CVPR Nashville JUNE 11-15, 2025

Revisiting Audio-Visual Segmentation with Vision-Centric Transformer

Shaofei Huang[1,2] Rui Ling[3] Tianrui Hui[1] Hongyu Li[4] Xu Zhou[5] Shifeng Zhang[5] Si Liu[3] Richang Hong[1] Meng Wang[1]

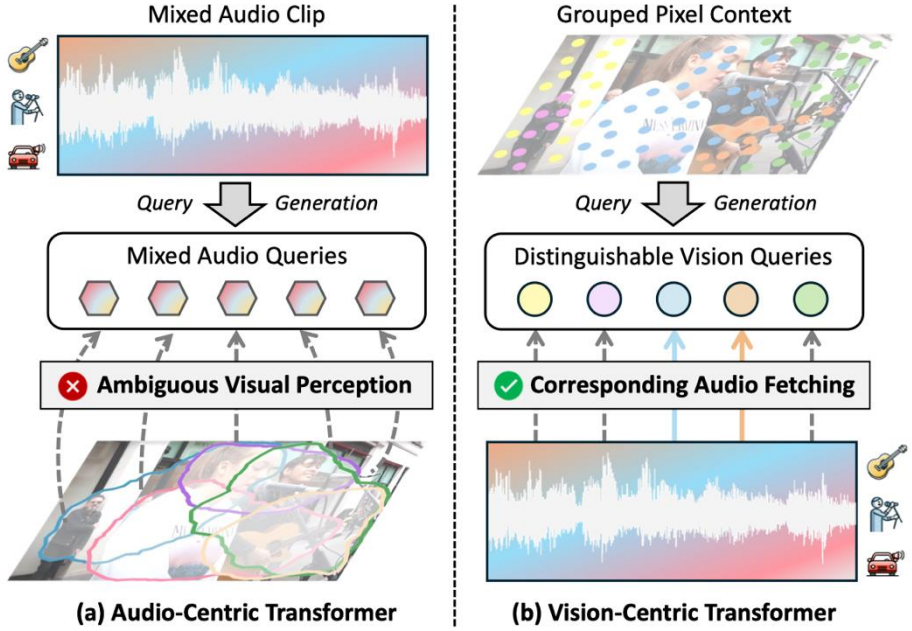
[1]HFUT [2]IIE, CAS [3]SCSE, BUAA [4]SAI, BUAA [5]Sangfor

Task

- Predict the pixel-level mask of the objects which make sounds in the given video clip according to the associated audio information

