ACL 2026 SAN DIEGO

Generalizable Prompt Tuning for Audio-Language Models via Semantic Expansion

Jaehyuk Jang*, Wonjun Lee*, Kangwook Ko* and Changick Kim

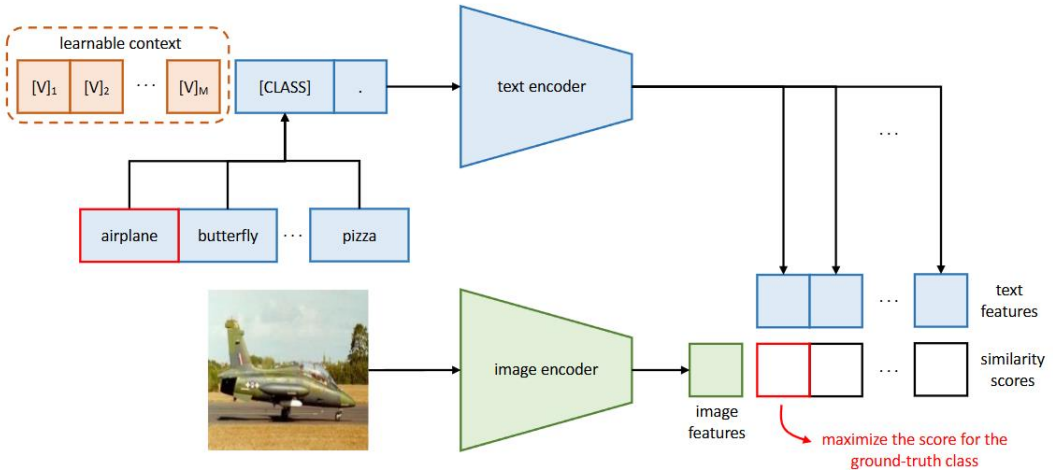
Computational Intelligence Lab., KAIST

CILAB KAIST

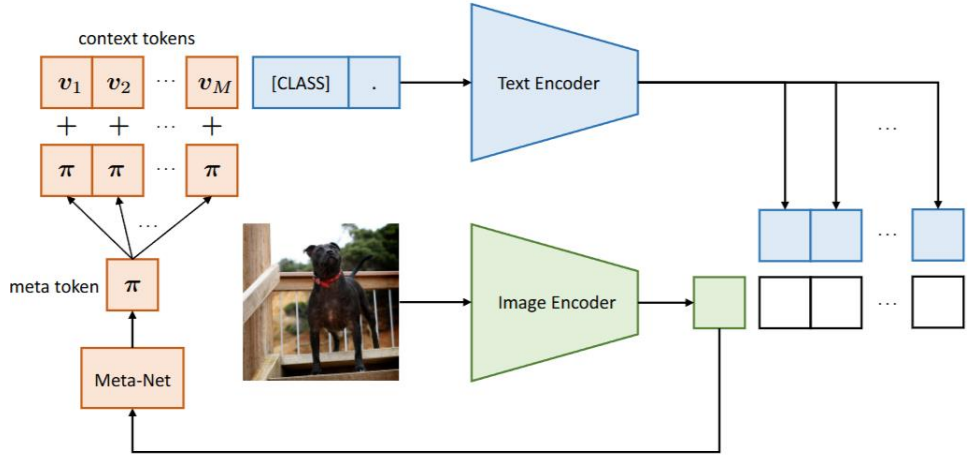
Introduction

Background: Prompt Tuning

- ✓ Prompt tuning
  - ✓ Learns lightweight prompt tokens while keeping the dual encoder frozen
  - ✓ Enables efficient few-shot adaptation without full model fine-tuning
- ✓ Why prompt tuning?
  - ✓ Parameter-efficient and label-efficient adaptation
  - ✓ Preserves pretrained knowledge
  - ✓ Easy to apply to downstream tasks
- ✓ Representative baselines
  - ✓ CoOp learns continuous context tokens
  - ✓ CoCoOp generates input-conditional prompts



