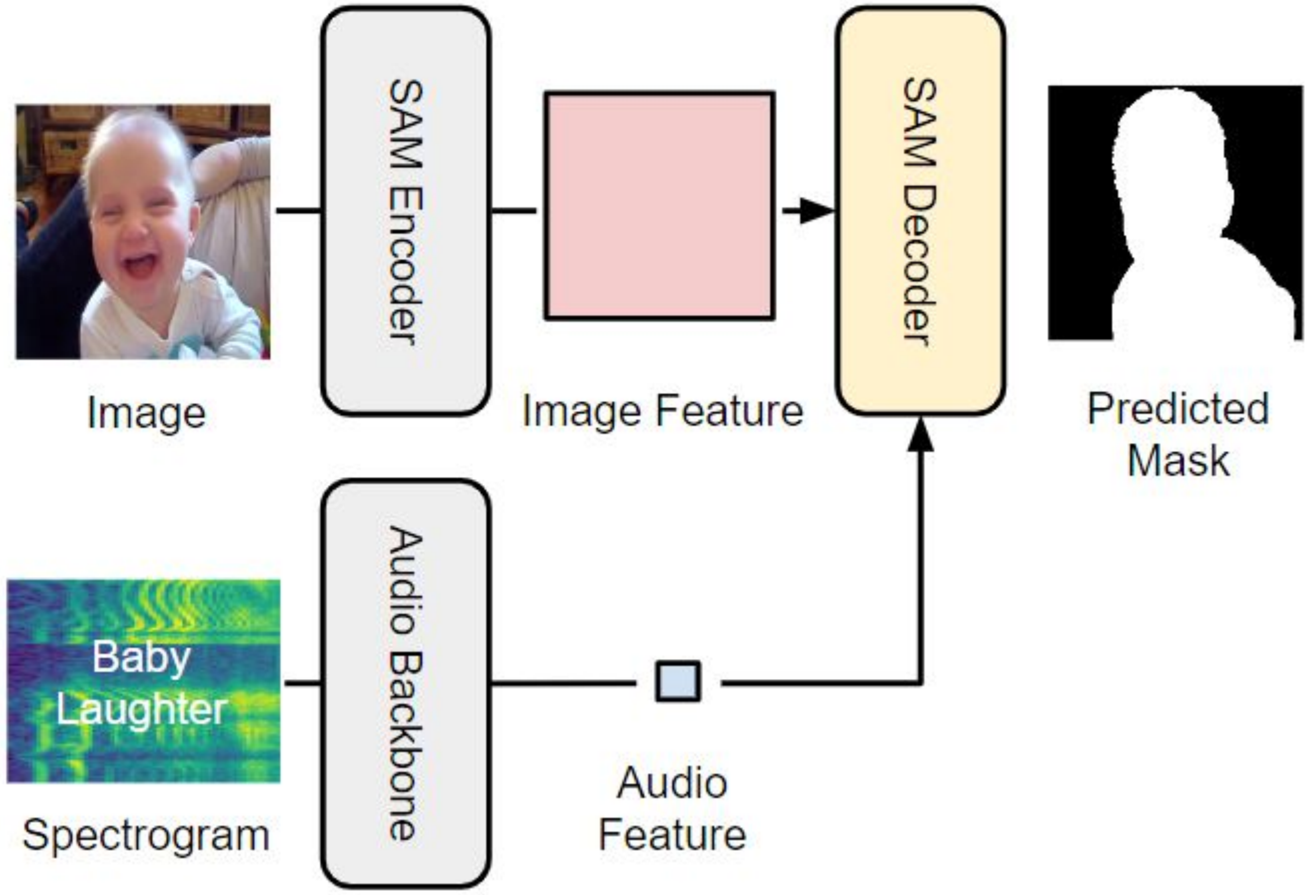


SAM decoder architecture

SGVR Lab KAIST

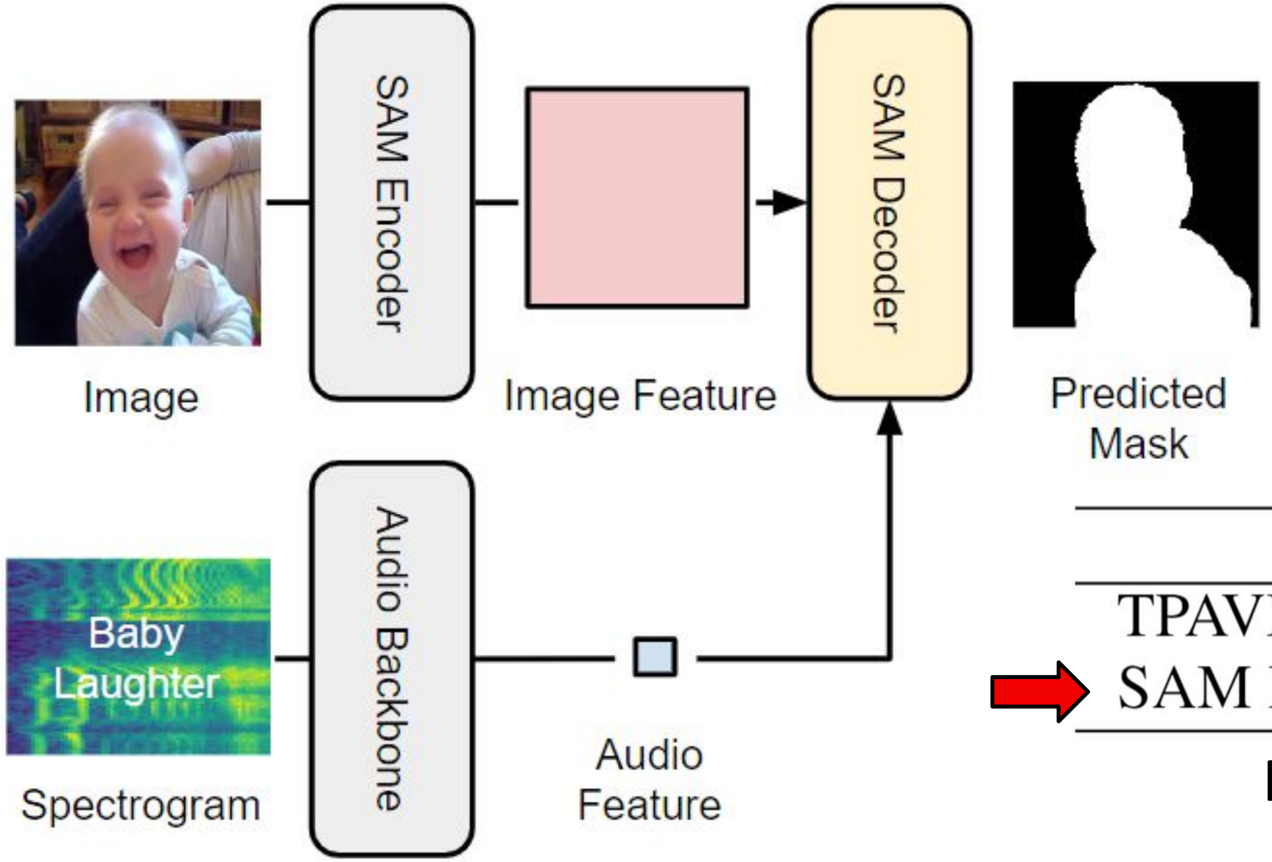
Naive Approach (SAM Baseline)

- We can replace point, box prompts to audio



Naive Approach (SAM Baseline): Limitations

1. **SAM Decoder is too shallow** to learn the audio-visual correspondence
2. Doesn’t utilize **the temporal relationship across the multiple frames**