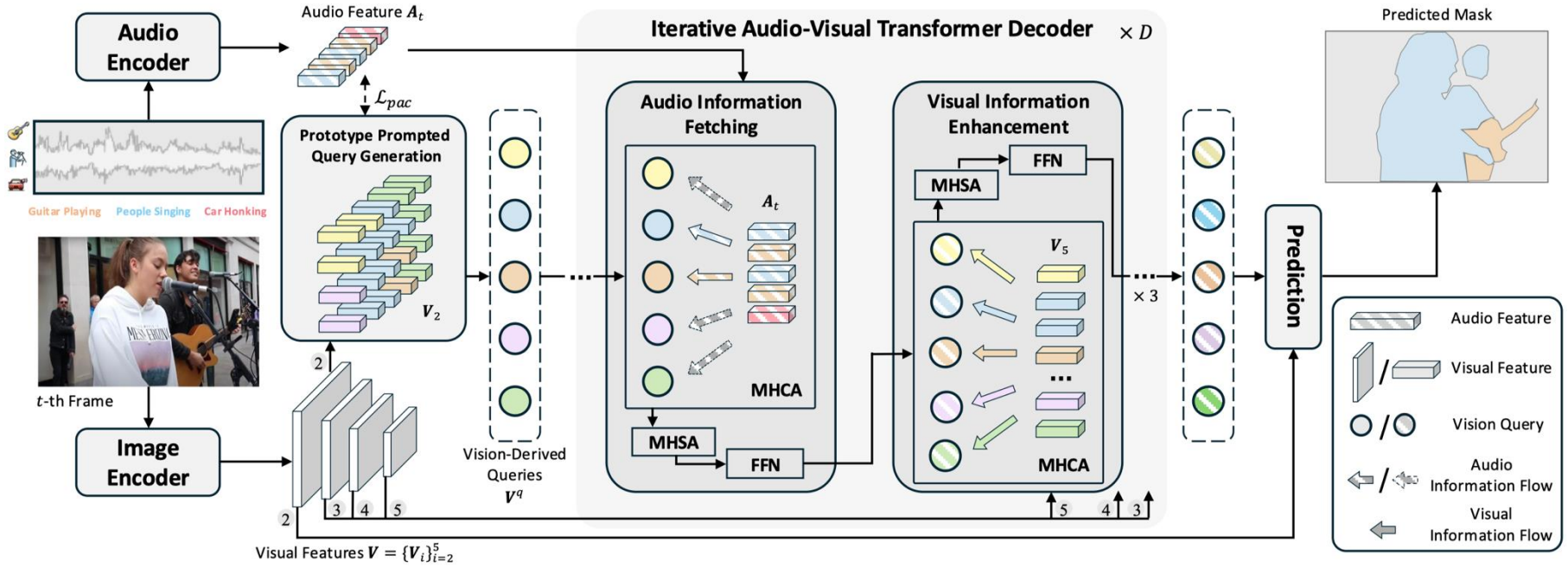
Motivation

- **Audio-Centric Transformer (ACT)**
  - Audio-derived queries from mixed sound sources causes perception confusion for audio-derived queries
  - Delayed integration of visual information may lead to the loss of visual details
- **Visual-Centric Transformer (VCT)**
  - Vision-derived queries focus on semantically distinct visual regions, showing superior discriminative ability
  - Interacting with audio features excludes unrelated sounds and alleviates the perceptual ambiguity