CoOp (IJCV '21)



CoCoOp (CVPR '22)

Background: Base-to-New Tradeoff

- ✓ Generalization protocol
  - ✓ Train prompts only on base (seen) classes
  - ✓ Evaluate the same prompts on both base and new (unseen) classes

- ✓ Base-to-New tradeoff (BNT)
  - ✓ Strong adaptation to base classes, poor generalization to unseen classes
  - ✓ A well-known Base-to-New Tradeoff in prompt tuning
  - ✓ Extensively studied in Vision-Language Models, but largely **unexplored in Audio-Language Models**

Key Observation: Semantic Structure Collapse

- ✓ Prompt tuning disrupts semantic alignment
  - ✓ Class embeddings become less aligned with their semantic neighbors
- ✓ Related sound concepts drift apart
  - ✓ Acoustically and semantically related concepts separate in the text space
- ✓ Sparse audio labels amplify the collapse
  - ✓ Audio datasets often contain only tens of class labels
  - ✓ Learned prompts become isolated prototypes rather than robust semantic anchors

Method

Semantic Expansion with LLM Generated Neighbors

- ✓ Main idea
  - ✓ Enrich each base class with LLM-generated semantic neighbors
  - ✓ Use neighbors as dense local support in the text embedding space

- ✓ Semantic neighbors
  - ✓ Generate N related sound concepts, acoustic variants, or paraphrases per class
    - ✓ airplane → aircraft, jet, engine noise
    - ✓ breathing → sighing, exhaling, inhaling

- ✓ Training-time regularization
  - ✓ Encode classes and neighbors with the same learnable prompt
  - ✓ Preserve local semantic geometry while optimizing for audio classification
  - ✓ Remove neighbors at inference time, keeping the original prompt-tuning pipeline

Margin-based Semantic Expansion Loss

- ✓ SEPT: Semantically Expanded Prompt Tuning
  - ✓ Intra-class Alignment
    - ✓ Pull each class embedding toward its own neighbors
  - ✓ Inter-class Separation
    - ✓ Push each class embedding away from other classes' neighbors
  - ✓ Margin Constraint
    - ✓ Margins are computed from the original pretrained text space
    - ✓ Prevent over-compression of positives and over-separation of negatives

$$\mathcal{L}_{\text{intra}}^{(i)} = \frac{1}{N}\sum_{n=1}^{N} \max\left(0,\ \|\mathbf{z}_i - \mathbf{p}_i^n\|_2 - m_{i,i,n}\right)$$

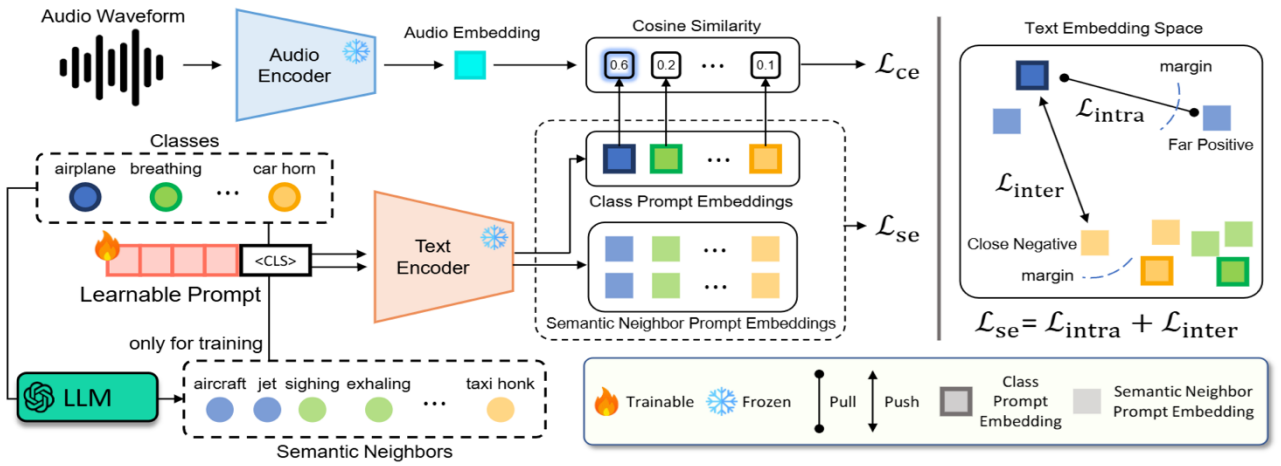
$$\mathcal{L}_{\text{inter}}^{(i,j)} = \frac{1}{N}\sum_{n=1}^{N} \max\left(0,\ m_{i,j,n} - \|\mathbf{z}_i - \mathbf{p}_j^n\|_2\right)$$

