How Can Contrastive Pre-training Benefit Audio-Visual Segmentation? A Study from Supervised and Zero-shot Perspectives

Jiarui Yu[1], Haoran Li[1], Yanbin Hao[1]*, Jinmeng Wu[2], Tong Xu[1], Shuo Wang[1] and Xiangnan He[1]

1 University of Science and Technology of China 2 Wuhan Institute of Technology

BMVC 2023

Motivation & Abstract

Sharing a similar spirit with the successful contrastive language-image pre-training (CLIP), audio-aware contrastive pre-training has also exhibited its powerful ability to align cross-model instances. In this paper, we aim to we explore the following question: **how can the instance-level alignment knowledge gained from contrastive pre-training benefit pixel-level audio-visual segmentation (AVS) ?** To address this question, we approach the problem from two perspectives:

* Supervised Audio-Visual Segmentation (AVS)

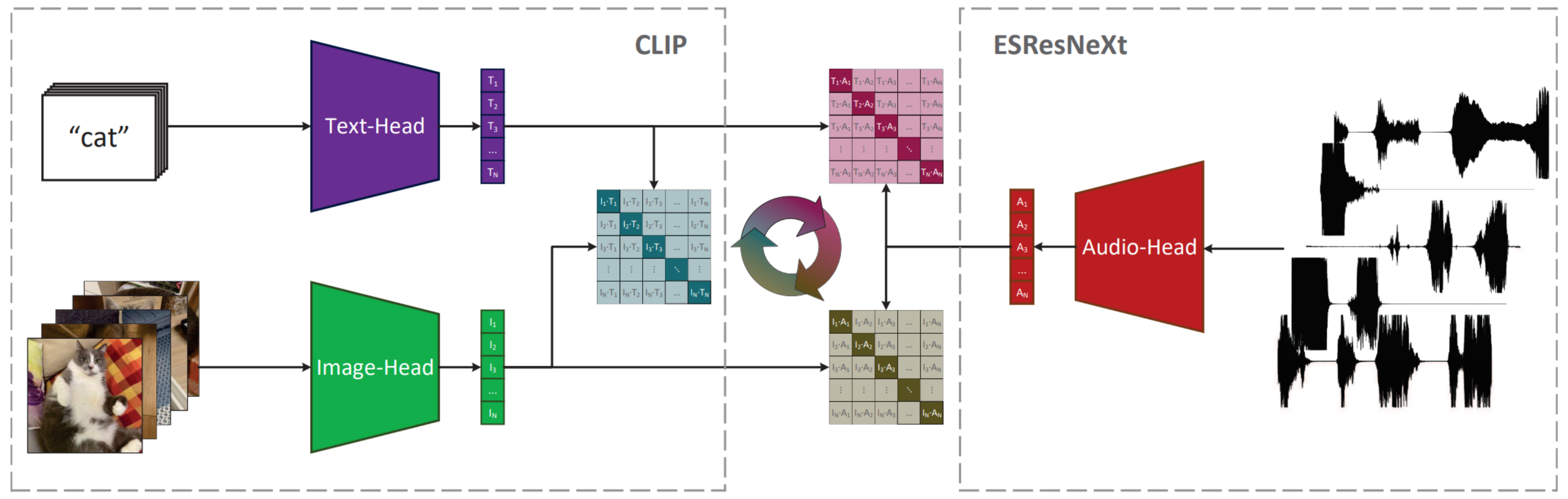
Transfer learning with pre-trained instance-level model AudioCLIP**,** leading to a simple yet effective model AC-FPN that enables pixel-level predictions for sounding objects.