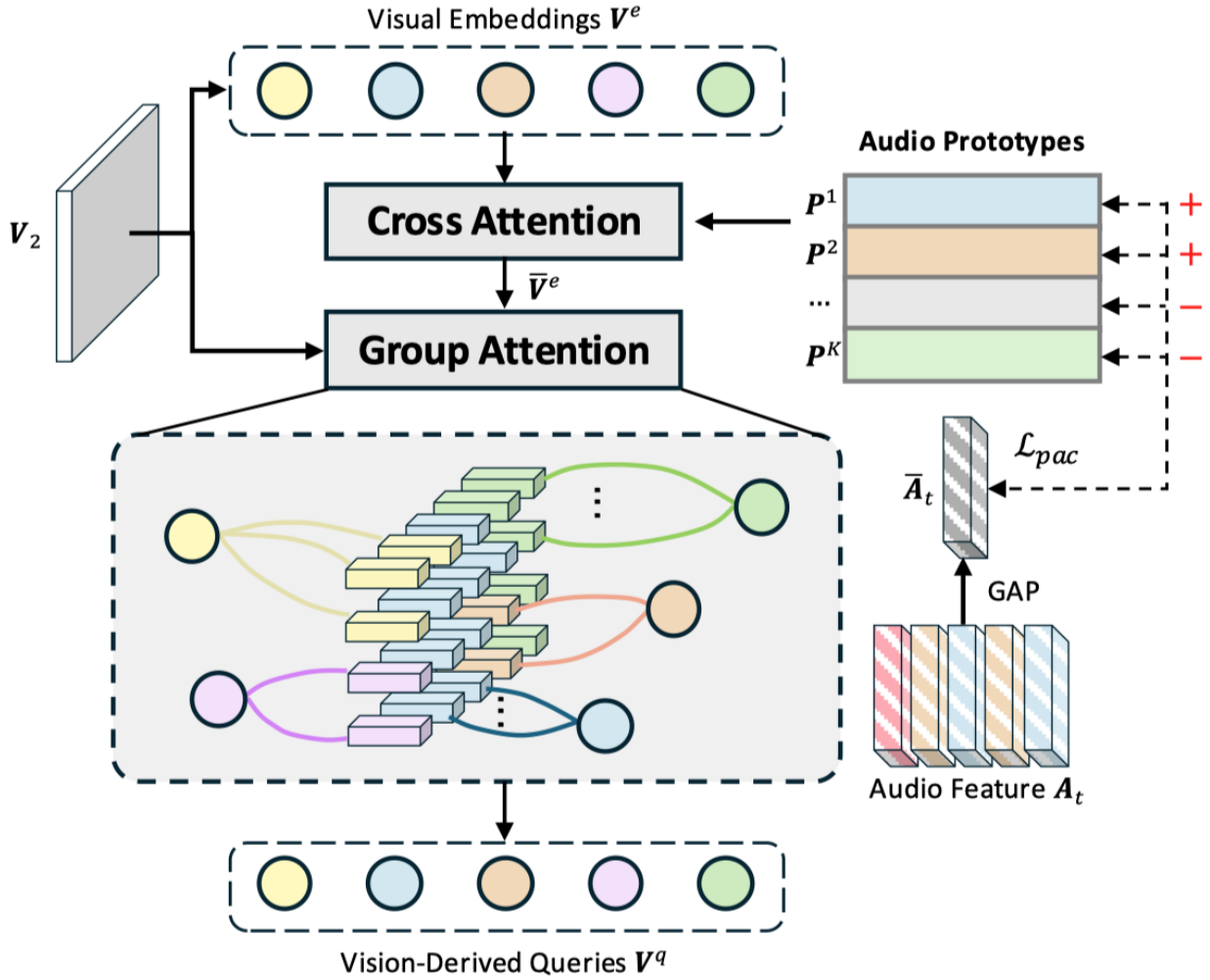
Framework

- Our VCT framework leverages **vision-derived queries** to fetch audio information and visual details in the **iterative audio-visual Transformer decoder**
  - **Audio information fetching**: obtain the corresponding sound information for each query region
  - **Visual information enhancement**: captures multi-scale fine-grained visual features for accurate mask predictions



Prototype Prompted Query Generation (PPQG)

- PPQG generates vision-derived queries that contain both rich visual details and audio semantics



Prototype Prompted Query Generation (PPQG)

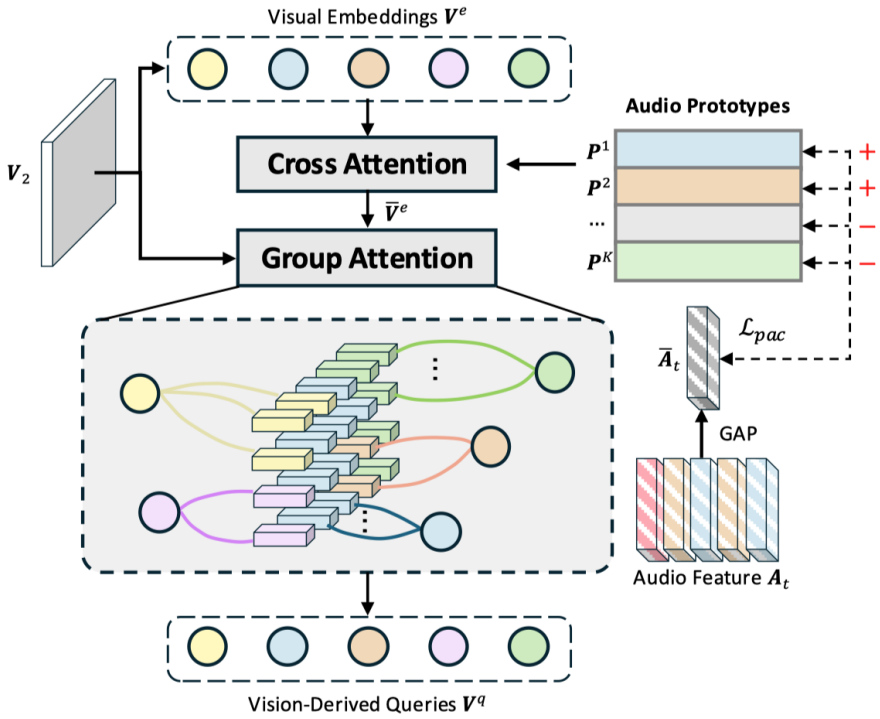
I. **Visual embedding aggregation:** perform spatial information aggregation obtain a set of visual embeddings as initial vision-derived queries

$$\boldsymbol{V}^h = \mathrm{Conv}_{1\times1}(\delta(\mathrm{Conv}_{3\times3}(\delta(\mathrm{Conv}_{1\times1}(\boldsymbol{V}_2))))),$$

$$\boldsymbol{V}^e = \mathrm{Reshape}(\mathrm{MLP}(\mathrm{Reshape}(\boldsymbol{V}^h))),$$

