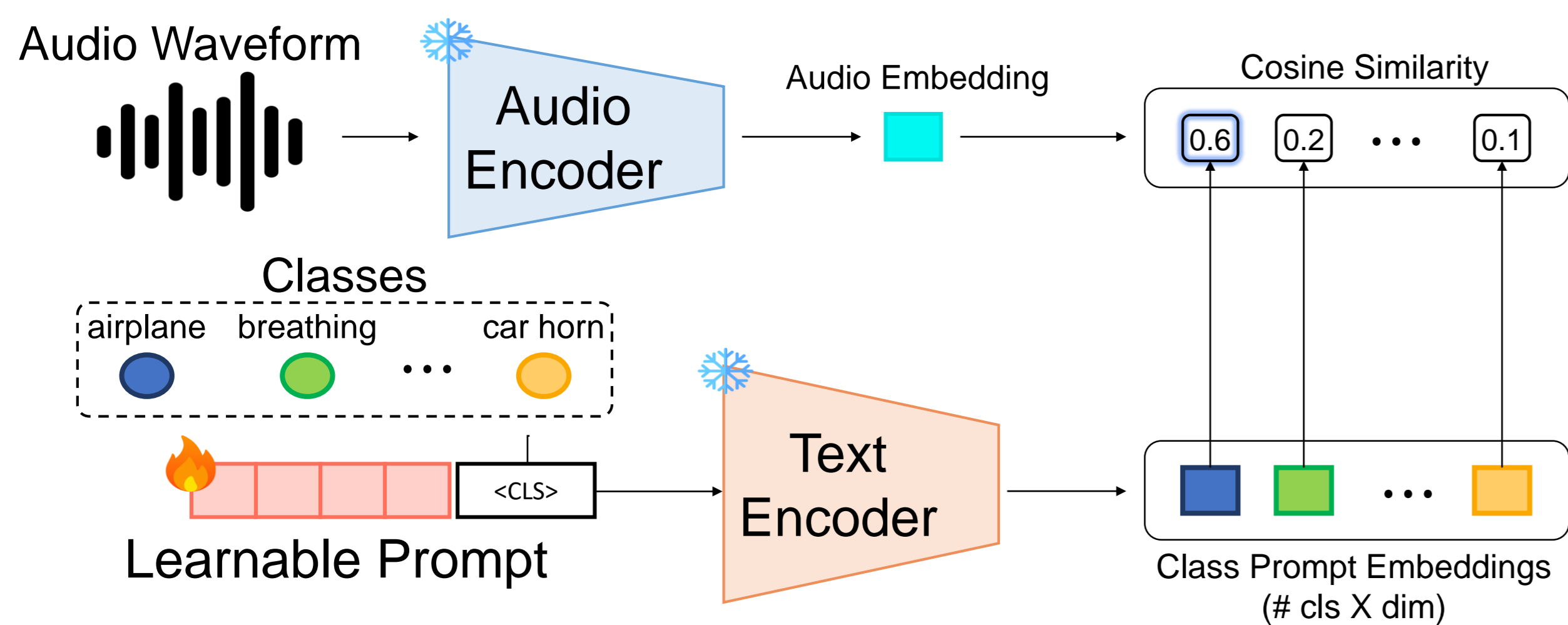


# Generalizable Prompt Tuning for Audio-Language Models via Semantic