* Zero-shot audio-visual segmentation (AVS)

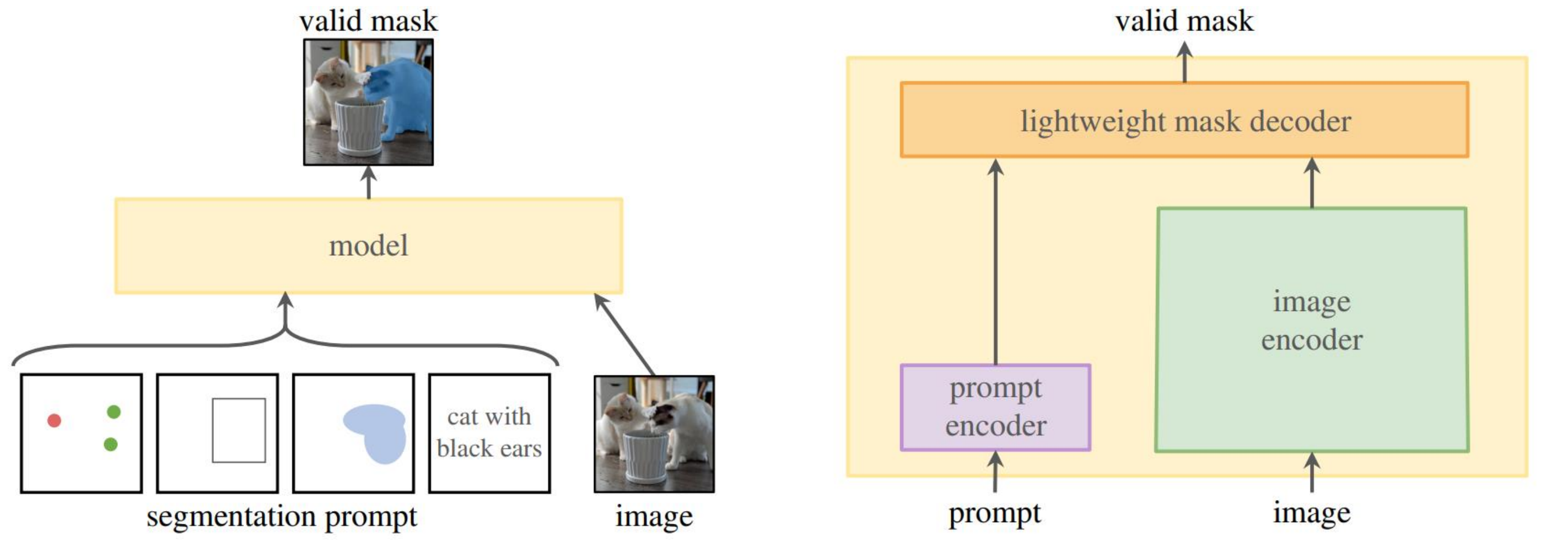
Promote the Segment-Anything-Model (SAM) for AVS by proposing three prompt formulizing strategies based on instance-level contrastive pre-training models.