II. **Audio prototype prompting**: define audio prototypes to prompt queries with audio event categories present in the scene.

$$\bar{\boldsymbol{V}}^e = \boldsymbol{V}^e + \mathrm{Softmax}(\frac{(\boldsymbol{V}^e\boldsymbol{W}_1^q)(\boldsymbol{P}\boldsymbol{W}_1^k)^{\mathrm{T}}}{\sqrt{C^h}})(\boldsymbol{P}\boldsymbol{W}_1^v).$$

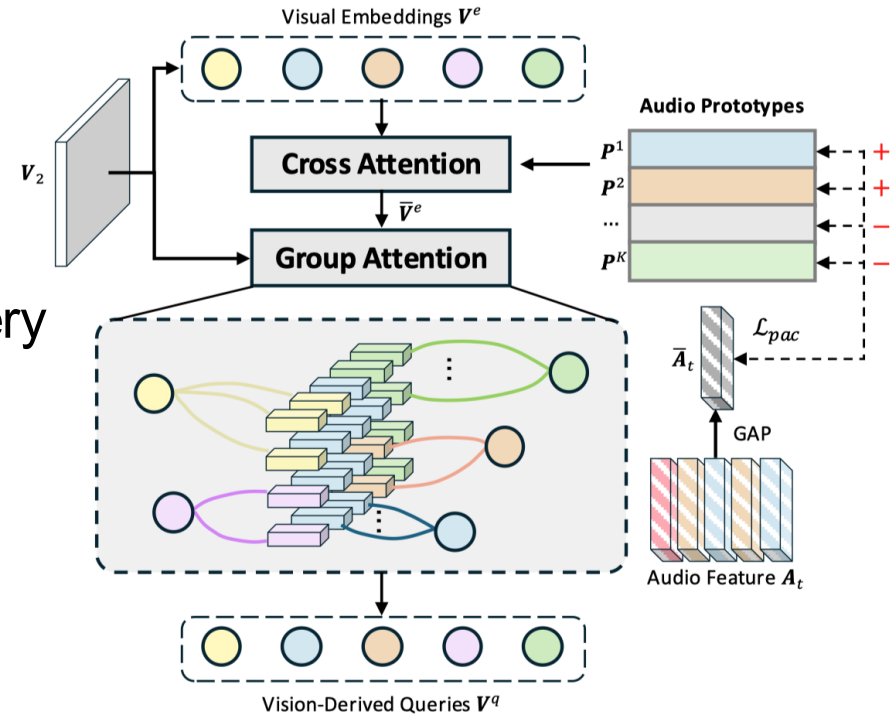


Prototype Prompted Query Generation (PPQG)

We design a **prototype-audio contrastive loss ($\mathcal{L}_{pac}$)** for audio prototype learning

$$\mathcal{L}_{\text{bce}}(\boldsymbol{M}_k, \boldsymbol{M}_k^*) = -\boldsymbol{M}_k^* \log(\boldsymbol{M}_k) - (1-\boldsymbol{M}_k^*)\log(1-\boldsymbol{M}_k),$$

$$\mathcal{L}_{\text{pac}}(\boldsymbol{M}, \boldsymbol{M}^*) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{L}_{\text{bce}}(\boldsymbol{M}_k, \boldsymbol{M}_k^*).$$