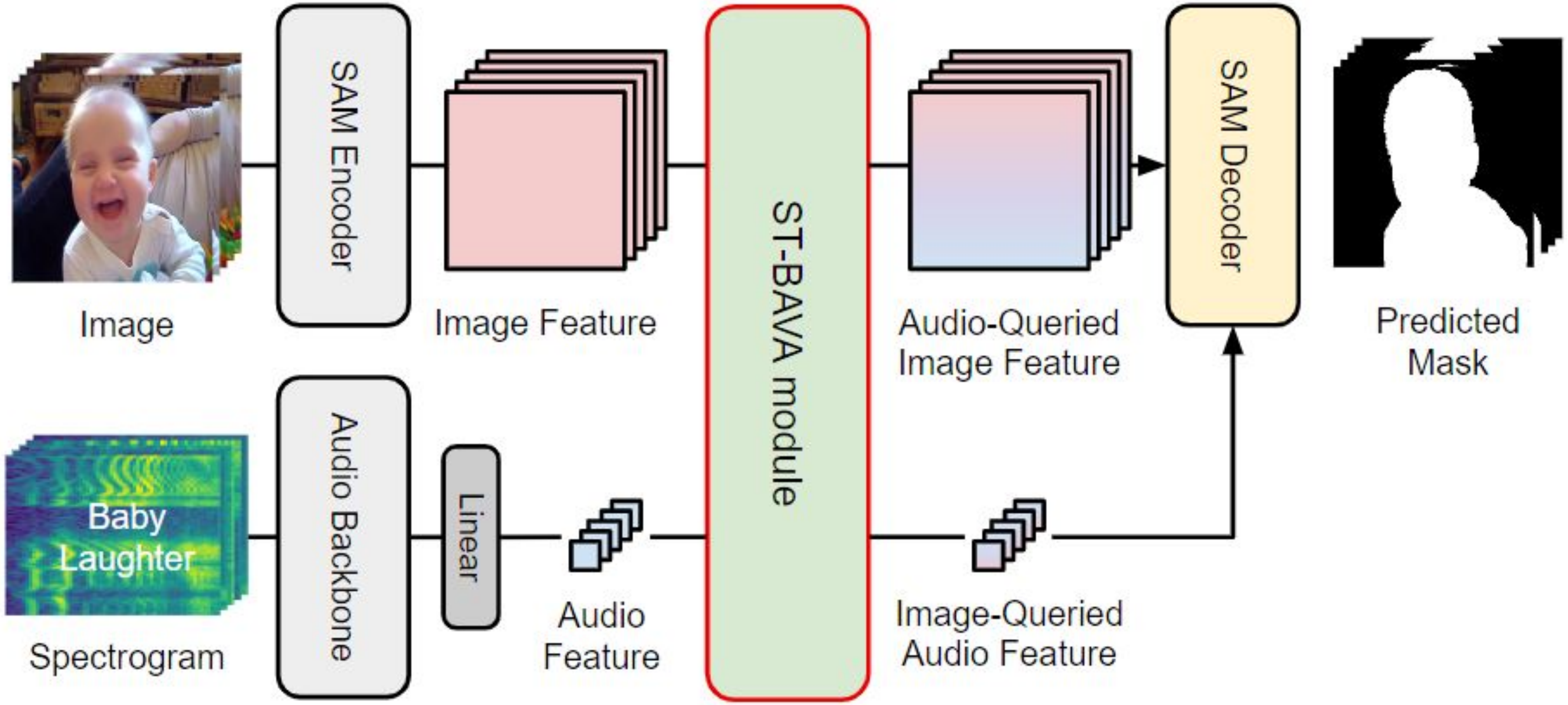
	mIoU	F-score
TPAVI [ECCV 22]	54.0	0.65
SAM Baseline	52.3	0.66

Results on AVS Benchmark

SGVR KAIST

Our approach

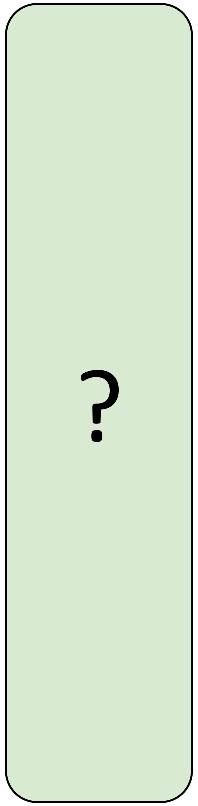
SAM + ST-BAVA



- Insert **ST-BAVA** module between the encoder and decoder

Our approach

SAM + ST-BAVA



- How to design **ST-BAVA** module ?