Related Work

AudioCLIP: Audio-language-image tri-modal alignment



Segment-Anything-Model: Promptable Segmentor

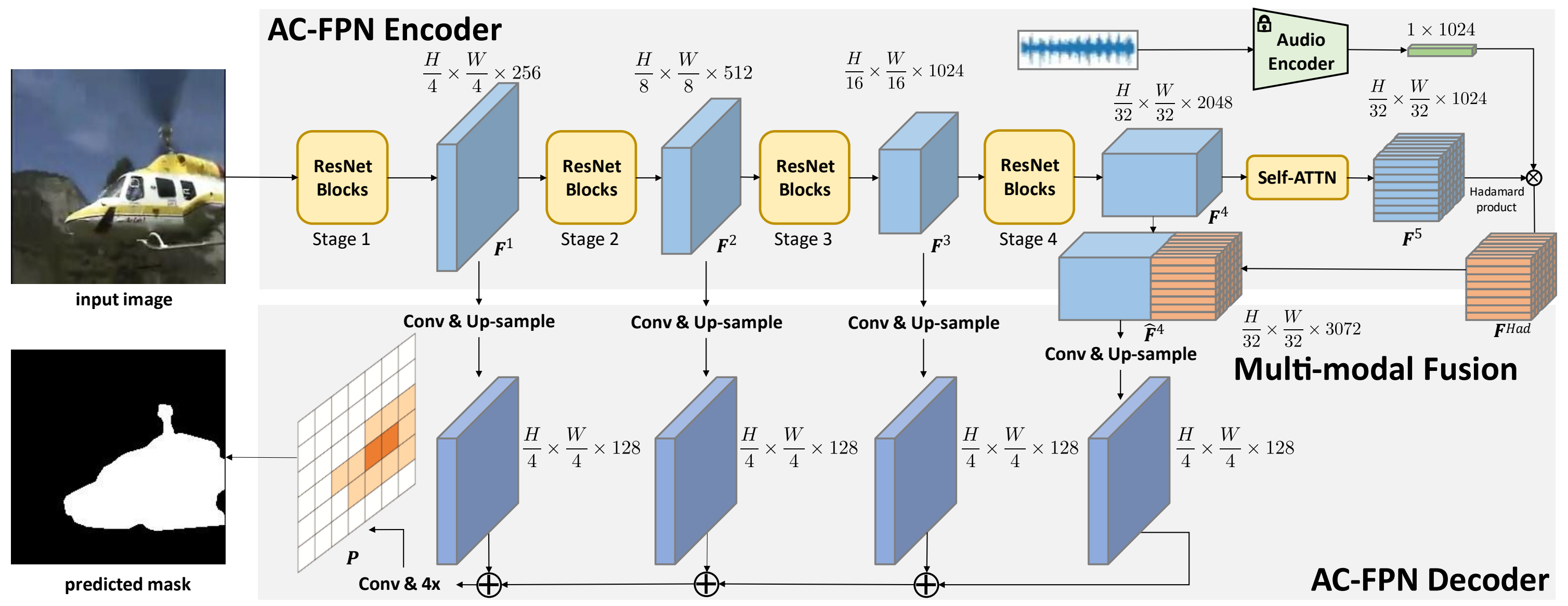


(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**)

Method

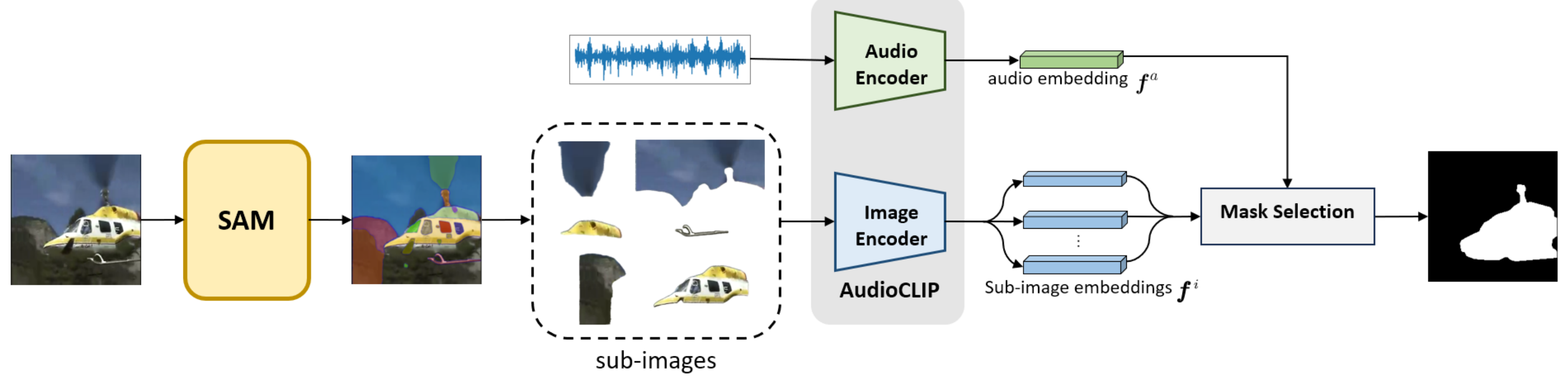
Supervised AVS

- * Mine visual-audio alignment information from visual feature map
- * Use simplest fusion strategy – Hadamard Dot