- ✓ Semantic Expansion Loss

$$\mathcal{L}_{\text{se}} = \frac{1}{K}\sum_{i=1}^{K}\left(\mathcal{L}_{\text{intra}}^{(i)} + \frac{1}{K-1}\sum_{j=1,j\neq i}^{K}\mathcal{L}_{\text{inter}}^{(i,j)}\right)$$

- ✓ Final Objective

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ce}} + \lambda \cdot \mathcal{L}_{\text{se}}$$



Overview of SEPT

Experiments

Setup

- ✓ Datasets
  - ✓ 11 audio classification benchmarks across sound events, instruments, emotion, scenes, music, and vocal sounds
  - ✓ Beijing-Opera NS-Instruments, ESC50, ESC50-Actions, UrbanSound8k, CREMA-D, RAVDESS, SESA, GT-Music-Genre, VocalSound, TUT2017

- ✓ Model and Baselines
  - ✓ Pengi Audio-Language Encoders
  - ✓ CoOp, CoCoOp, KgCoOp, DePT

- ✓ Evaluation protocols
  - ✓ Base-to-new generalization: train prompts on base classes, evaluate on base and unseen new classes
  - ✓ Cross-dataset evaluation: train prompts on a source dataset and directly test on a different target dataset

- ✓ Few-shot setting
  - ✓ 16-shot per base class

Results: Base-to-New Generalization

- ✓ Generalization under category shift
  - ✓ Train prompts on base classes and evaluate them on both seen base classes and unseen new classes within the same dataset
- ✓ Consistent gains across baselines
  - ✓ Average harmonic mean over 11 datasets
    - ✓ CoOp: 42.83 → 49.70
    - ✓ CoCoOp: 46.26 → 50.65
    - ✓ KgCoOp: 36.39 → 49.79
    - ✓ DePT: 46.79 → 49.06
- ✓ Better new-class generalization
  - ✓ Conventional prompt tuning overfits base classes
  - ✓ SEPT improves new-class accuracy while keeping competitive base performance

Method	Avg. over 11 datasets			Beijing-Opera			NS-Instruments			ESC50		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
Pengi	40.68	38.21	38.46	64.16	68.55	66.15	43.32	30.01	35.46	15.70	13.70	14.33
CoOp	**65.00**	34.09	42.83	97.27	61.38	74.87	41.11	23.88	30.08	**61.97**	11.83	19.46
+ SEPT	64.36	**42.98**	**49.70**	**97.88**	**71.87**	**82.45**	**43.33**	**37.78**	**40.29**	59.00	**15.80**	**24.70**
CoCoOp	**69.13**	36.83	46.26	**97.86**	70.80	**81.84**	47.42	**37.66**	**41.84**	**70.13**	13.63	22.78
+ SEPT	68.63	**42.59**	**50.65**	97.85	**71.06**	81.60	**52.15**	29.96	37.95	69.10	**17.33**	**27.62**
KgCoOp	37.99	37.42	36.39	67.26	61.40	62.96	39.80	39.27	39.11	14.10	10.33	11.76
+ SEPT	**58.92**	**45.28**	**49.79**	**94.44**	**67.99**	**78.28**	**51.24**	**40.86**	**45.28**	**47.60**	**18.33**	**26.33**
DePT	63.86	39.91	46.79	**97.26**	62.25	**75.32**	43.75	28.21	34.01	**60.80**	14.00	22.53
+ SEPT	**64.57**	**41.63**	**49.06**	96.32	**62.74**	75.00	**47.66**	**37.90**	**41.60**	56.80	**16.03**	**24.91**