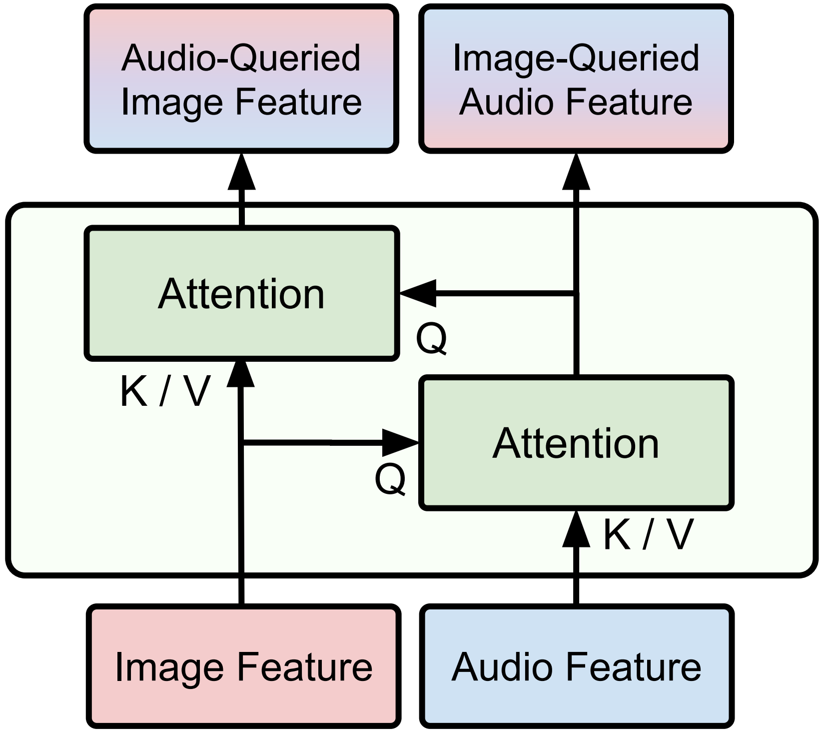
SGVR Lab KAIST

ST-BAVA | Architecture

- **Auditory extension** of SAM

  => **Bidirectional** attention

  between audio-visual features



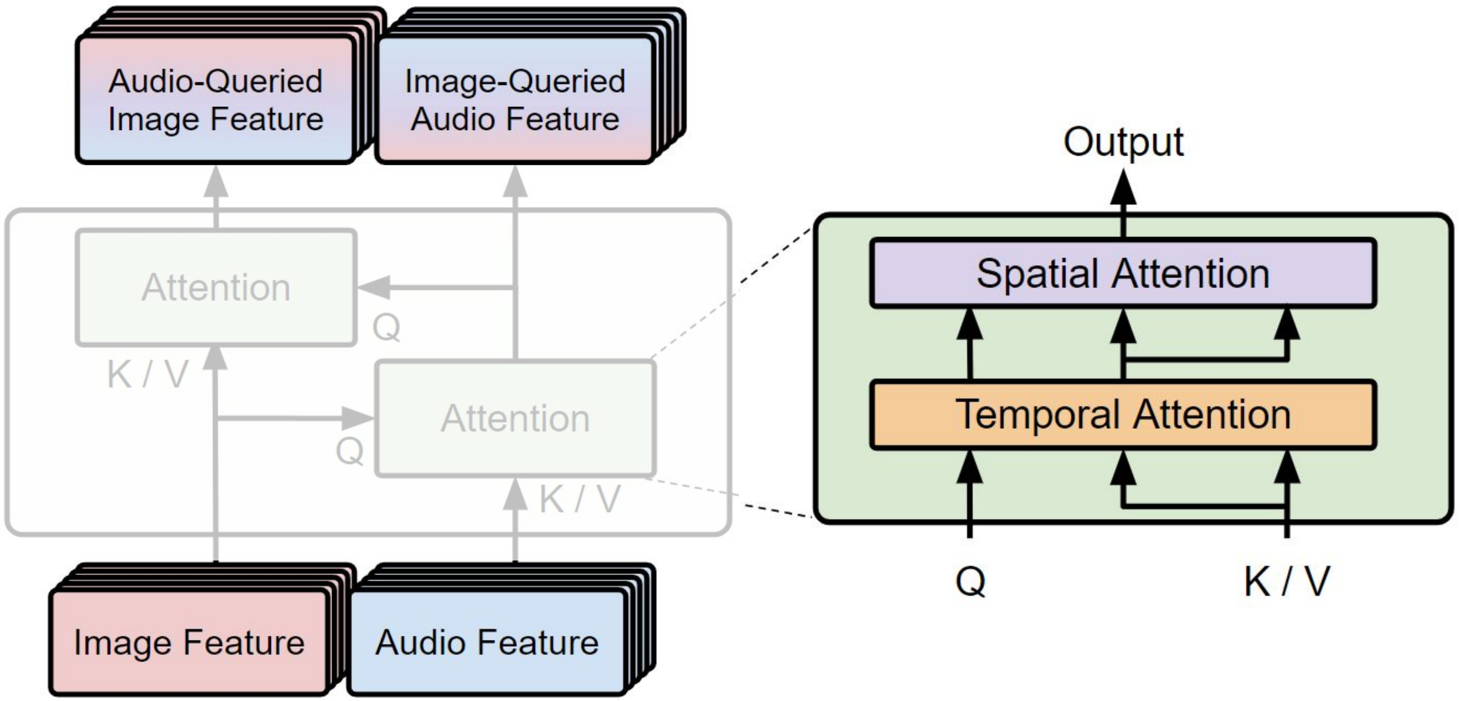
SGVR Lab KAIST

ST-BAVA | Architecture

- **Temporal extension** of SAM

  => Spatio-Temporal attention*

  between audio-visual features

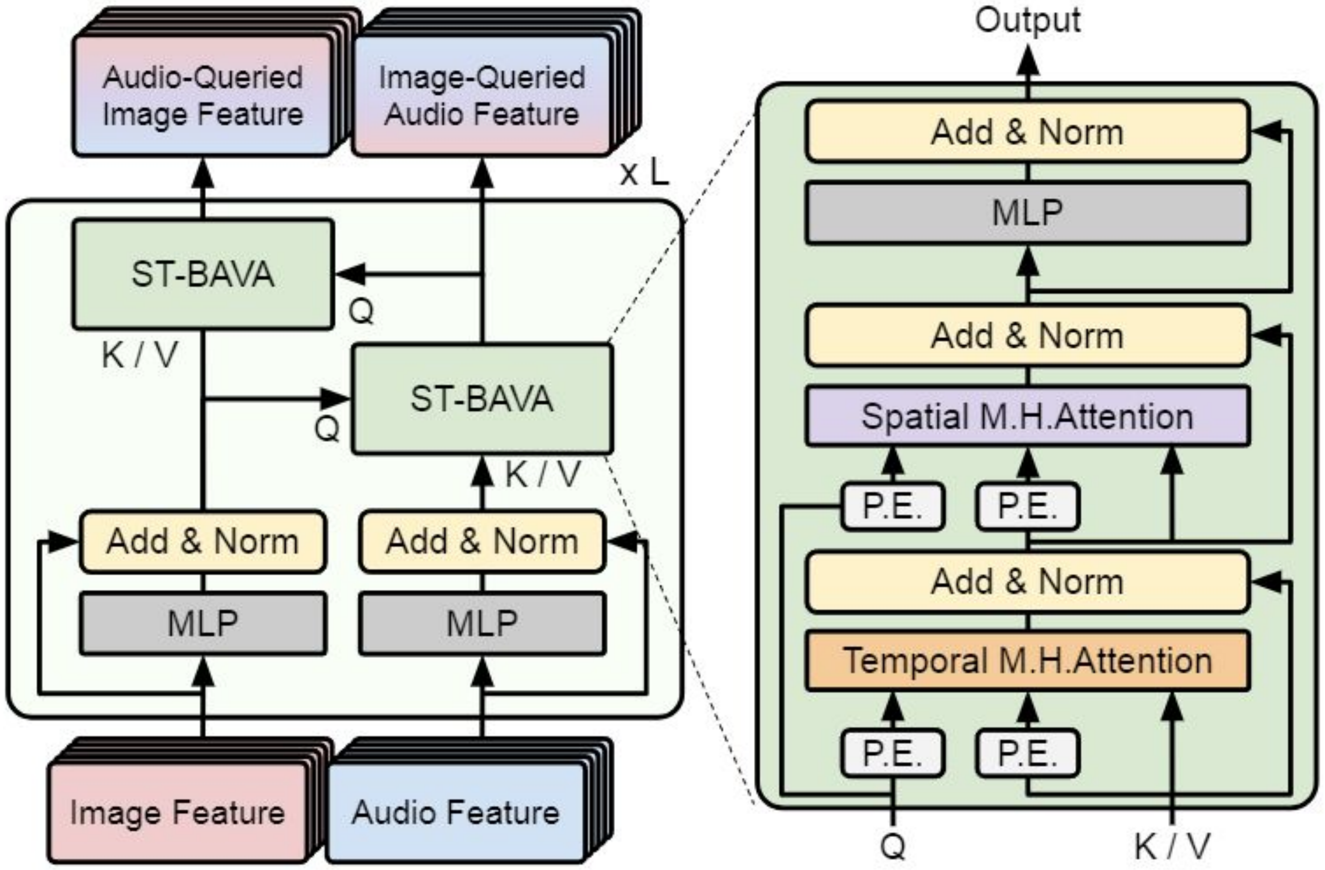


SGVR Lab KAIST

*Divided spatio-temporal attention reduces the memory requirements (details in appendix)

ST-BAVA | Architecture

- Spatio-Temporal , Bidirectional Audio-Visual Attention



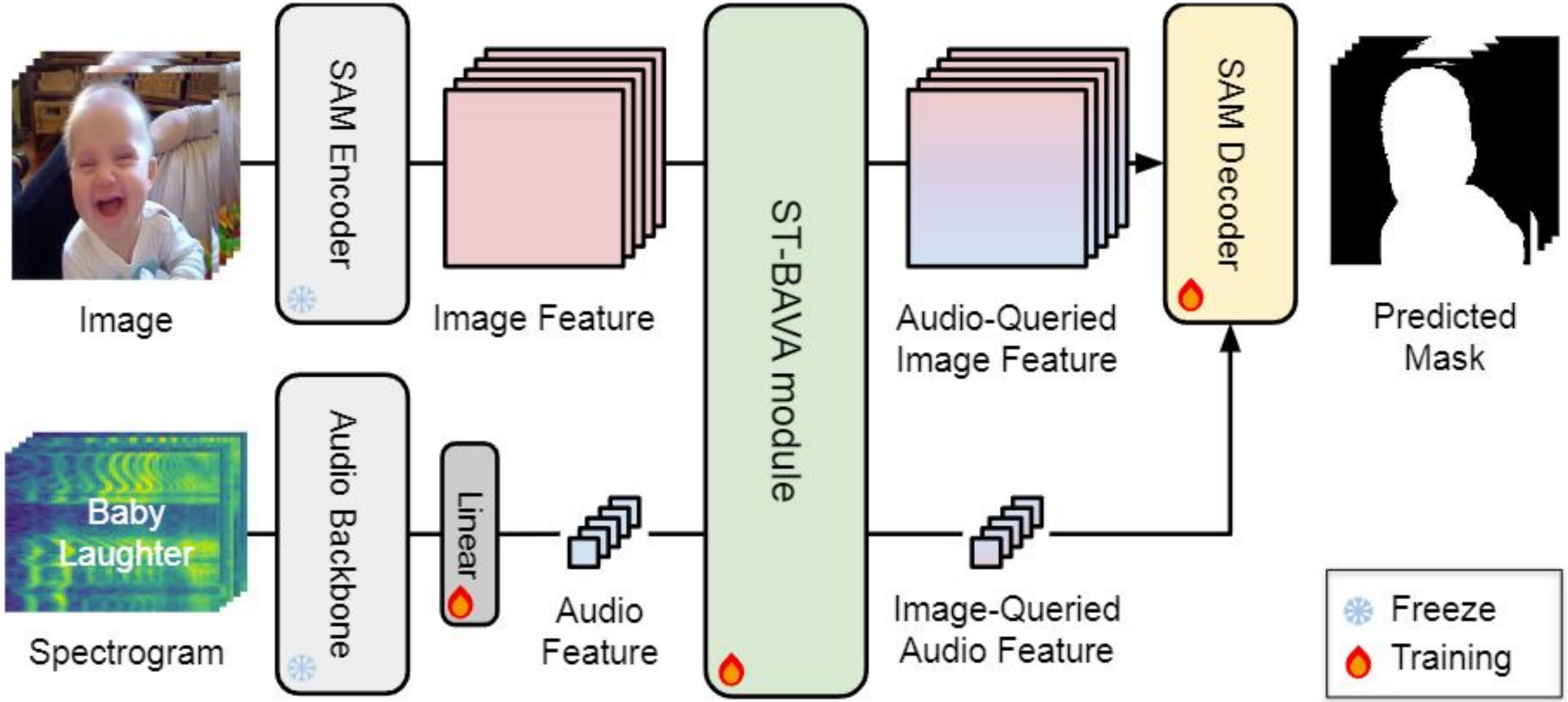
SGVR Lab KAIST

Our approach

SAM + ST-BAVA

	mIoU	F-score
TPAVI [ECCV 22]	54.0	0.65
SAM Baseline	52.3	0.66
SAM + ST-BAVA	**69.0**	**0.78**