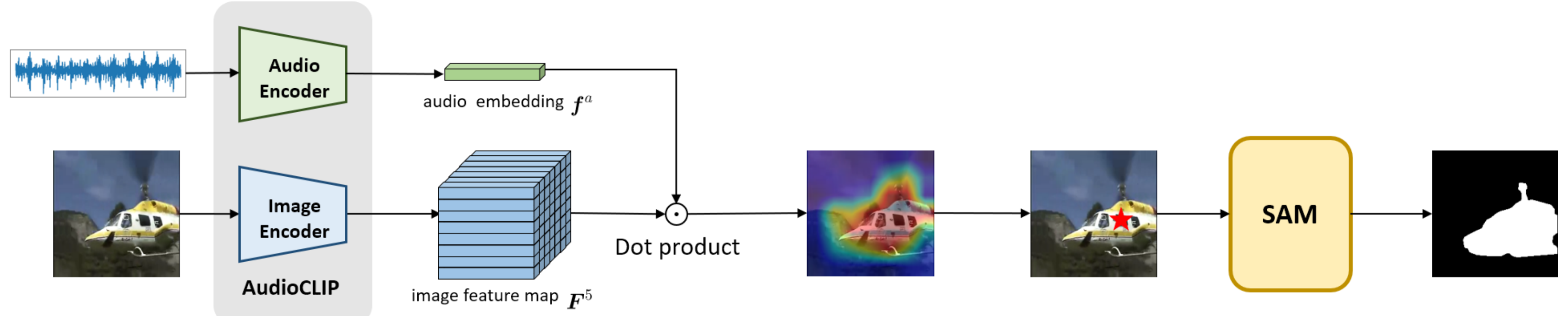

AC-FPN Encoder
input image
ResNet Blocks
Stage 1
Stage 2
Stage 3
Stage 4
$\frac{H}{4}\times\frac{W}{4}\times256$
$\frac{H}{8}\times\frac{W}{8}\times512$
$\frac{H}{16}\times\frac{W}{16}\times1024$
$\frac{H}{32}\times\frac{W}{32}\times2048$
$\frac{H}{32}\times\frac{W}{32}\times1024$
1×1024
F^1
F^2
F^3
F^4
F^5
F^{Had}
$\hat{F}^4$
Audio Encoder
Self-ATTN
Hadamard product
$\frac{H}{32}\times\frac{W}{32}\times3072$
Conv & Up-sample
Multi-modal Fusion
$\frac{H}{4}\times\frac{W}{4}\times128$
P
Conv & 4x
predicted mask
AC-FPN Decoder

Zero-shot AVS

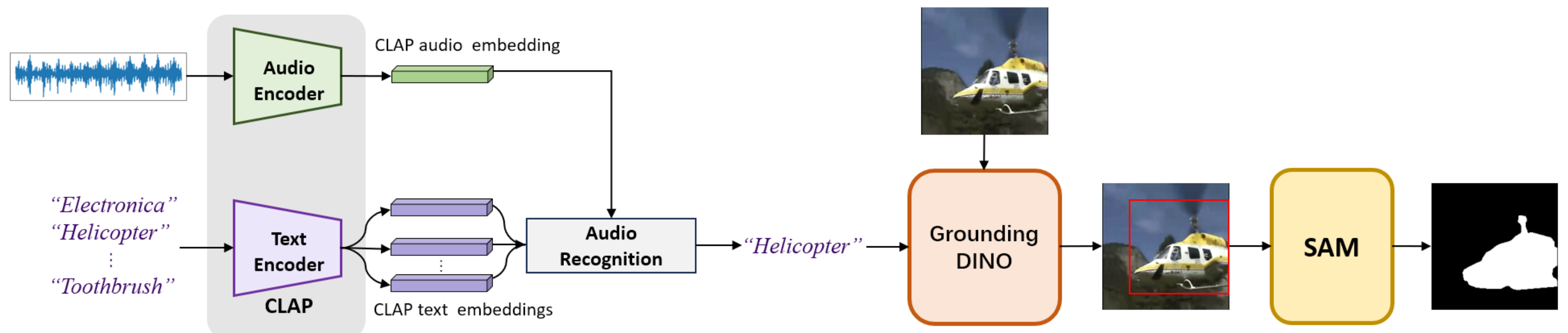
*** Strat. 1 No-Prompt:** Segment Everything + AudioCLIP filter