Method	ESC50-Actions			UrbanSound8k			CREMA-D			RAVDESS		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
Pengi	25.50	28.00	26.44	30.68	39.51	34.14	52.25	32.36	39.96	25.38	32.16	28.37
CoOp	**85.33**	49.33	61.97	**61.86**	31.22	40.94	52.50	9.01	15.29	51.01	27.75	35.88
+ SEPT	82.50	**52.67**	**63.90**	61.22	**36.34**	**44.95**	**52.75**	**42.06**	**44.17**	**54.55**	**33.92**	**41.39**
CoCoOp	**89.17**	41.50	56.15	**64.25**	**30.77**	**41.05**	**56.80**	**26.07**	**35.62**	56.82	26.87	36.39
+ SEPT	87.50	**51.50**	**64.58**	61.06	25.93	36.26	54.06	19.72	28.21	**63.55**	**37.57**	**46.77**
KgCoOp	36.50	30.67	32.88	26.14	26.15	25.40	42.70	55.43	42.48	27.53	27.46	27.42
+ SEPT	**67.83**	**51.50**	**57.97**	**56.96**	**38.77**	**45.32**	**53.06**	**69.53**	**59.92**	**45.58**	**32.31**	**37.79**
DePT	**81.33**	44.67	57.12	**63.21**	**37.28**	**46.05**	**51.56**	33.19	33.54	52.65	**32.01**	**39.28**
+ SEPT	79.67	**52.67**	**63.02**	61.11	36.22	44.79	51.44	**35.60**	**38.93**	**52.90**	29.07	37.49

Method	SESA			GT-Music-Genre			VocalSound			TUT2017		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
Pengi	75.00	89.19	81.48	35.64	15.15	21.26	59.28	53.01	55.97	20.52	18.65	19.53
CoOp	**85.78**	74.77	79.33	**57.76**	20.54	29.76	**75.26**	45.01	56.00	45.11	20.25	27.53
+ SEPT	78.92	**83.78**	**81.22**	56.11	**23.23**	**32.15**	74.11	**54.14**	**62.38**	**47.55**	**21.14**	**29.05**
CoCoOp	85.78	72.97	78.68	**60.73**	24.24	32.68	**82.47**	40.03	53.19	49.01	20.55	28.61
+ SEPT	**88.73**	**86.49**	**87.42**	49.80	**46.32**	**47.99**	79.57	**61.06**	**68.84**	**51.60**	**21.54**	**29.94**
KgCoOp	68.14	79.28	72.37	27.39	15.82	19.77	52.50	49.91	51.10	15.83	15.90	15.10
+ SEPT	**79.90**	**82.88**	**81.23**	**42.90**	**20.88**	**28.08**	**70.12**	**56.05**	**62.23**	**38.48**	**18.94**	**25.30**
DePT	80.39	**91.89**	**85.72**	50.17	22.90	31.08	72.24	49.00	58.28	**49.12**	**23.67**	**31.75**
+ SEPT	**87.25**	82.88	84.90	**55.78**	**23.91**	**33.13**	**75.34**	**57.70**	**65.24**	46.05	23.25	30.70

Results: Cross-dataset Transfer

- ✓ Generalization beyond category shift
  - ✓ Train prompts on a source dataset and test on a different target dataset
  - ✓ Measures robustness under dataset-level distribution shift

- ✓ SEPT consistently improves transfer accuracy
  - ✓ Pluggd into CoOp:
    - ✓ ESC50-Actions → UrbanSound8K: 19.02 → 24.21
    - ✓ RAVDESS → CREMA-D: 15.74 → 23.89
    - ✓ NS-Instruments → Beijing-Opera: 31.92 → 41.66

