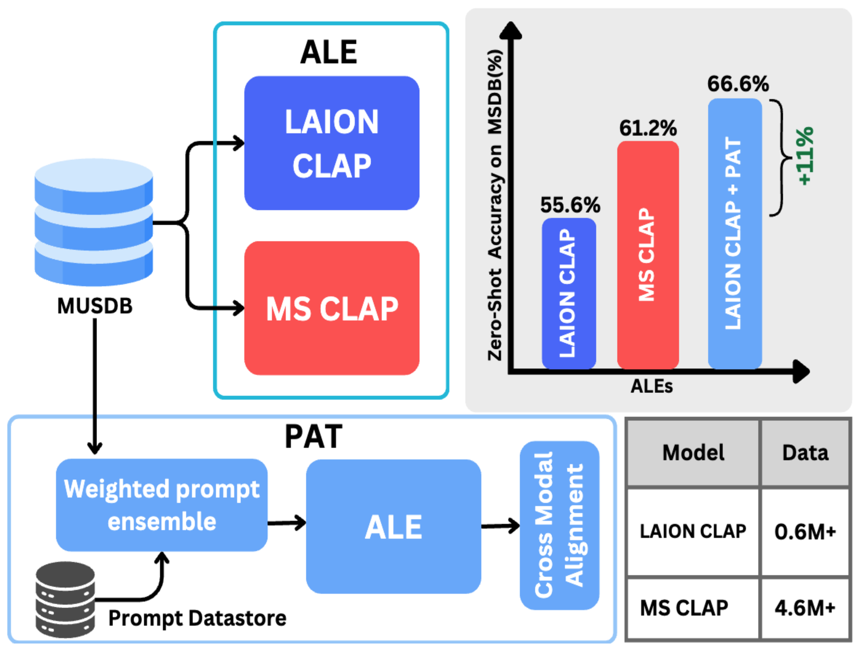
PAT:Parameter-Free Audio-Text Aligner to Boost Zero-Shot Audio Classification

Ashish Seth, Ramaneswaran Selvakumar, Sonal Kumar, Sreyan Ghosh, Dinesh Manocha
University of Maryland, College Park, USA

INTRODUCTION AND MOTIVATION

- Significant progress has been made to enhance zero-shot performances of prior Audio-Language Encoders (ALEs) for audio-classification tasks
- While these models show zero-shot improvement, it comes at the additional cost of pre-training them with either more refined learning objectives or volume of training data



MAIN CONTRIBUTION

Our main contribution are as follows

- We propose **PAT (P**arameter free **A**udio-**T**ext aligner**)**, a novel approach to improve zero-shot audio classification performance in a *training-free* fashion. With **PAT,** we introduce a cross-modal interaction approach aimed at improving audio-text alignment by enhancing both audio and textual representations
- We evaluate **PAT** across multiple ALEs on 18 audio classification datasets and show that **PAT** achieves **0.42%-27.0%** improvement
- We further investigate **PAT** 's robustness to *noisy audio* to show that **PAT** consistently outperforms our baselines under varied noise augmentation settings.