Results on AVS Benchmark



- Shows meaningful performance improvement

SGVR Lab KAIST

Feature similarity analysis

Input video:
Male Speech

Before
ST-BAVA

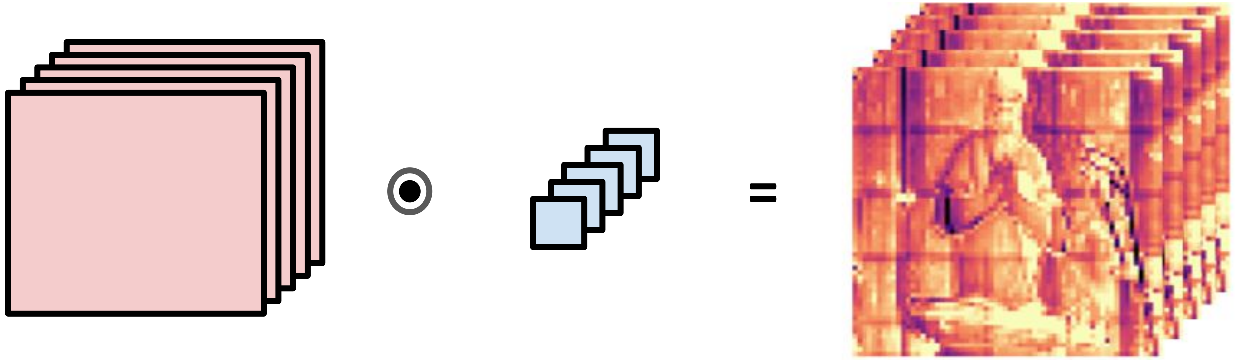


Image Feature ⊙ Audio Feature = Similarity map

Irregular patterns

After
ST-BAVA

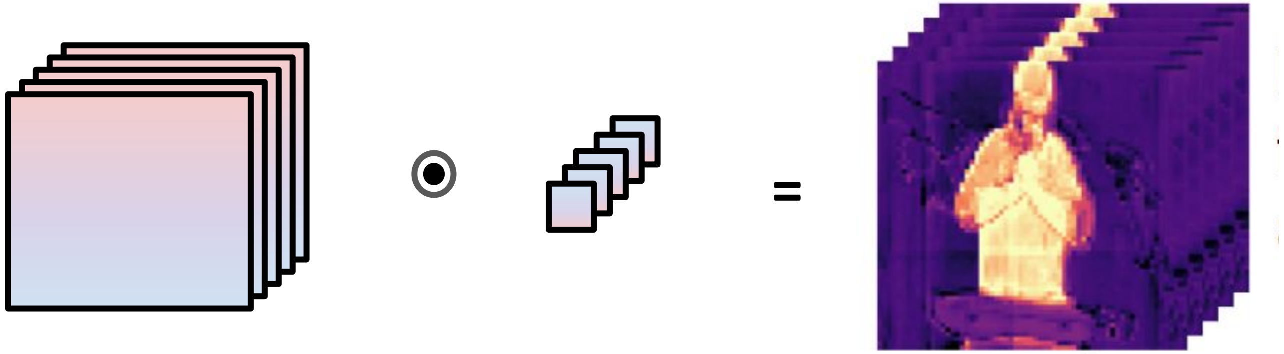


Image Feature ⊙ Audio Feature = Similarity map

Correct separation of the sound source

SGVR Lab KAIST

Experimental results

SGVR Lab
KAIST

Dataset - AVSBench

- 5 second per video with 1 FPS
- Two subsets
  - **Single sound source** subset
  - **Multiple sound sources** subset

subset	classes	videos	train/valid/test	annotated frames
Single-source	23	4,932	3,452* /740/740	10,852
Multi-sources	23	424	296/64/64	2,120

SGVR Lab KAIST

Audio-visual segmentation, ECCV 2022

Evaluation Metric

- **Accuracy between the ground truth mask and model's prediction**
  - mIoU, F-score (details in Appendix)
- Training loss: Binary Cross Entropy with GT and prediction mask

Ground truth

Model Prediction

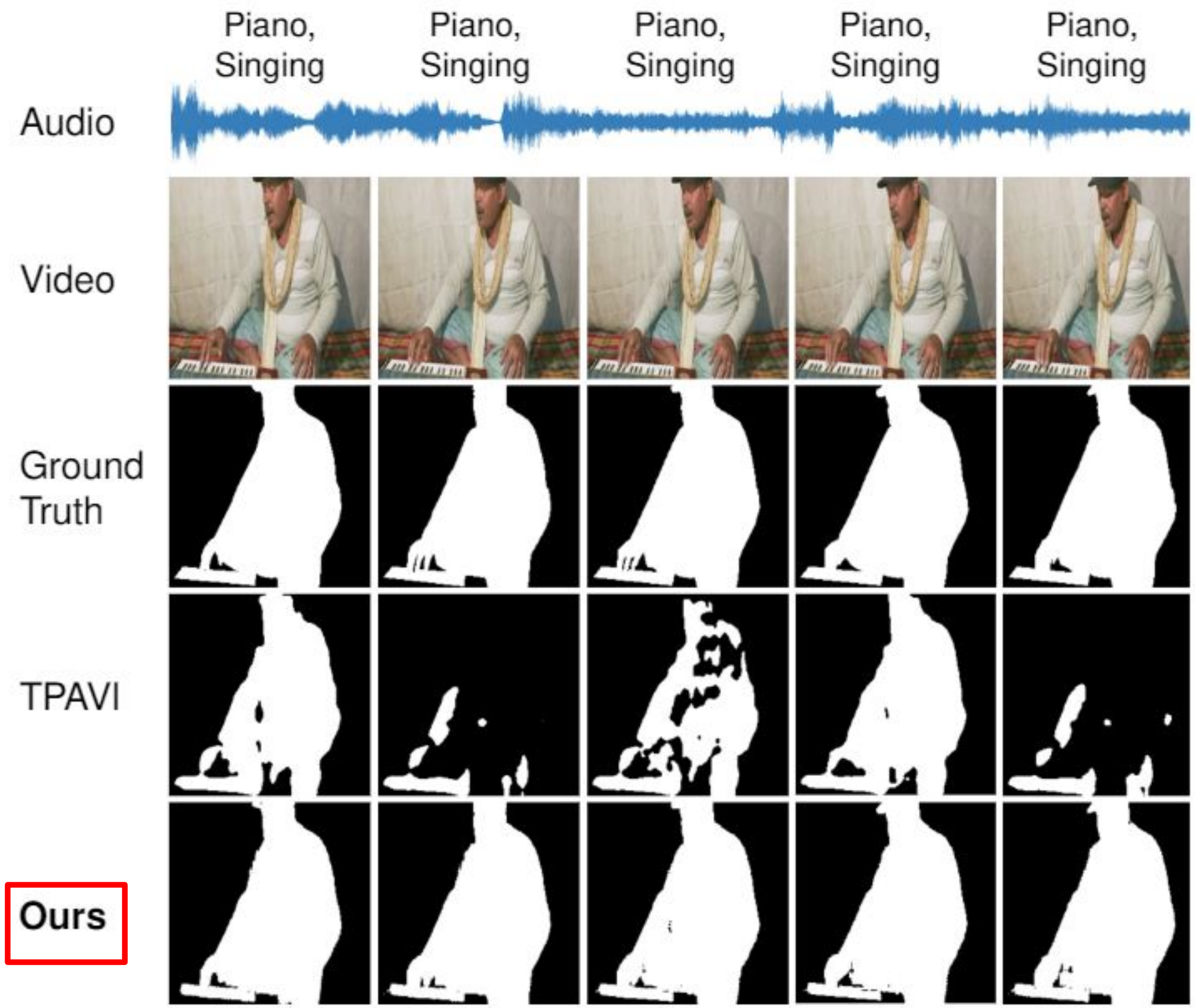
Results | Comparison to SOTA

- Quantitative comparison with non-SAM based methods on the AVSBench
- Ours shows the highest performance in all metrics

