CVPR JUNE 3-7, 2026 DENVER COLORADO

SOUPLE: ENHANCING AUDIO-VISUAL LOCALIZATION AND SEGMENTATION WITH LEARNABLE PROMPT CONTEXTS

Khanh Binh Nguyen[1]
Deakin University

Chae Jung Park[2*]
National Cancer Center

SOUPLE: LEARNABLE PROMPT CONTEXTS

BACKGROUND

AUDIO-VISUAL SOUND SOURCE LOCALIZATION

Pinpointing origins of sound in intricate visual scenes remains challenging. It typically relies on aligning audio-visual representations as self-supervision signals in contrastive learning.

CLIP IN AUDIO-VISUAL LEARNING

Large-scale pre-trained models exhibit robust multimodal representations. Extensions like AudioCLIP or Wav2CLIP incorporate audio.

How to effectively ground audio with visual representation without text labels?

PRESENTING AUTHOR
Khanh Binh Nguyen

SOUPLE: LEARNABLE PROMPT CONTEXTS

LIMITATION OF EXISTING WORK

Previous work like ACL-SSL adapts CLIP by replacing the classification token with an audio-embedded token:

"a photo of a [V_a]"

> *This approach fails in some conditions because the fixed prompt "a photo of a" lacks semantic meaning relevant to the audio token [V_a].*

SEMANTIC MISALIGNMENT

PRESENTING AUTHOR
Khanh Binh Nguyen

SOUPLE: LEARNABLE PROMPT CONTEXTS

PROMPT LEARNING FOR CLIP

Instead of using discrete handcrafted text, we propose **SouPLe (Sound-aware Prompt Learning)**, introducing instance-conditional, learnable context tokens.