Expansion

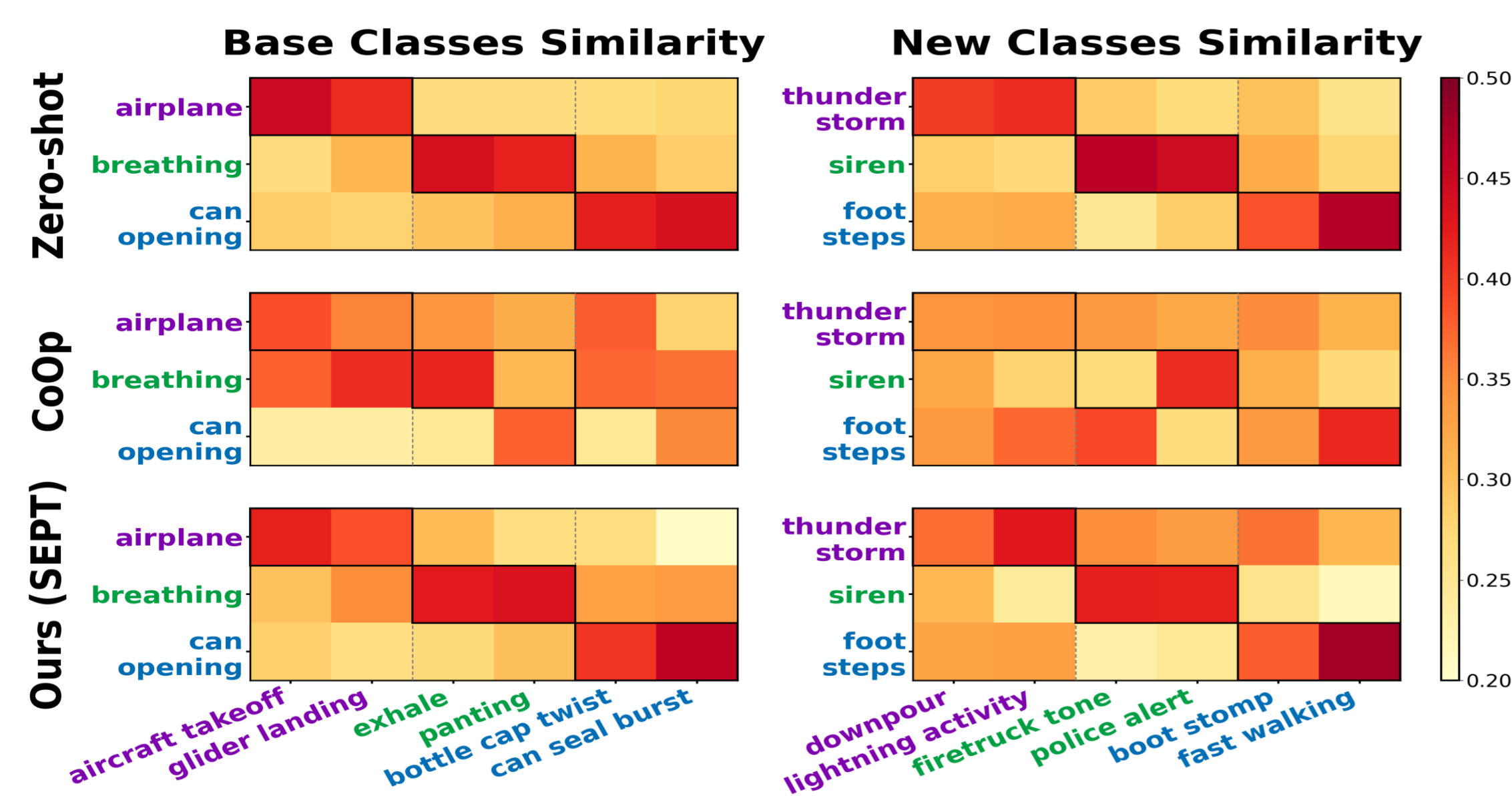
Jaehyuk Jang\* Wonjun Lee\* Kangwook Ko\* Changick Kim