*** Strat. 2 Point-Prompt:** AudioCLIP Heatmap → Point → Mask



*** Strat. 3 Box-Prompt:** CLAP → Category → Box → Mask



Results

Results on Supervised AVS

- * Higher performance
- * Fewer parameters

Method	S4		MS3		Fixed Params. ↓	Tunable Params. ↓
	mIoU ↑	F-score ↑	mIoU ↑	F-score ↑		
TPAVI-ResNet50 [30]	72.79	.848	47.88	.578	72.1M	91.4M
AC-FPN (Hadamard)	77.12	.874	**49.95**	.635	**32.1M**	**68.0M**
AC-FPN (Concatenation)	**77.29**	**.879**	48.63	**.637**	**32.1M**	68.2M

Results on Zero-shot AVS

- * Box-Prompt is best
- * Point-Prompt works without category list

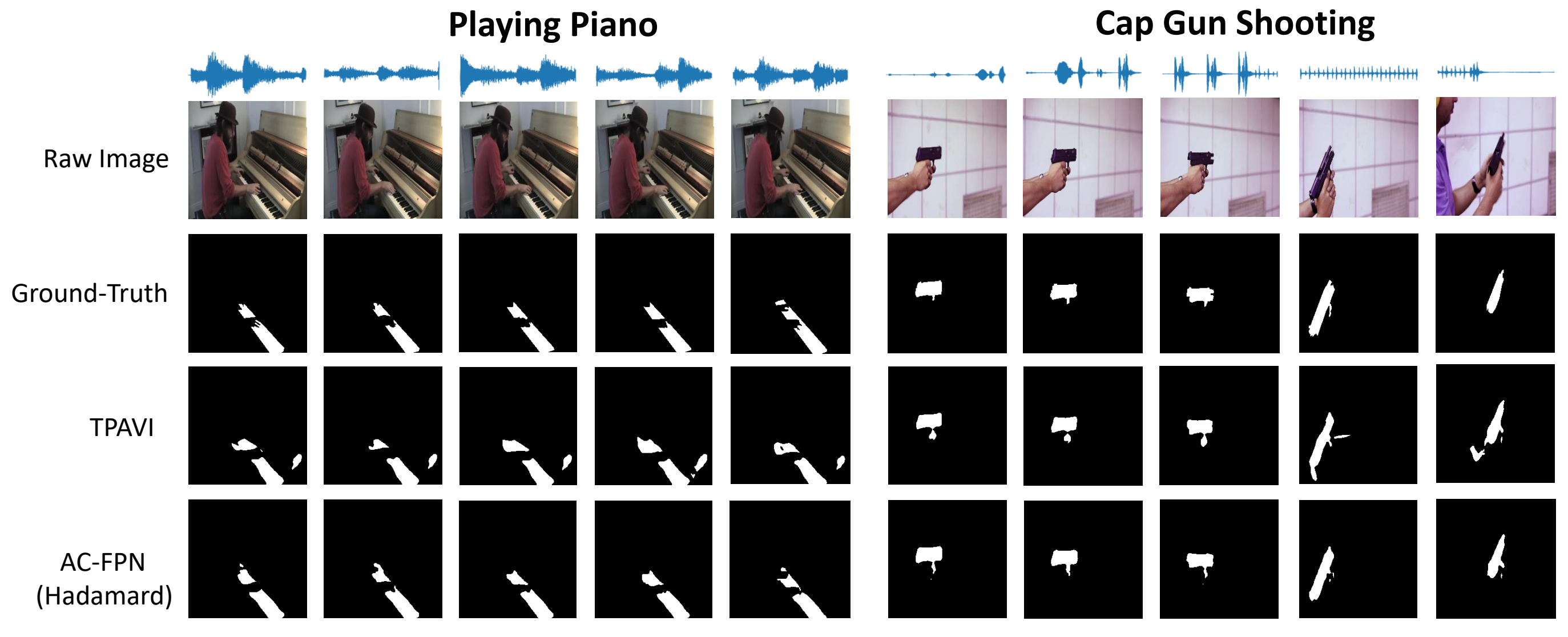
Method	S4		MS3	
	mIoU	F-score	mIoU	F-score
Random-SAM	7.0	.240	11.5	.187
Full-mask	19.0	.226	12.7	.170
No-Prompt	23.8	.358	19.7	.242
Point-Prompt(global)	27.2	.424	19.4	.279
Point-Prompt(local)	30.7	.416	20.0	.270
Point-Prompt(dense)	40.3	.515	28.8	.333
Box-Prompt	**51.2**	**.615**	**41.8**	**.478**