III. **Pixel context grouping:** group pixel context for each query through hard assignment to increase distinguishability

$$\boldsymbol{R} = \text{Softmax}((\bar{\boldsymbol{V}}^e \boldsymbol{W}_2^q)(\boldsymbol{V}^h \boldsymbol{W}_2^k)^{\text{T}} + \boldsymbol{G}),$$

$$\hat{\boldsymbol{R}} = \text{One-hot}(\arg\max_N(\boldsymbol{R})) + \boldsymbol{R} - \text{sg}(\boldsymbol{R}),$$

$$\boldsymbol{V}^q = \bar{\boldsymbol{V}}^e + (\text{Norm}(\hat{\boldsymbol{R}})(\boldsymbol{V}^h \boldsymbol{W}_2^v))\boldsymbol{W}^o,$$

Quantitative Results

Method	Reference	Backbone	Image Size	AVS-Semantic (AVSS)		Single-Source (S4)		Multi-Source (MS3)	
				$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
TPAVI [42, 43]	[ECCV'22]	ResNet-50	224×224	-	-	72.8	84.8	46.9	57.8
		PVT-v2		29.8	35.2	78.7	87.9	54.0	64.5
AQFormer [15]	[IJCAI'23]	ResNet-50	224×224	-	-	77.0	86.4	55.7	66.9
		PVT-v2		-	-	81.6	89.4	61.1	72.1
ECMVAE [28]	[ICCV'23]	ResNet-50	224×224	-	-	76.3	86.5	48.7	60.7
		PVT-v2		-	-	81.7	90.1	57.8	70.8
CATR [20]	[ACMMM'23]	ResNet-50	224×224	-	-	74.8	86.6	52.8	65.3
		PVT-v2		32.8	38.5	81.4	89.6	59.0	70.0
AVSC [22]	[ACMMM'23]	ResNet-50	224×224	-	-	77.0	85.2	49.6	61.5
		PVT-v2		-	-	80.6	88.2	58.2	65.1
BAVS [23]	[TMM'24]	ResNet-50	224×224	24.7	29.6	78.0	85.3	50.2	62.4
		PVT-v2		32.6	36.4	82.0	88.6	58.6	65.5
		Swin-B		33.6	37.5	82.7	89.8	59.6	65.9
COMBO [40]	[CVPR'24]	ResNet-50	224×224	33.3	37.3	81.7	90.1	54.5	66.6
		PVT-v2		42.1	46.1	84.7	91.9	59.2	71.2
CPM [6]	[ECCV'24]	ResNet-50	224×224	34.5	39.6	81.4	90.5	59.8	71.0
TeSO [37]	[ECCV'24]	Swin-B	384×384	39.0	45.1	83.3	93.3	66.0	80.1
SelM [19]	[ACMMM'24]	ResNet-50	224×224	31.9	37.2	76.6	86.2	54.5	65.6
		PVT-v2		41.3	46.9	83.5	91.2	60.3	71.3
AVSBias [33]	[ACMMM'24]	Swin-B	384×384	44.4	49.9	83.3	93.0	67.2	80.8
VCT (Ours)	-	ResNet-50	224×224	**37.5**	**42.2**	**81.8**	**90.6**	**61.9**	**74.7**
		PVT-v2	224×224	**44.7**	**49.5**	**84.8**	**92.1**	**62.0**	**75.0**
		Swin-B	224×224	**47.9**	**52.9**	**84.7**	**92.3**	**67.5**	**79.3**
		Swin-B	384×384	**51.2**	**55.5**	**86.2**	**93.4**	**67.6**	**81.4**

Quantitative Results

Framework	Queries	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
ACT	Audio-Derived Queries	33.2	37.0
VCT	Naive Vision Queries	35.2	39.3
	w/ Cross-Attention	35.8	39.8
	w/ Group-Attention	36.3	40.5
	***w/* Audio Prototypes**	**37.5**	**42.2**

Table 2. Ablation study of generating vision-derived queries in our PPQG module on the AVSS subset.

Method	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
w/o Audio Prototypes	36.3	40.5
w/ Audio Prototypes, *w/o* loss	36.3	40.4
w/ Audio Prototypes, *w/* visual loss	36.5	40.8
***w/* Audio Prototypes, *w/* PAC loss**	**37.5**	**42.2**

Table 3. Ablation study of audio prototype prompting in PPQG.