Computational Intelligence Lab., KAIST

## Introduction



### Prompt Tuning in ALMs



### Observation

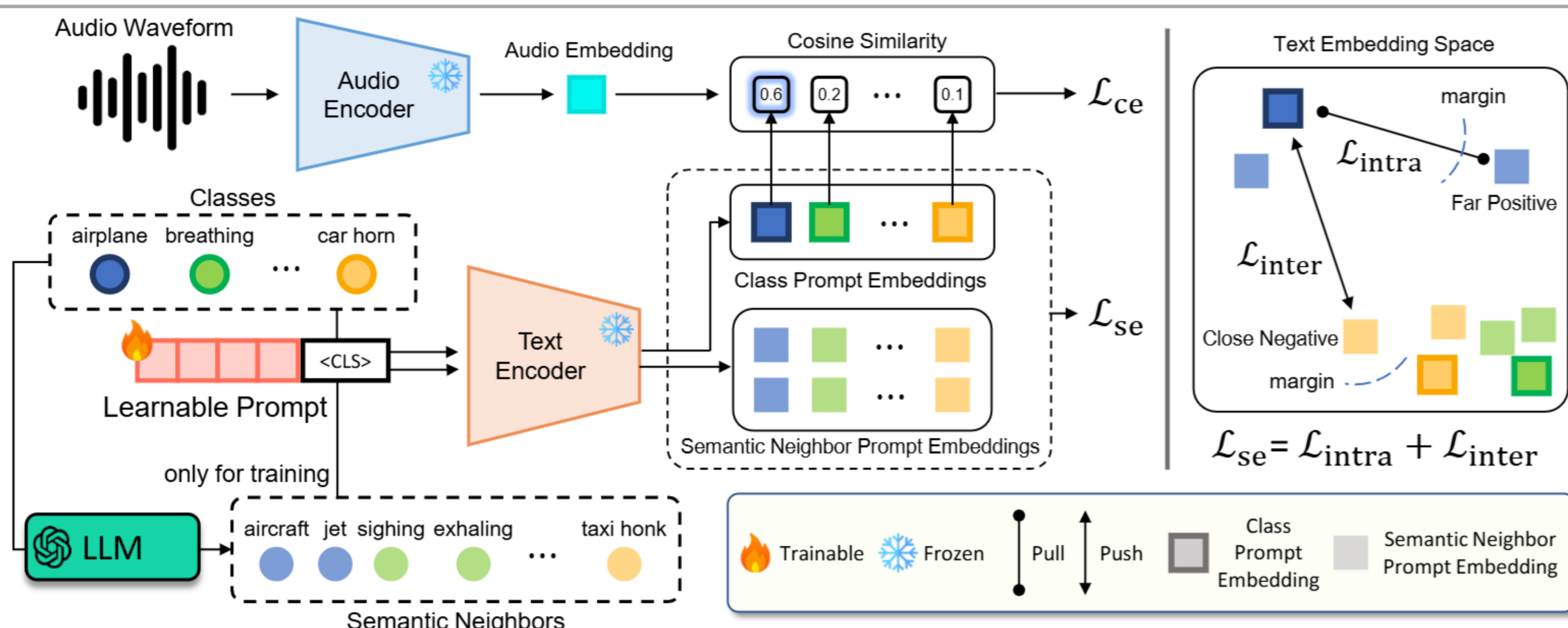
## Prompt Tuning for Audio-Language Models (ALMs)

- ✓ Efficient few-shot adaptation of frozen ALMs by learning lightweight prompt tokens instead of full model fine-tuning
- ✓ However, prompt tuning improves base (seen) class performance but often fails to generalize to new (unseen) classes, known as the Base-to-New Tradeoff (BNT)

## Key Observation

- ✓ Learned prompts disrupt semantic alignment between classes and their neighbors
  - Acoustically and semantically related sound concepts drift apart in the text embedding space
- ✓ Sparse audio label spaces provide weak semantic support, causing learned prompts to become isolated prototypes rather than robust semantic anchors

## Proposed Method



### Overview of proposed SEPT

#### Semantic Neighbors

- For each base class, generate  $N$  semantically related sound concepts using an LLM
- Neighbors provide dense local support around each class in text embedding space

#### Semantic Expansion Loss

- ✓ Intra-class Alignment
  - Pull class embeddings toward their own semantic neighbors
- ✓ Inter-class Separation
  - Push class embeddings away from other classes' neighbors
- ✓ Margin Constraint
  - Margins from the pretrained text space preserve the original semantic geometry
  - Avoids over-compressing positives and over-separating negatives

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

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

- ✓ Training Objective

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

#### Key Advantage

- ✓ **Plug-and-play:** compatible with existing prompt tuning
- ✓ **Model-agnostic:** can be applied without modifying the frozen encoders
- ✓ **Efficient:** improves generalization without increasing inference cost