Contact with us

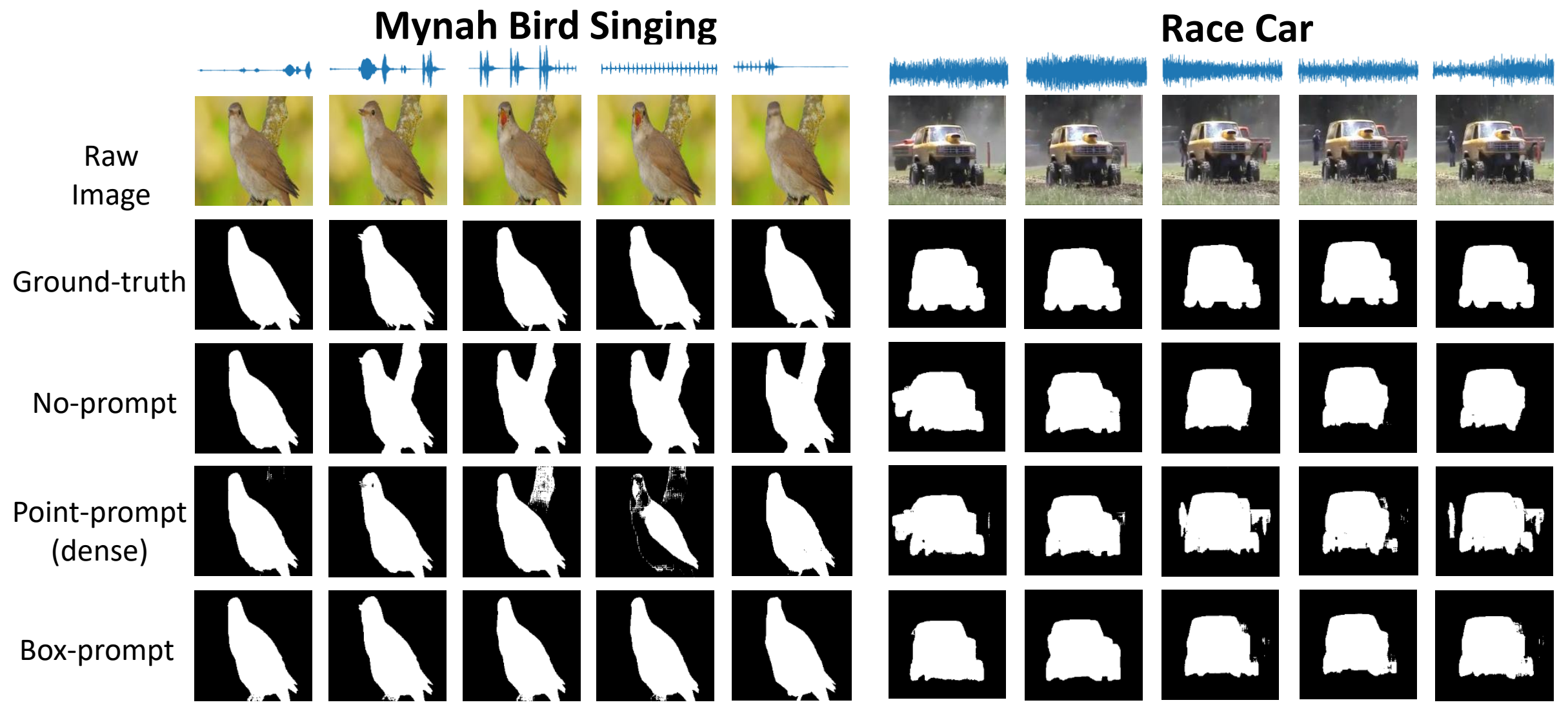
yjr@mail.ustc.edu.cn
lihaoran747@126.com