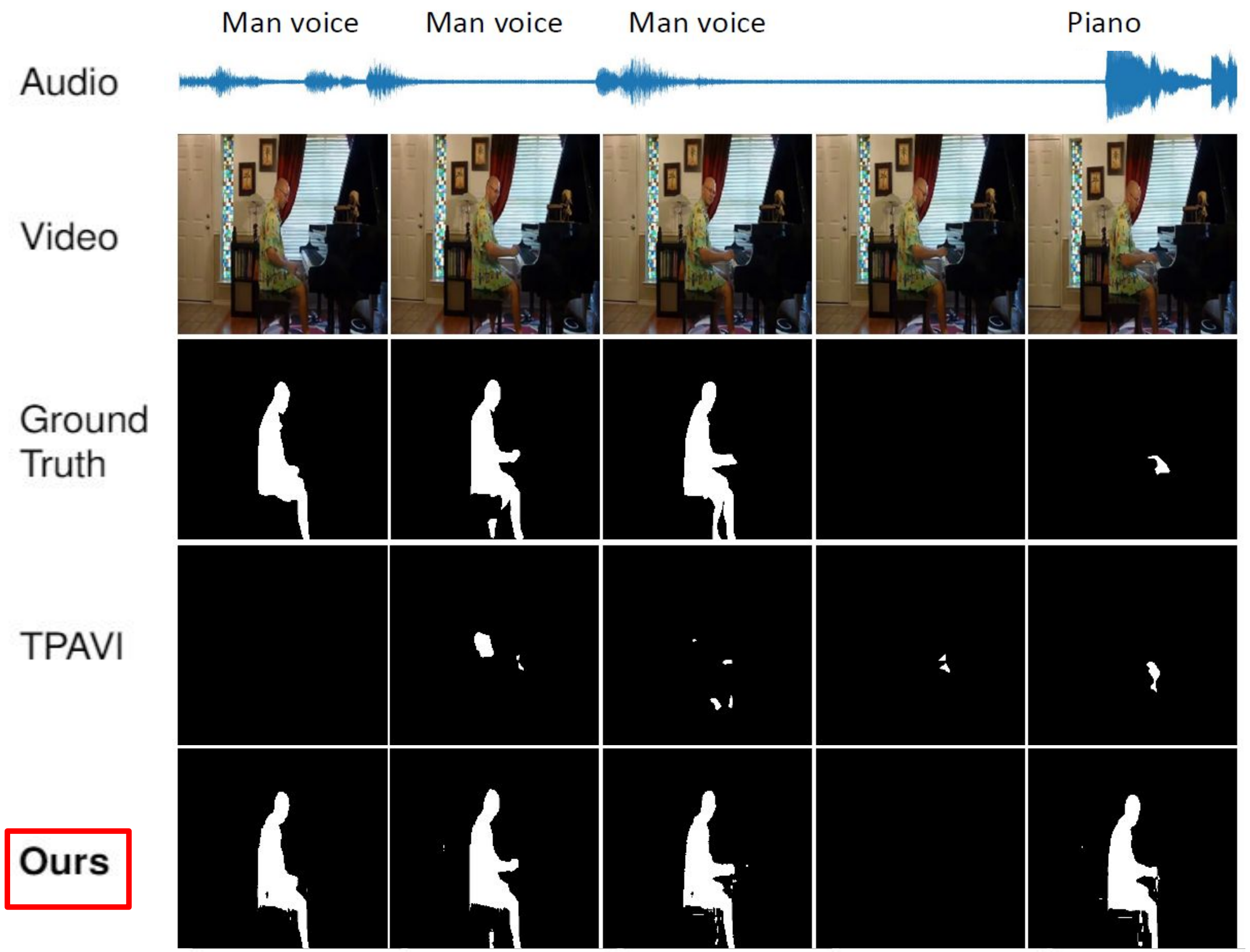
Methods	Single-source		Multi-sources	
	mIoU	F-score	mIoU	F-score
TPAVI [ECCV 22]	78.7	0.88	54.0	0.65
CATR [MM 23]	81.4	0.90	59.0	0.70
AQFormer [IJCAI 23]	81.6	0.89	61.1	0.72
ECMVAE [ICCV 23]	81.7	0.90	57.8	0.71
SAM + ST-BAVA (Ours)	**82.5**	**0.91**	**69.0**	**0.78**

Results on AVS Benchmark

Qualitative results



Qualitative results



SGVR Lab KAIST

Ablation study | Model components

- Baseline - use spatial attention, not the temporal and bidirectional
- Utilizing all attention components performs best

Methods	Single-source mIoU	Single-source F-score	Multi-Sources mIoU	Multi-Sources F-score
Baseline (Spatial Attn.)	76.65	0.857	61.54	0.703
+ Bidirectional Attn.	80.72	0.892	65.37	0.752
+ Temporal Attn.	80.09	0.887	65.17	0.749
Full	**82.46**	**0.906**	**69.01**	**0.776**

Results on AVS Benchmark

Results | Comparison to concurrent works

- **Temporal-Aware** ST-BAVA (ours) outperforms **concurrent SAM-based methods without temporal-awareness**

Methods	Single-source		Multi-sources	
	mIoU	F-score	mIoU	F-score
GAVS [AAAI 24]	80.1	0.90	63.7	0.77
SAMA-AVS [WACV 24]	81.5	0.89	63.1	0.69
ST-BAVA (Ours)	**82.5**	**0.91**	**69.0**	**0.78**

Results on AVS Benchmark

SGVR Lab KAIST

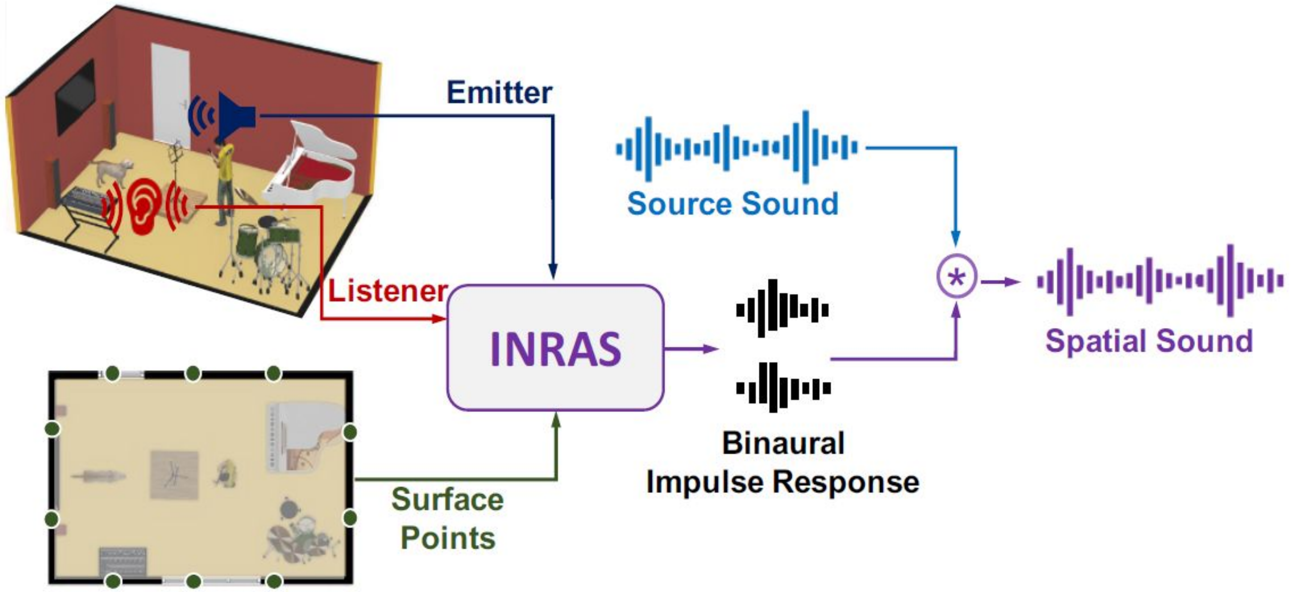
Conclusion

Summary

- Extend SAM into temporal and auditory dimensions for AVS
- Propose a Spatio-Temporal, Bidirectional Audio-Visual Attention (ST-BAVA) module to leverage the audio-visual correspondence across the video sequence
- Achieve meaningful performance enhancement on the AVS benchmark

Future work

- **Acoustic rendering** technology using room geometry and acoustics
- **Applications: VR / AR** (accurately reproduce audio-visual scenes)
- Plan to **utilize the recent 3D representation techniques**, such as NeRF [1] or Gaussian Splatting [2]



[1] NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis, ECCV 2020
[2] 3D Gaussian Splatting for Real-Time Radiance Field Rendering, SIGGRAPH 2023
Image from INRAS: Implicit Neural Representation for Audio Scenes, NeurIPS 2024

SGVR Lab KAIST

Publications

- **First-authored**
  - **Ju-hyeong Seon**, Woobin Im, Sebin Lee, Jumin Lee, Sung-Eui Yoon, **Extending Segment Anything Model into Auditory and Temporal Dimensions for Audio-Visual Segmentation**, *ICIP 2024* (Under review)
  - **Ju-hyeong Seon**, Jaeyoon Kim, Joo Young Kim, Young Ju Lee, Hye-kyung Han, Sung-Eui Yoon, 비디오 내 음원 위치 추정 모델의 성능 향상을 위한 클래스 인지 대조 학습 기법 제안, *한국정보과학회 KTCP 2023 (KCI 저널)*

- **Co-authored**
  - Guoyuan An, **Ju-hyeong Seon**, InKyu An, Yuchi Huo, Sung-Eui Yoon, **Topological RANSAC for instance verification and retrieval without fine-tuning**, *NeurIPS 2023*
  - Jumin Lee*, Sebin Lee*, Changho Jo, Woobin Im, **Ju-hyeong Seon**, and Sung-Eui Yoon, **SemCity: Semantic Scene Generation with Triplane Diffusion**, *CVPR 2024* (Accepted)