## Experiments

### Base-to-new Generalization

Method	Avg. over 11 datasets			Beijing-Opera			NS-Instruments			ESCS50		
	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	<b>65.00</b>	34.09	42.83	97.27	61.38	74.87	41.11	23.88	30.08	<b>61.97</b>	11.83	19.46
+ SEPT	64.36	<b>42.98</b>	<b>49.70</b>	<b>97.88</b>	<b>71.87</b>	<b>82.45</b>	<b>43.33</b>	<b>37.78</b>	<b>40.29</b>	<b>59.00</b>	<b>15.80</b>	<b>24.70</b>
CoCoOp	<b>69.13</b>	36.83	46.26	<b>97.86</b>	70.80	<b>81.84</b>	47.42	<b>37.66</b>	<b>41.84</b>	<b>70.13</b>	13.63	22.78
+ SEPT	68.63	<b>42.59</b>	<b>50.65</b>	97.85	<b>71.06</b>	81.60	<b>52.15</b>	29.96	37.95	69.10	<b>17.33</b>	<b>27.62</b>
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	<b>58.92</b>	<b>45.28</b>	<b>49.79</b>	<b>94.44</b>	<b>67.99</b>	<b>78.28</b>	<b>51.24</b>	<b>40.86</b>	<b>45.28</b>	<b>47.60</b>	<b>18.33</b>	<b>26.33</b>
DePT	63.86	39.91	46.79	<b>97.26</b>	62.25	<b>75.32</b>	43.75	28.21	34.01	<b>60.80</b>	14.00	22.53
+ SEPT	<b>64.57</b>	<b>41.63</b>	<b>49.06</b>	96.32	<b>62.74</b>	75.00	<b>47.66</b>	<b>37.90</b>	<b>41.60</b>	56.80	<b>16.03</b>	<b>24.91</b>

Method	ESCS50-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	<b>85.33</b>	49.33	61.97	<b>61.86</b>	31.22	40.94	52.50	9.01	15.29	51.01	27.75	35.88
+ SEPT	82.50	<b>52.67</b>	<b>63.90</b>	61.22	<b>36.34</b>	<b>44.95</b>	<b>52.75</b>	<b>42.06</b>	<b>44.17</b>	<b>54.55</b>	<b>33.92</b>	<b>41.39</b>
CoCoOp	<b>89.17</b>	41.50	56.15	<b>64.25</b>	<b>30.77</b>	<b>41.05</b>	<b>56.80</b>	<b>26.07</b>	<b>35.62</b>	56.82	26.87	36.39
+ SEPT	87.50	<b>51.50</b>	<b>64.58</b>	61.06	25.93	36.26	54.06	19.72	28.21	<b>63.55</b>	<b>37.57</b>	<b>46.77</b>
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	<b>67.83</b>	<b>51.50</b>	<b>57.97</b>	<b>56.96</b>	<b>38.77</b>	<b>45.32</b>	<b>53.06</b>	<b>69.53</b>	<b>59.92</b>	<b>45.58</b>	<b>32.31</b>	<b>37.79</b>
DePT	<b>81.33</b>	44.67	57.12	<b>63.21</b>	<b>37.28</b>	<b>46.05</b>	<b>51.56</b>	33.19	33.54	52.65	<b>32.01</b>	<b>39.28</b>
+ SEPT	79.67	<b>52.67</b>	<b>63.02</b>	61.11	36.22	44.79	51.44	<b>35.60</b>	<b>38.93</b>	<b>52.90</b>	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	<b>85.78</b>	74.77	79.33	<b>57.76</b>	20.54	29.76	<b>75.26</b>	45.01	56.00	45.11	20.25	27.53
+ SEPT	78.92	<b>83.78</b>	<b>81.22</b>	56.11	<b>23.23</b>	<b>32.15</b>	74.11	<b>54.14</b>	<b>62.38</b>	<b>47.55</b>	<b>21.14</b>	<b>29.05</b>
CoCoOp	85.78	72.97	78.68	<b>60.73</b>	24.24	32.68	<b>82.47</b>	40.03	53.19	40.01	20.55	28.61
+ SEPT	<b>88.73</b>	<b>86.49</b>	<b>87.42</b>	49.80	<b>46.32</b>	<b>47.99</b>	79.57	<b>61.06</b>	<b>68.84</b>	<b>51.60</b>	<b>21.54</b>	<b>29.94</b>
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	<b>79.90</b>	<b>82.88</b>	<b>81.23</b>	<b>42.90</b>	<b>20.88</b>	<b>28.08</b>	<b>70.12</b>	<b>56.05</b>	<b>62.23</b>	<b>38.48</b>	<b>18.94</b>	<b>25.30</b>
DePT	80.39	<b>91.89</b>	<b>85.72</b>	50.17	22.90	31.08	72.24	49.00	58.28	<b>49.12</b>	<b>23.67</b>	<b>31.75</b>
+ SEPT	<b>87.25</b>	82.88	84.90	55.78	<b>23.91</b>	<b>33.13</b>	<b>75.34</b>	<b>57.70</b>	<b>65.24</b>	46.05	23.25	30.70