ARCHITECTURE

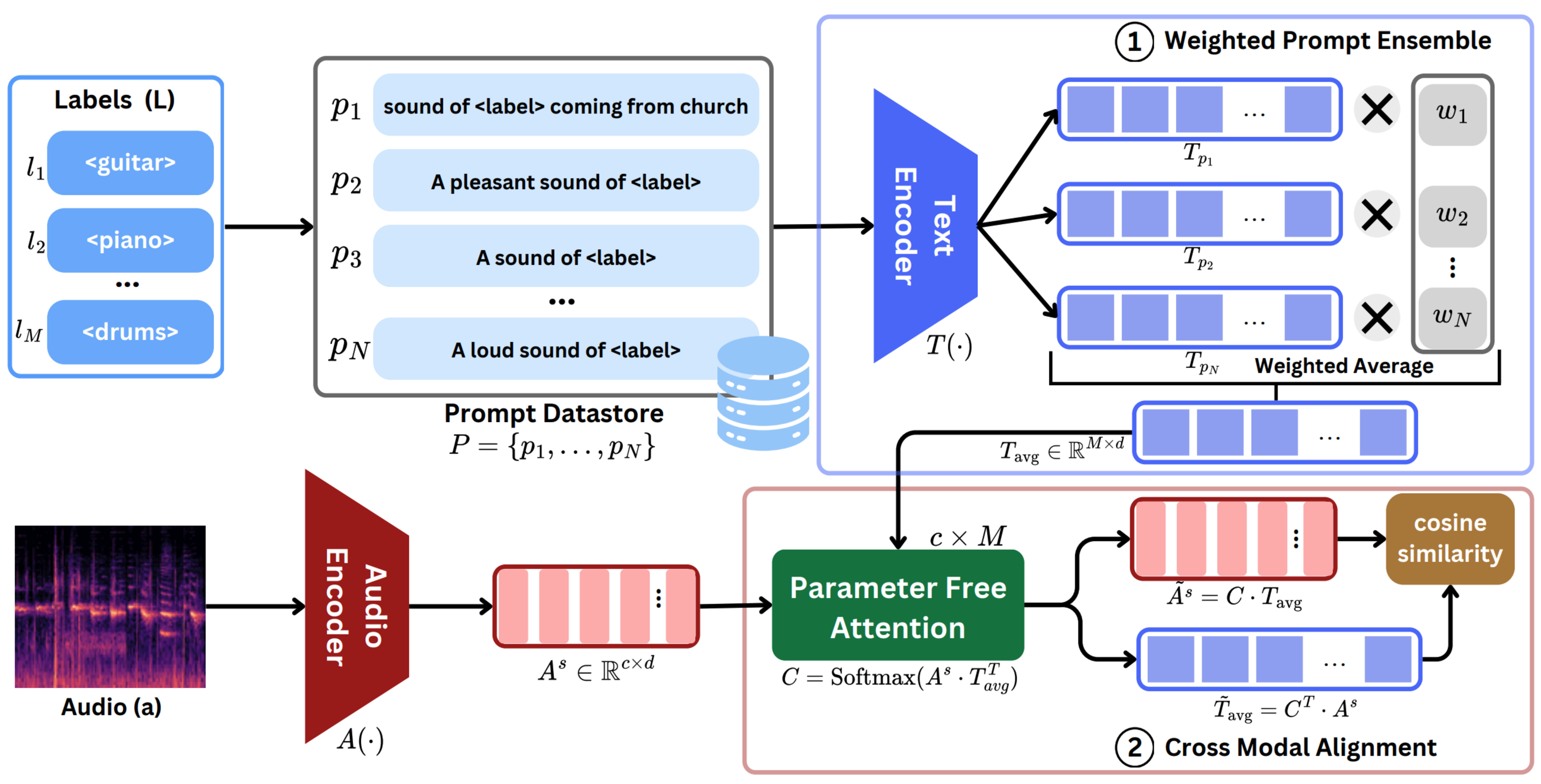


Illustration of **PAT**. **PAT** improves the zero-shot capabilities of ALEs by enriching audio-text representations in a training-free fashion.

- We utilizes an in-house generic prompt datastore to transform class labels into diverse textual descriptions, which are then encoded by a text encoder.
- Further, each prompt is assigned a unique score based on the level of uncertainty it introduces during zero-shot prediction (less uncertainty results in a higher score).
- A weighted average is then performed to generate task-specific, semantically rich textual representations.
- Next, the enriched textual representations are used to guide the enhancement of audio representations using a novel zero-shot **cross model alignment.**
- Precisely, frame-level audio representations are paired with enhanced textual representations to compute a parameter-free attention map, which is used in performing audio and text-guided transformations.