Feel free to scan our QR code and follow our work.

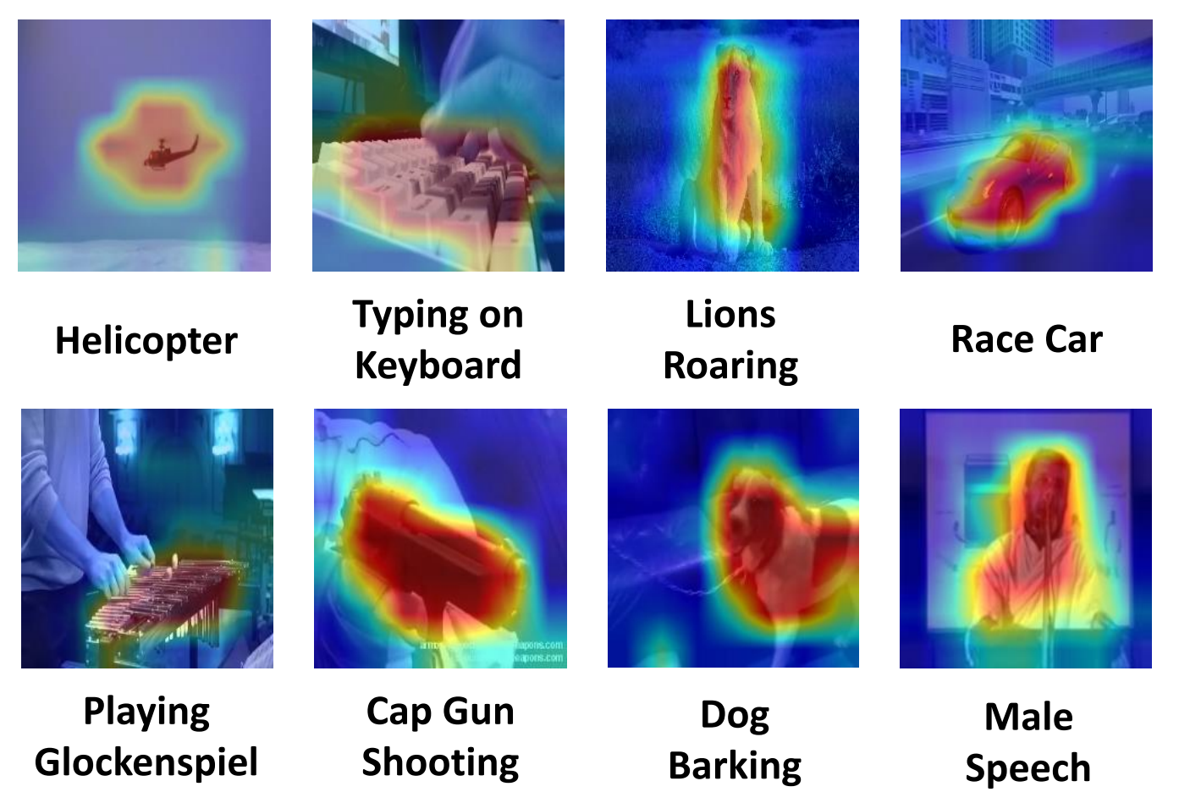
Visualization



Supervised AVS



Zero-shot AVS



Heatmap with different sounding objects