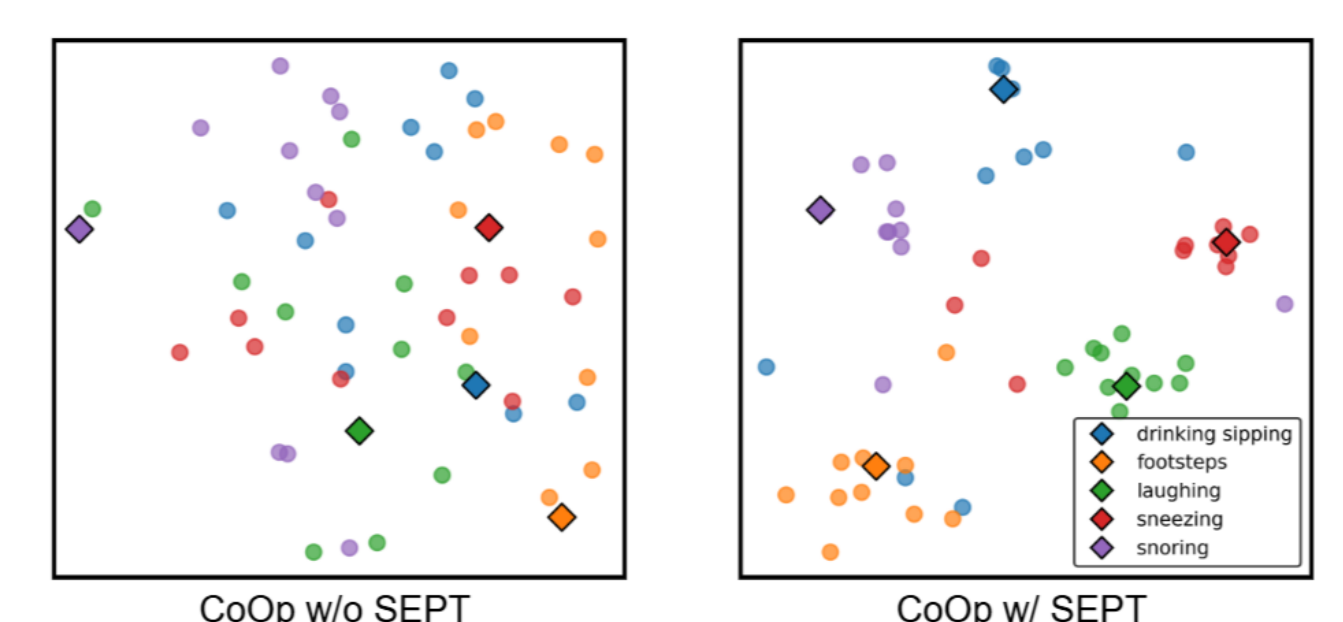
### Cross-dataset Evaluation

Method	Sound Event Classif.		Emotion Recog.		Instrument Classif.	
	ESCS50-A. (Source)	UrbanS. (Target)	RAV. (Source)	CREM. (Target)	NS-Inst. (Source)	Beijing. (Target)
CoOp	<b>70.08</b>	19.02	30.68	15.74	35.30	31.92
+ SEPT	69.42	<b>24.21</b>	<b>31.70</b>	<b>23.89</b>	<b>37.21</b>	<b>41.66</b>
CoCoOp	77.83	16.43	33.81	8.98	37.45	36.44
+ SEPT	<b>78.17</b>	<b>23.18</b>	<b>31.50</b>	<b>19.14</b>	<b>39.63</b>	<b>37.99</b>
KgCoOp	15.92	13.34	14.19	14.08	16.55	29.96
+ SEPT	<b>54.50</b>	<b>21.84</b>	<b>25.05</b>	<b>20.60</b>	<b>39.16</b>	<b>50.01</b>
DePT	70.83	21.86	30.96	17.53	36.89	37.72
+ SEPT	68.08	<b>24.46</b>	<b>27.83</b>	<b>27.07</b>	<b>36.30</b>	<b>44.49</b>

### Ablation study

	$\mathcal{L}_{intra}$	$m^{intra}$	$\mathcal{L}_{inter}$	$m^{inter}$	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	<b>49.70</b>

### t-SNE visualization



### 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	<b>49.70</b>