QUANTITATIVE RESULTS

Model → / Dataset ↓	L-CLAP ZS	L-CLAP PAT	LM-CLAP ZS	LM-CLAP PAT	MSCLAP-22 ZS	MSCLAP-22 PAT	MSCLAP-23 ZS	MSCLAP-23 PAT	Wav2CLIP ZS	Wav2CLIP PAT	CompA ZS	CompA PAT
Sound												
ESC-50	89.00	**93.00** +4.00%	85.60	**92.65** +7.05%	76.95	**78.35** +1.40%	91.80	**94.80** +3.00%	24.85	**31.60** +6.08%	91.35	**93.20** +1.85%
USD-8K	76.00	**80.00** +4.00%	28.09	**39.93** +11.84%	72.54	**74.80** +2.26%	77.70	**82.50** +4.80%	20.97	**22.69** +1.72%	73.53	**78.32** +4.79%
TUT	36.00	**39.00** +3.00%	28.09	**39.93** +11.84%	24.44	**25.61** +1.17%	45.00	**47.00** +2.00%	11.54	**15.18** +3.64%	40.12	**46.28** +6.16%
VS	78.20	**80.00** +1.80%	74.46	78.91 +4.45%	43.78	**54.94** +11.16%	79.00	**79.60** +0.60%	22.72	**24.06** +1.31%	65.22	**71.26** +6.04%
DCASE	44.88	**50.81** +5.93%	**56.76**	55.94 -0.82%	13.93	**23.77** +9.84%	45.90	**45.96** +0.06%	09.63	**17.21** +7.58%	33.20	**34.29** +1.09%
Gunshot Tri	10.23	**22.72** +12.49%	13.64	**29.52** +15.88%	17.05	**23.86** +6.81%	25.00	**25.00** +0.00%	**25.00**	25.00 +0.00%	25.00	**26.15** +1.15%
SESA	67.72	**74.28** +6.56%	72.38	**79.04** +6.66%	66.67	**68.47** +1.80%	70.48	**71.61** +1.13%	29.52	**56.10** +26.58%	64.76	**69.42** +4.66%
AudioSet	31.88	**36.98** +5.10%	33.12	38.21 +5.09%	16.10	**17.81** +1.71%	25.33	**28.73** +3.40%	18.03	**20.12** +2.09%	33.24	**35.12** +1.88%
FSD50K	46.45	**48.76** +2.31%	47.12	**49.10** +2.08%	32.50	**33.80** +1.30%	44.49	**45.52** +1.02%	42.31	**44.14** +2.07%	42.18	**43.22** +1.04%
Cochlscene	38.56	**48.66** +10.10%	50.66	55.35 +4.69%	25.94	**33.51** +7.57%	85.00	**85.22** +0.22%	13.09	**16.11** +3.02%	31.95	**38.21** +6.26%
Music												
Beijing Op.	45.34	**68.64** +23.30%	75.00	**75.42** +0.42%	54.24	**73.72** +19.48%	71.19	**71.61** +0.42%	26.69	**34.32** +7.63%	61.86	**63.21** +1.35%
GTZAN	43.40	**54.20** +10.80%	63.92	**63.93** +0.01%	19.19	**20.75** +1.56%	56.24	**58.56** +2.32%	**30.00**	27.76 -2.24%	50.22	**52.17** +1.95%
MUSDB	55.60	**66.00** +10.40%	73.20	**73.20** +0.00%	47.20	**47.75** +0.55%	61.20	**62.40** +1.20%	51.60	**52.20** +0.60%	56.80	**59.55** +2.75%
Medley	82.50	**92.00** +9.50%	87.88	94.30 +6.42%	84.41	**86.20** +1.79%	45.00	**47.00** +2.00%	42.20	**47.08** +4.88%	56.27	**57.24** +0.97%
Mri. St	10.81	**37.35** +26.54%	47.40	**47.80** +0.40%	14.50	**14.80** +0.30%	44.09	**47.12** +3.03%	06.09	**19.49** +13.40%	06.25	**07.42** +1.17%
Mri. To	25.10	**34.38** +9.28%	27.59	**31.62** +4.03%	16.50	**16.63** +0.13%	22.02	**26.18** +4.16%	15.57	**24.95** +9.38%	17.43	**18.79** +1.36%
NSynth Inst	37.20	**38.00** +0.80%	31.67	**36.49** +4.82%	26.26	**29.63** +3.37%	63.30	**66.30** +3.00%	**24.39**	21.72 -2.67%	27.86	**29.24** +1.38%
NSynth Src	37.00	**41.00** +4.00%	43.92	**46.38** +2.46%	37.06	**41.45** +4.39%	49.70	**61.45** +11.75%	38.28	**42.01** +3.73%	53.66	**55.97** +2.31%

Performance comparison between PAT and vanilla ZS classification across 6 ALEs and 18 diverse audio classification tasks