ACL-SSL

~~"a photo of [V_a]"~~

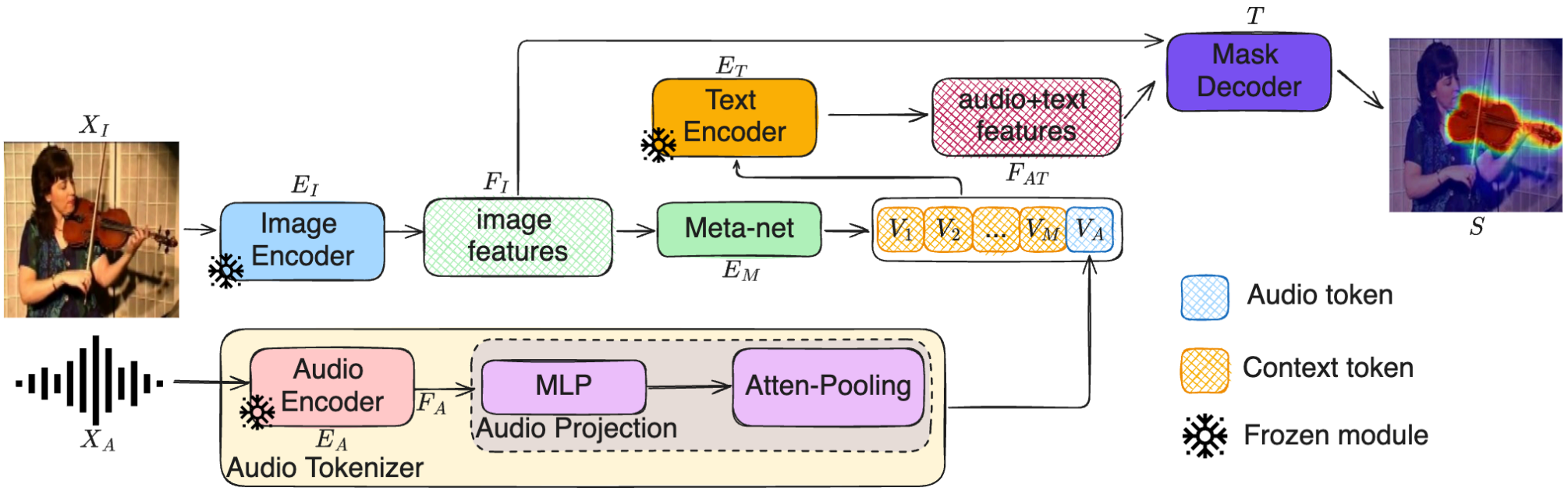
SouPLe

$[V_1][V_2]...[V_m][V_a]$

PRESENTING AUTHOR
Khanh Binh Nguyen

SOUPLE: LEARNABLE PROMPT CONTEXTS

METHODOLOGY: SOUPLE PIPELINE



PRESENTING AUTHOR
Khanh Binh Nguyen

SOUPLE: LEARNABLE PROMPT CONTEXTS

QUANTITATIVE RESULTS: VGG-SS

METHOD	cIoU ↑	AUC ↑
Wav2CLIP	37.71	39.93
AudioCLIP	44.15	46.23
ACL-SSL (Baseline)	**49.46**	**46.32**
SouPLe (Ours)	**53.21** (+3.75)	**48.15** (+1.83)

PRESENTING AUTHOR
Khanh Binh Nguyen

SOUPLE: LEARNABLE PROMPT CONTEXTS

QUANTITATIVE RESULTS: SOUNDNET

METHOD	cIoU ↑	AUC ↑
Wav2CLIP	26.00	29.60
AudioCLIP	47.20	45.22
ACL-SSL (Baseline)	**80.80**	**64.62**
SouPLe (Ours)	**84.80** (+4.00)	**67.64** (+3.02)

PRESENTING AUTHOR
Khanh Binh Nguyen

SOUPLE: LEARNABLE PROMPT CONTEXTS

OPEN-SET EVALUATION

110 HEARD CATEGORIES

HEARD SET

cIoU

54.76

+6.32% vs ACL-SSL

110 UNHEARD CATEGORIES

UNHEARD SET

cIoU

48.40

+6.42% vs ACL-SSL

PRESENTING AUTHOR
Khanh Binh Nguyen

SOUPLE: LEARNABLE PROMPT CONTEXTS

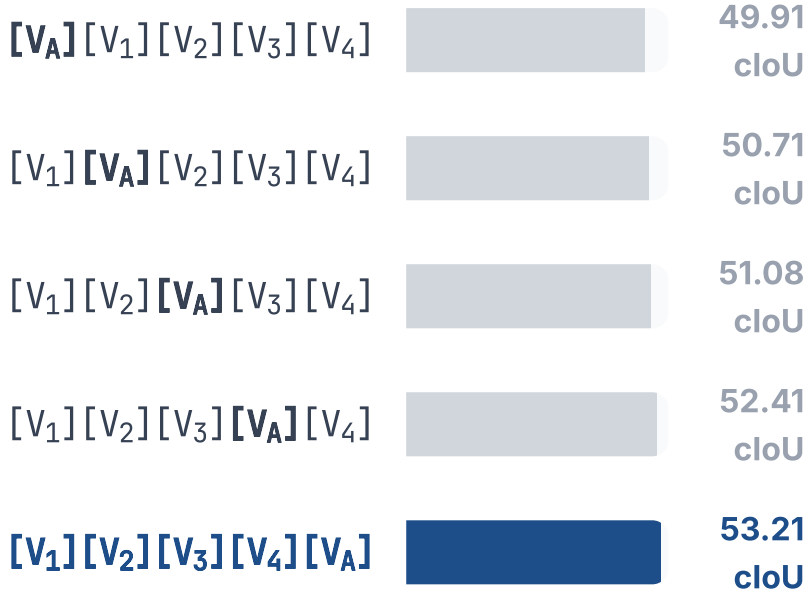
QUALITATIVE COMPARISONS

PRESENTING AUTHOR
Khanh Binh Nguyen

SOUPLE: LEARNABLE PROMPT CONTEXTS

ABLATION STUDIES

TOKEN ORDER



CONTEXT LENGTH

ctx=4	**53.21** cIoU
ctx=8	52.01 cIoU
ctx=16	51.08 cIoU

PRESENTING AUTHOR
Khanh Binh Nguyen

CONCLUSION

We introduced **SouPLe**, utilizing prompt learning to bridge the semantic gap in self-supervised audio-visual localization.

By generating **instance-conditional** learnable context tokens via visual guidance, it gracefully surpasses current state-of-the-art baselines globally.

THANK YOU

Questions?

PRESENTING AUTHOR
Khanh Binh Nguyen