SGVR Lab KAIST

References

- G. Bertasius, H. Wang, and L. Torresani, “Is space-time atten-tion all you need for video understanding?,” in ICML, 2021, number 3, p. 4.
- J. Liu, Y. Wang, C. Ju, C. Ma, Y. Zhang, et al., “Annotation-free audio-visual segmentation,” Proc. WACV, 2024.
- Y. Wang, W. Liu, G. Li, J. Ding, D. Hu, and X. Li, “Prompting segmentation with sound is generalizable audio-visual source localizer,” arXiv preprint arXiv:2309.07929, 2023.
- Q. Shen, X. Yang, and X. Wang, “Anything-3d: Towards single-view anything reconstruction in the wild,” arXiv preprint arXiv:2304.10261, 2023.
- J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel, “Medical sam adapter: Adapting segment anything model for medical image segmentation,” arXiv preprint arXiv:2304.12620, 2023.
- Wang, W. Zhou, Y. Mao, and H. Li, “Detect any shadow: Segment anything for video shadow detection,” arXiv preprint arXiv:2305.16698, 2023

References

- C. Liu, P. P. Li, X. Qi, H. Zhang, L. Li, D. Wang, and X. Yu, “Audio-visual segmentation by exploring cross-modal mutual semantics,” in Proc. ACM MM, 2023, pp. 7590–7598
- S. Huang, H. Li, Y. Wang, H. Zhu, J. Dai, J. Han, et al., “Discovering sounding objects by audio queries for audio visual segmentation,” arXiv preprint arXiv:2309.09501, 2023

Thank you for listening

SGVR Lab
KAIST

Appendix

SGVR Lab
KAIST

SAM for AVS

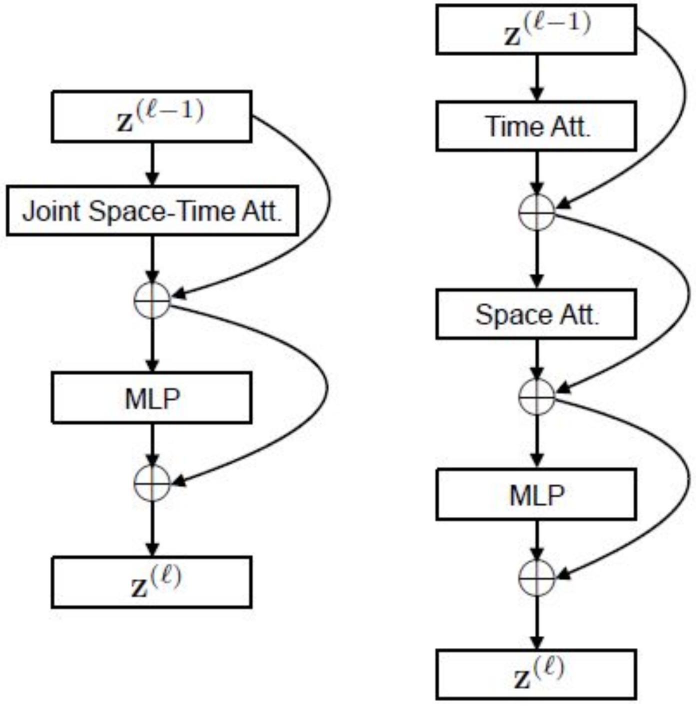
Original SAM

- **Original SAM**
  - Boxes or Points as query
  - Users manually give queries for segmentation

Related work

Divided Space-Time Attention in video classification

- Efficient and effective performance
- Not explored yet in Audio-Visual Learning



Joint Space-Time Attention (ST)

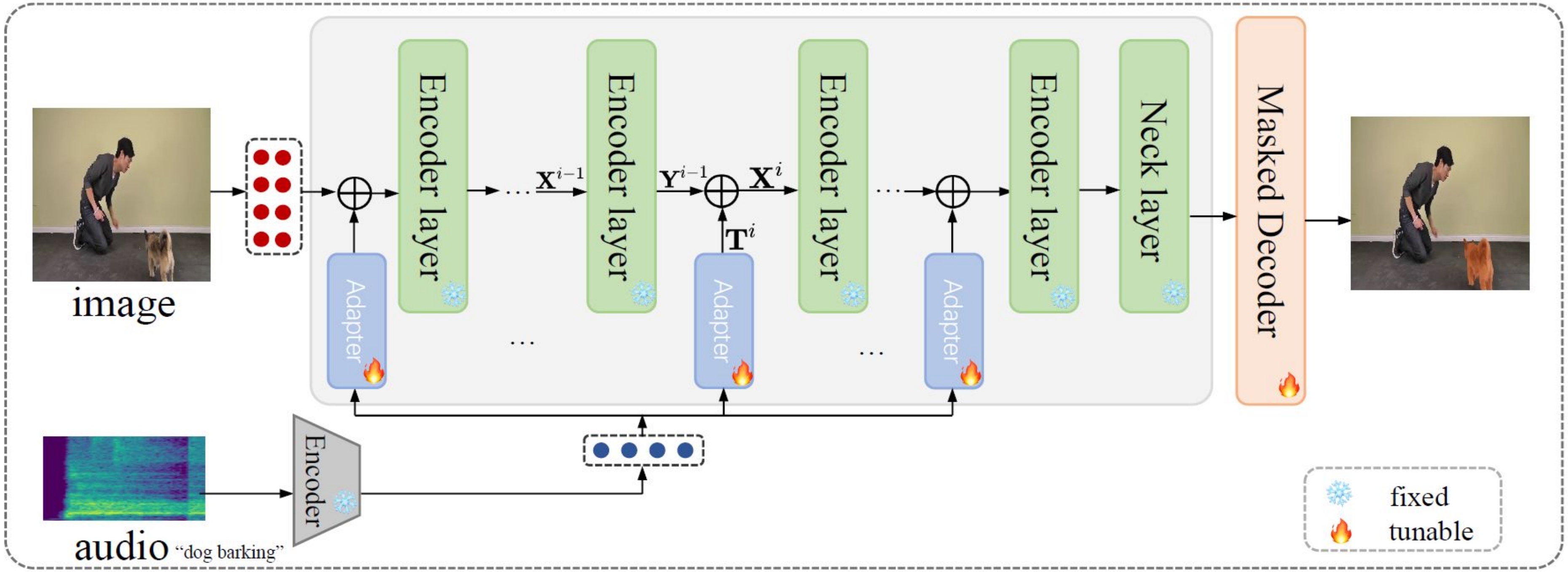
Divided Space-Time Attention (T+S)

SGVR Lab KAIST

Is Space-Time Attention All You Need for Video Understanding?, ICML 2021

SAM in Audio-Visual Segmentation

- Recent approaches use prompt tuning of SAM with adaptors[1,2]
- Didn't utilize the temporal information, limiting SAM's performance on AVS - Predict per image, not per video



SGVR Lab KAIST

[1] Annotation-free Audio-Visual Segmentation, WACV 2024
[2] Prompting Segmentation with Sound is Generalizable Audio-Visual Source Localizer, ICCVW 2023

Adapter

- We use Adapters* to help the subsequent operation of ST-BAVA
- Designed to inject audio feature in the image encoding stage