	Audio Feature	$\mathcal{M}_{\mathcal{J}}$	$\mathcal{M}_{\mathcal{F}}$
Visual Feature	Multiply	33.9	37.9
	Concatenation	35.3	39.3
	Addition	36.3	40.5
Audio Proto	Replace	36.6	40.8
VCT	**Full Model**	**37.5**	**42.2**

Table 4. Ablation study of audio feature incorporation.

Qualitative Results

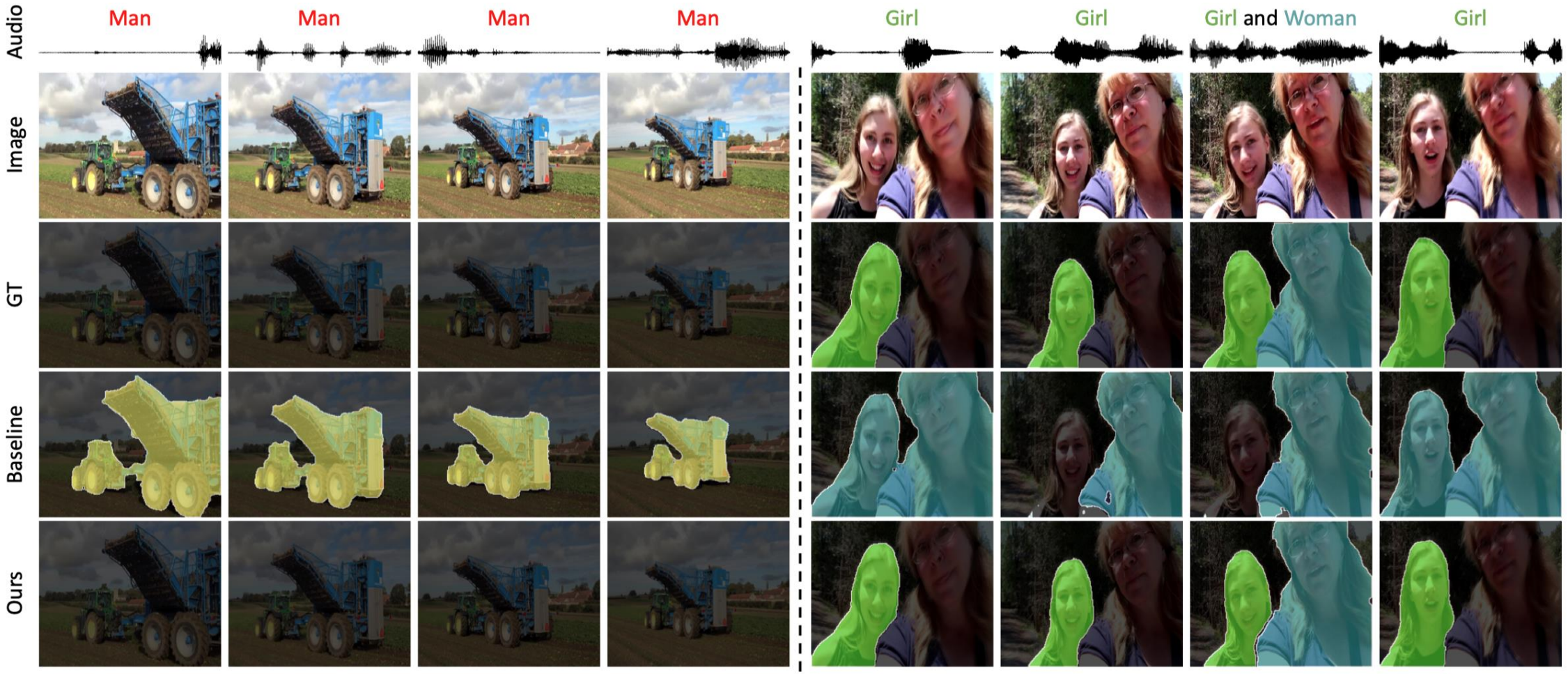


Figure 4. Qualitative comparison between our full model and the ACT baseline. Existing sounds correspond to masks of the same colors.