Dataset	Gaussian Noise ZS	Gaussian Noise PAT	Pitch Shift ZS	Pitch Shift PAT	Polarity Inversion ZS	Polarity Inversion PAT	Gain ZS	Gain PAT	High Pass ZS	High Pass PAT
Sound										
ESC-50	91.80	**94.20** +2.40%	78.05	**80.10** +2.05%	91.85	**94.40** +2.55%	92.05	**94.85** +2.80%	82.35	**86.15** +3.80%
USD8K	77.26	**82.70** +5.44%	63.61	**70.31** +6.70%	77.43	**82.69** +5.26%	77.08	**82.67** +5.59%	71.12	**76.77** +5.65%
TUT	44.94	**45.74** +0.80%	**26.05**	26.04 -0.01%	45.68	**47.34** +1.66%	38.95	**41.97** +3.02%	**35.80**	35.00 -0.80%
VS	**81.31**	77.86 -3.45%	**76.61**	69.64 -6.97%	**78.98**	78.00 -0.98%	79.00	**79.44** +0.44%	74.07	**76.16** +2.09%
DCASE	38.32	**42.21** +3.89%	31.76	**34.01** +2.25%	38.93	**45.69** +6.76%	43.24	**45.28** +2.04%	33.40	**37.70** +4.30%
Gunshot Tri.	**25.00**	**25.00** +0.00%	**25.00**	**25.00** +0.00%	**25.00**	**25.00** +0.00%	**25.00**	**25.00** +0.00%	19.32	**22.72** +3.40%
SESA	67.62	**69.52** +1.90%	62.86	**64.76** +1.90%	67.62	**69.52** +1.90%	68.57	**69.52** +0.95%	48.57	**58.10** +9.53%
AudioSet	30.40	**31.15** +0.75%	22.37	**23.06** +0.69%	**30.40**	29.22 -1.18%	28.78	**30.23** +1.45%	23.89	**24.38** +0.49%
FSD50K	44.54	**45.74** +1.20%	37.16	**43.87** +6.71%	44.39	**44.96** +0.57%	**44.56**	43.79 -0.77%	37.94	**43.73** +5.79%
Cochlscene	**85.07**	84.36 -0.71%	60.18	**61.42** +1.24%	85.07	**85.17** +0.10%	81.97	**82.09** +0.12%	73.34	**75.15** +1.81%
Music										
Beijing Op.	70.34	**70.62** +0.28%	61.02	**62.74** +1.72%	71.19	**71.61** +0.42%	69.49	**69.61** +0.12%	**65.68**	64.86 -0.82%
GTZAN	55.77	**58.53** +2.76%	47.56	**50.38** +2.82%	56.43	**58.26** +1.83%	55.20	**57.79** +2.59%	47.36	**50.54** +3.18%
MUSDB	**63.60**	53.60 -10.00%	58.00	**61.60** +3.60%	**68.00**	56.80 -11.20%	**68.00**	58.40 -9.60%	46.80	**55.20** +8.40%
Medley	**96.61**	95.96 -0.65%	92.09	**92.46** +0.37%	95.98	**96.42** +0.44%	95.97	**96.53** +0.56%	**93.37**	90.94 -2.43%
Mri. St.	42.63	**48.93** +6.30%	33.15	**44.94** +11.79%	44.09	**47.12** +3.03%	41.31	**44.67** +3.36%	34.73	**37.30** +2.57%
Mri. To.	24.97	**26.61** +1.64%	13.54	**17.12** +3.58%	22.02	**26.18** +4.16%	19.08	**26.78** +7.70%	**17.24**	16.59 -0.65%
NSynth Inst	52.86	**53.85** +0.99%	60.11	**64.77** +4.66%	63.89	**66.33** +2.44%	61.89	**64.62** +2.73%	46.22	**47.87** +1.65%
NSynth Src	39.75	**47.85** +8.10%	49.44	**59.37** +9.93%	49.76	**61.45** +11.69%	49.49	**60.64** +11.15%	47.39	**56.46** +9.07%

Zero-shot performance measure of MSCLAP-23 using PAT across 18 audio classification tasks under noisy setting.

QUALITATIVE RESULTS

Dataset	Prompt	Score
ESC-50	The sound of <label> coming from a cliff edge.	0.0035
	A sound of a <label> coming from a parking lot	0.0033
NSynth Inst	A major sound of a <label>	0.0038
	A minimal sound of a <label>	0.0037

We present top two highest scoring prompt by PAT for MSCLAP-23 on ESC-50 and Nsynth-Inst

ABLATION RESULTS

- Performance comparison of MS-CLAP with across different combinations of original, audio-guided, and text-guided logits with Cross-Modal Alignment.

Logits	Audio Guided	Text Guided	Average Accuracy
✓	×	×	58.28
✓	×	✓	59.24
✓	✓	×	59.55
×	✓	✓	58.29
✓	✓	✓	**60.76**

- Performance comparison of MS-CLAP using various components of PAT including Prompt Ensemble, Weighted Prompt Ensemble, Cross-Modal Alignment

Method	Average Accuracy
MS-CLAP	58.28
MS-CLAP+PE	58.91
MS-CLAP+WPE	59.23
MS-CLAP+PE+CMA	59.32
MS-CLAP+ PAT (WPE+CMA)	**60.76**