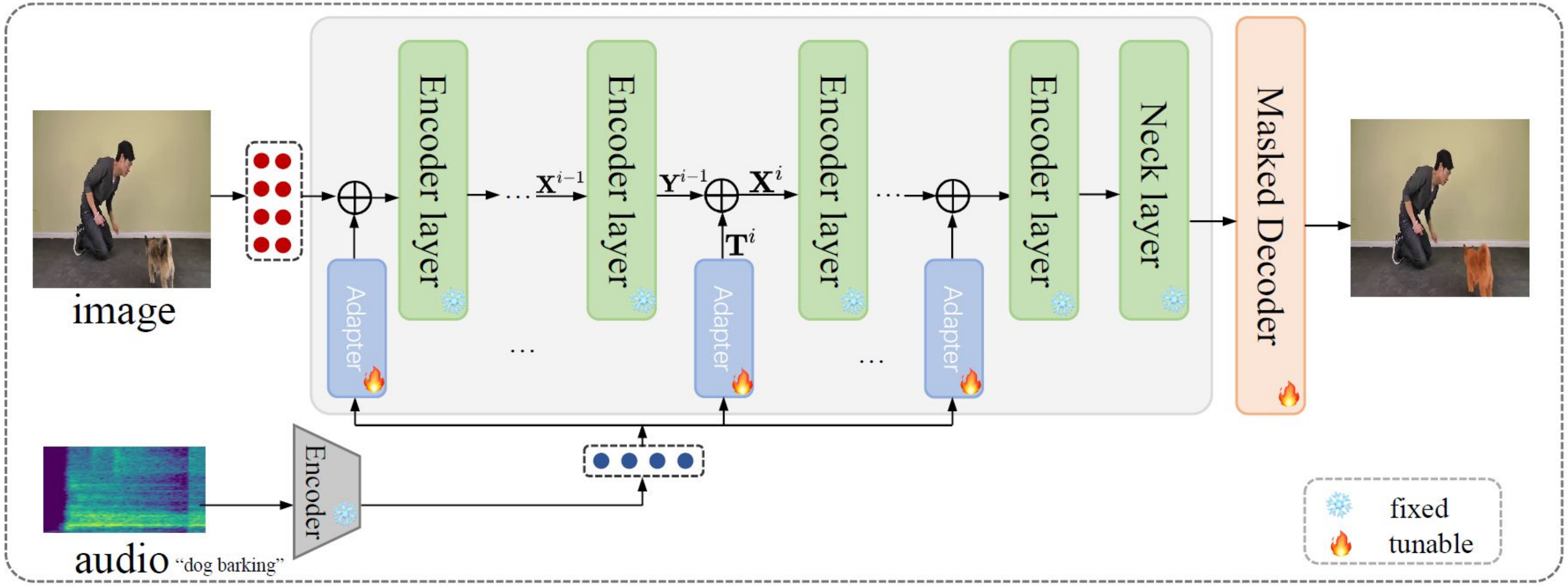
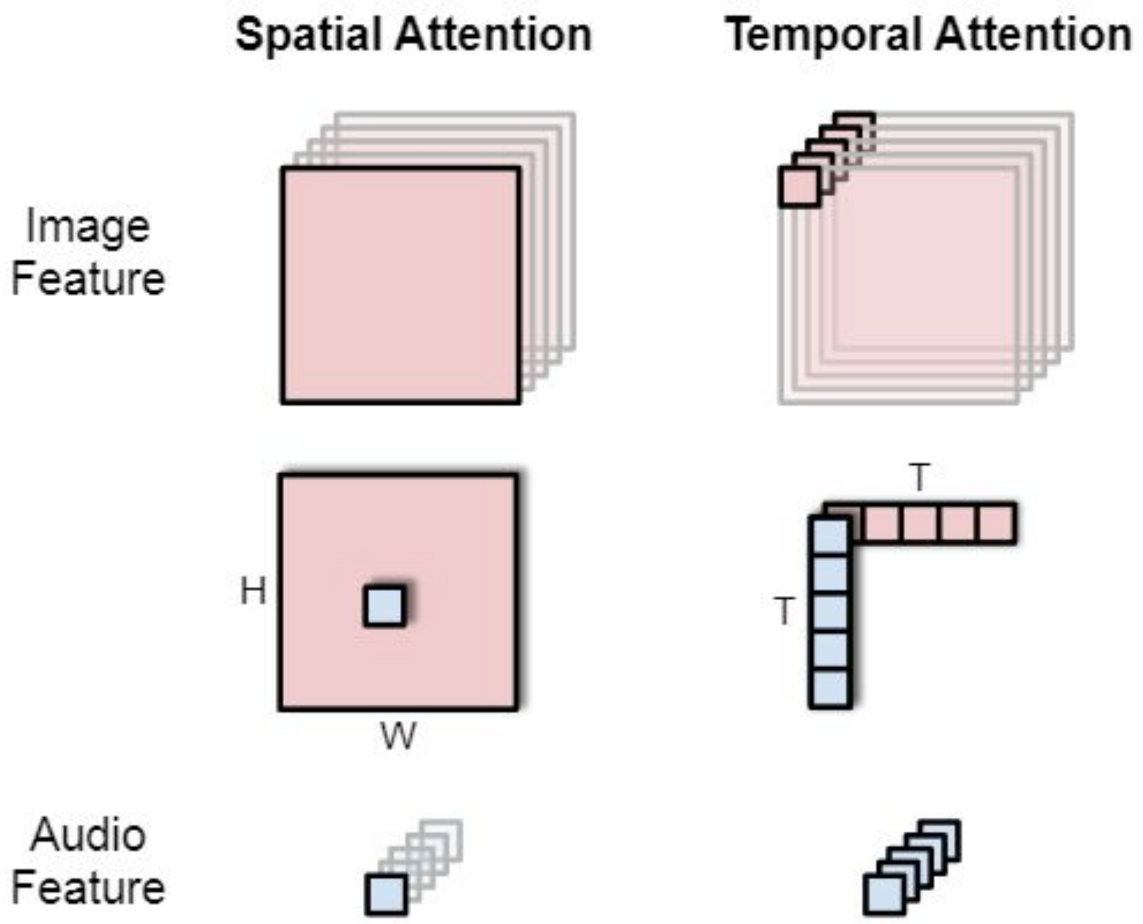


Figure from *Annotation-free audio-visual segmentation, WACV 2024

SGVR Lab KAIST

ST-BAVA | Attention components

- **Spatial attention** captures the audio-visual relationship per frame
- **Temporal attention** captures the relationship across consecutive frames per pixel



Evaluation Metric

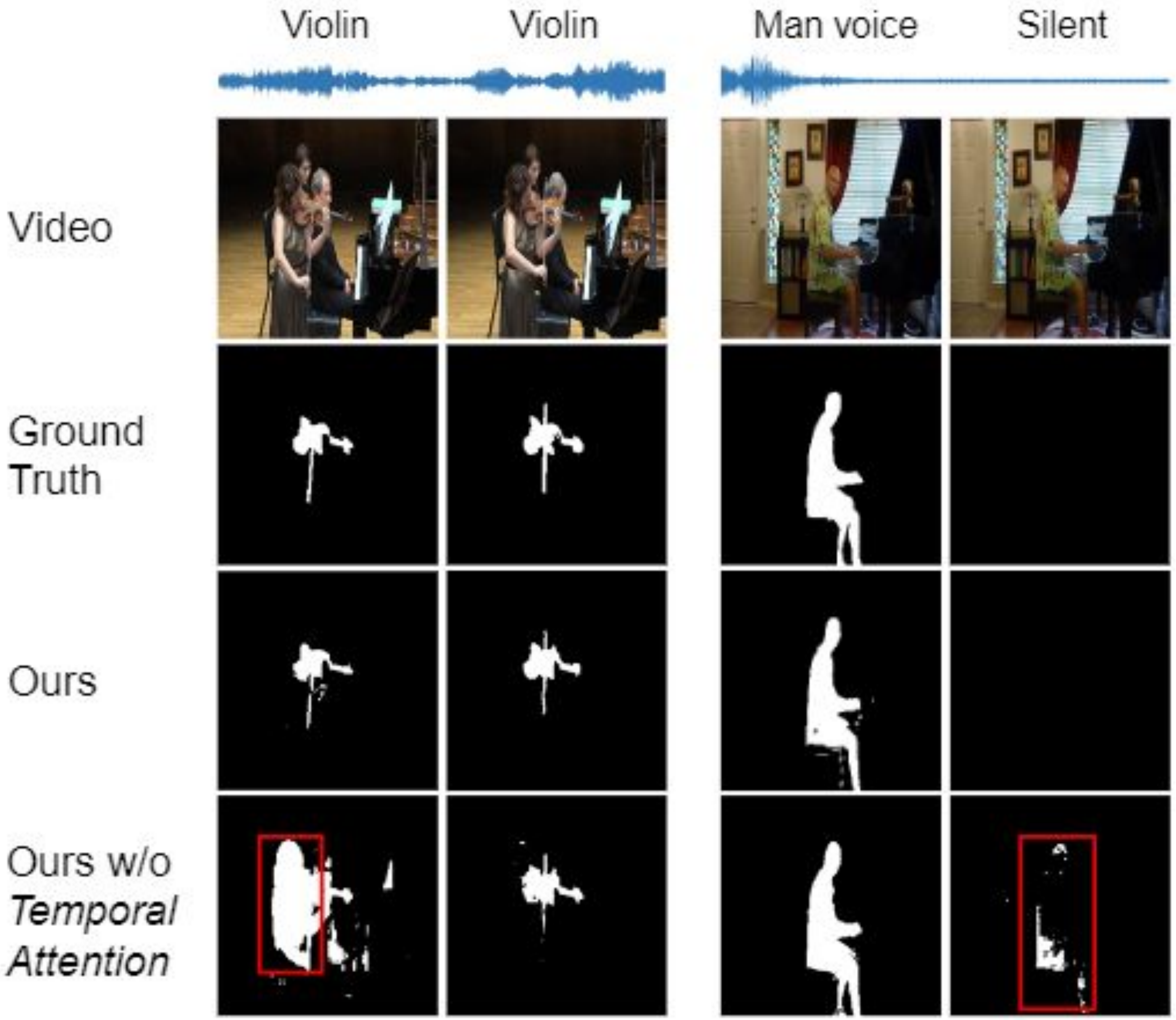
- mIoU = Inter(y, y_pred) / Union(y, y_pred)
- F-score = $\frac{(1+\beta^2)\times precision \times recall}{\beta^2 \times precision + recall}$
  - Precision = Inter(y, y_pred) / y_pred
  - Recall = Inter(y, y_pred) / y