Qualitative Results

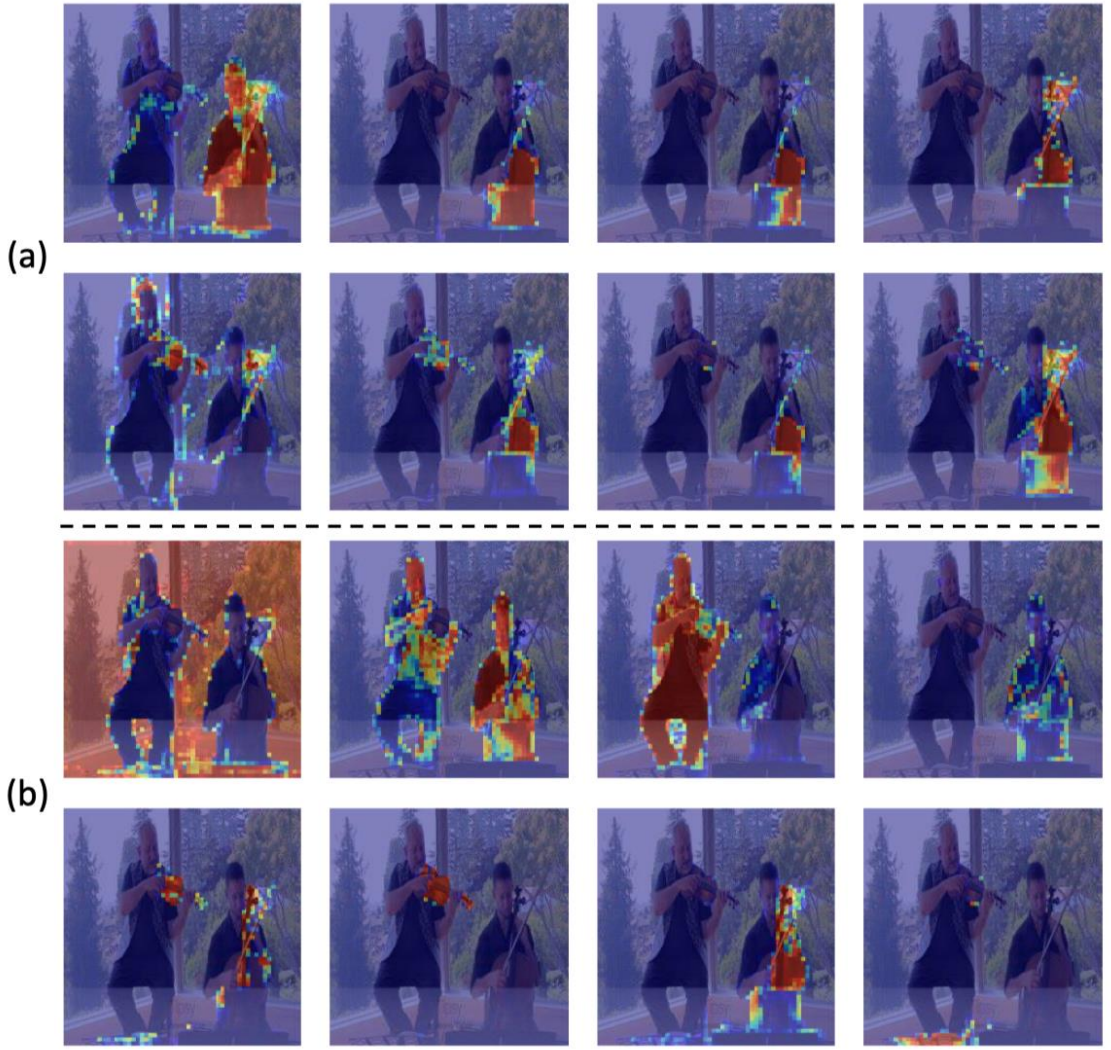


Figure 5. Visualization of logit maps from different types of queries. (a) Audio-derived queries. (b) Vision-derived queries.

CVPR Nashville JUNE 11-15, 2025

Thank You!

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

SANGFOR
深信服科技