Method	Sound Event Classif.		Emotion Recog.		Instrument Classif.	
	ESC50-A. (Source)	UrbanS. (Target)	RAV. (Source)	CREM. (Target)	NS-Inst. (Source)	Beijing. (Target)
CoOp	**70.08**	19.02	30.68	15.74	35.30	31.92
+ SEPT	69.42	**24.21**	**31.70**	**23.89**	**37.21**	**41.66**
CoCoOp	77.83	16.43	**33.81**	8.98	37.45	36.44
+ SEPT	**78.17**	**23.18**	31.50	**19.14**	**39.63**	**37.99**
KgCoOp	15.92	13.34	14.19	14.08	16.55	29.96
+ SEPT	**54.50**	**21.84**	**25.05**	**20.60**	**39.16**	**50.01**
DePT	**70.83**	21.86	**30.96**	17.53	**36.89**	37.72
+ SEPT	68.08	**24.46**	27.83	**27.07**	36.30	**44.49**

- ✓ Component ablations
  - ✓ Naive pull-push regularization without margins can distort the embedding space
  - ✓ Margin constraints stabilize semantic expansion and produce the best harmonic mean
- ✓ Compute efficiency
  - ✓ No additional trainable parameters over CoOp
  - ✓ No additional inference latency; neighbors are used only during training
- ✓ Qualitative structure
  - ✓ t-SNE shows tighter clusters between new classes and their semantic neighbors
  - ✓ Semantic structure learned from base classes transfers to unseen classes

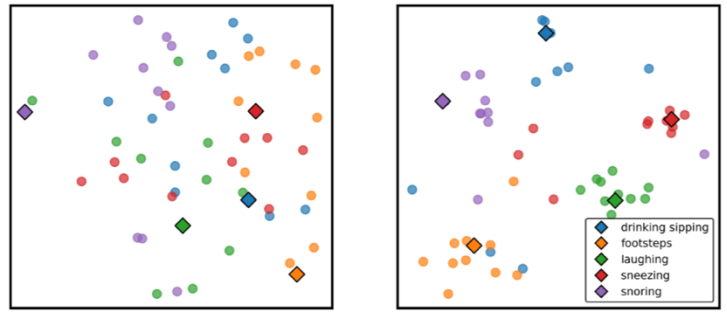
Ablation study

	$\mathcal{L}_{\text{intra}}$	m^{intra}	$\mathcal{L}_{\text{inter}}$	m^{inter}	Base	New	H
1					65.00	34.09	42.83
2	✓				58.31	33.58	41.17
3	✓	✓			64.99	36.90	44.52
4			✓		62.01	36.68	43.48
5			✓	✓	62.84	38.79	46.04
6	✓		✓		61.82	32.89	41.47
7	✓	✓	✓	✓	64.37	42.93	**49.70**

Efficiency

	# Params	Training Time (s)	Inference Time (ms)	H
CoCoOp	107,072	681.5	40.9	46.26
KgCoOp	8,192	96.0	2.8	36.39
DePT	114,738	108.1	2.8	46.79
CoOp	8,192	94.0	2.8	42.83
+ SEPT	8,192	170.7	2.8	**49.70**

t-SNE visualization



CoOp w/o SEPT CoOp w/ SEPT

Conclusion

Takeaway

- ✓ Preserve semantic structure, not only base-class accuracy
  - ✓ Prompt tuning in ALMs suffers from BNT because learned prompts disrupt semantic neighborhoods
- ✓ Semantic neighbors provide dense local support
  - ✓ LLM-generated neighbors compensate for sparse audio label spaces and maintain text-space geometry
- ✓ Margin-based expansion enables stable regularization
  - ✓ Intra-class alignment and inter-class separation improve generalization while preserving pretrained relations
- ✓ Simple, effective, and plug-and-play
  - ✓ SEPT improves base-to-new generalization and cross-dataset transfer
  - ✓ No additional inference overhead

Thank you!