y: Ground truth

y_pred: Prediction

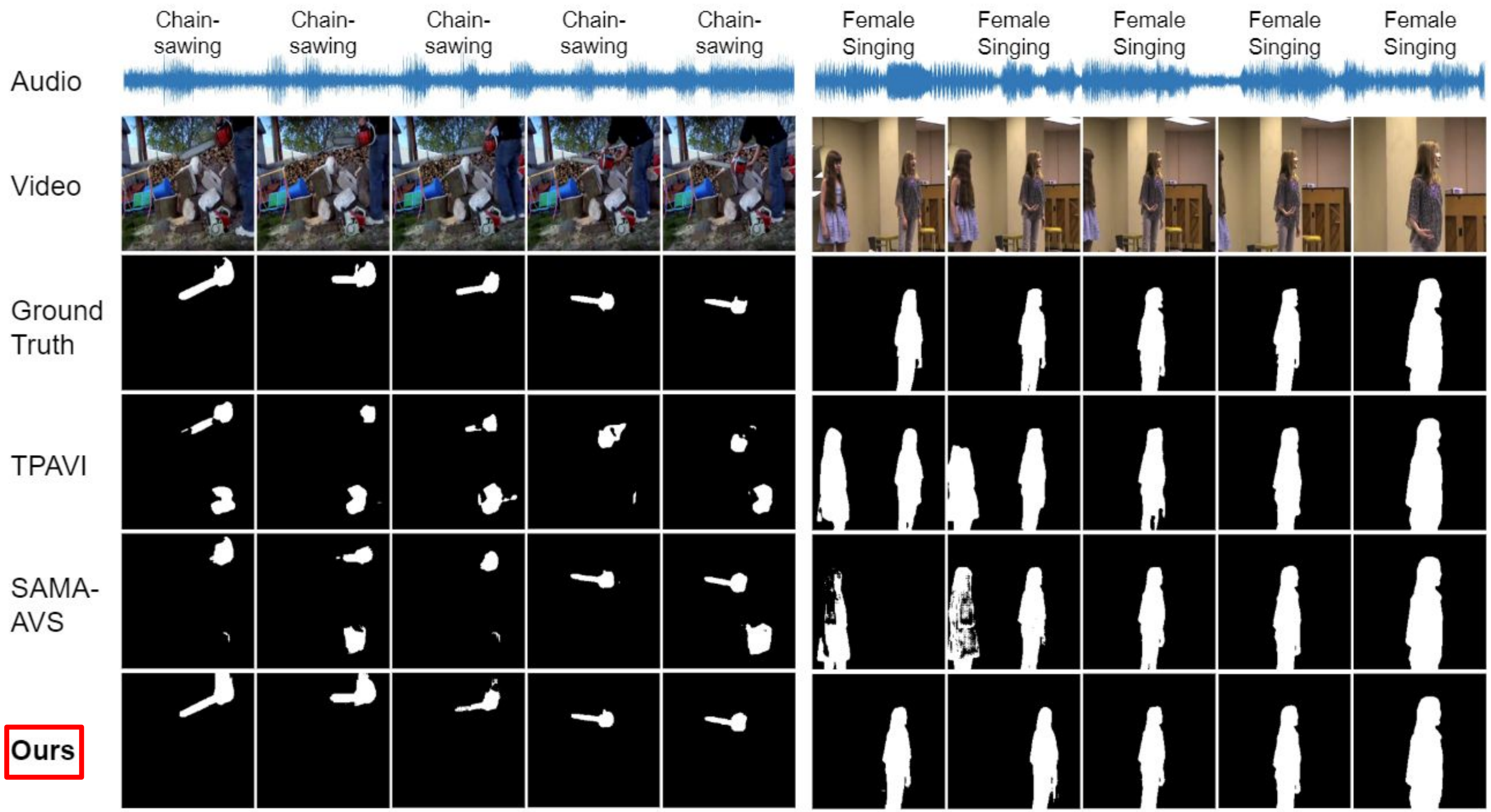
SGVR Lab KAIST

Ablation study | Model components



- Qualitative results show the effects of temporal attention in ST-BAVA

Qualitative results



SGVR Lab
KAIST

Ablation study | Intermediate feature fusion module

- TPAVI[14] is a fusion module proposed in other AVS work
  - **Doesn't use bidirectional attention**, showing not good results
- CMRAN[33], HAN[34], JCA[35] are proposed in other A-V tasks
  - **Don't utilize the spatial visual features**, showing not good result

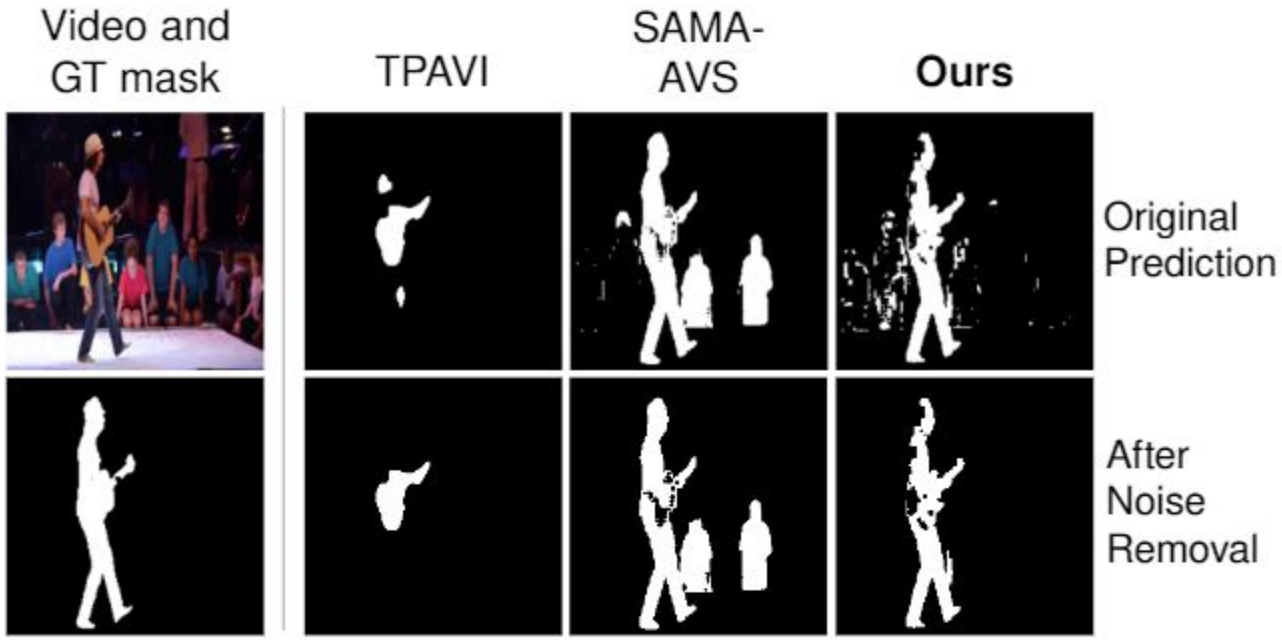
Approach	Methods	S4		MS3	
		mIoU	F-score	mIoU	F-score
Audio Prompts with Training	w/o fusion module [1]	81.53	0.886	63.14	0.691
	+ TPAVI [14]	81.68	0.902	64.78	0.749
	+ HAN [33]	80.56	0.896	64.14	0.739
	+ CMRAN [34]	81.46	0.899	65.09	0.747
	+ JCA [35]	81.99	0.903	65.44	0.751
	+ ST-BAVA (Ours)	**82.46**	**0.906**	**69.01**	0.776

[1] Audio-Visual Segmentation, ECCV 2022
[33] Cross-modal relation-aware networks for audio-visual event localization, ACM MM 2020
[34] Unified multisensory perception: Weakly-supervised audio-visual video parsing, ECCV 2020
[35] A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition, CVPR 2022

Failure case

Weakness on distinguishing the semantically similar visual objects

- SAM doesn’t have good understanding on the object semantics
- Auxiliary consideration to the object semantic could be introduced



SGVR Lab KAIST