

Oregon State University

Classifying Audio Timbre Without Audio Using Text-only Training

EvoMUSART 2026

Peter McCabe, Patrick J. Donnelly

November 22, 2026

soundbendor lab

Introduction

Motivation 1

The Data Collection Problem

- (Modern) Machine learning is data hungry
- Niche data are difficult to obtain
- Audio, images, video are expensive to acquire, label, and store

Text-only Training offers a solution to all these problems

Text-only Training Definition

Text-only Training refers to use of text as the only training data for a machine learning problem not typically trained on language data.

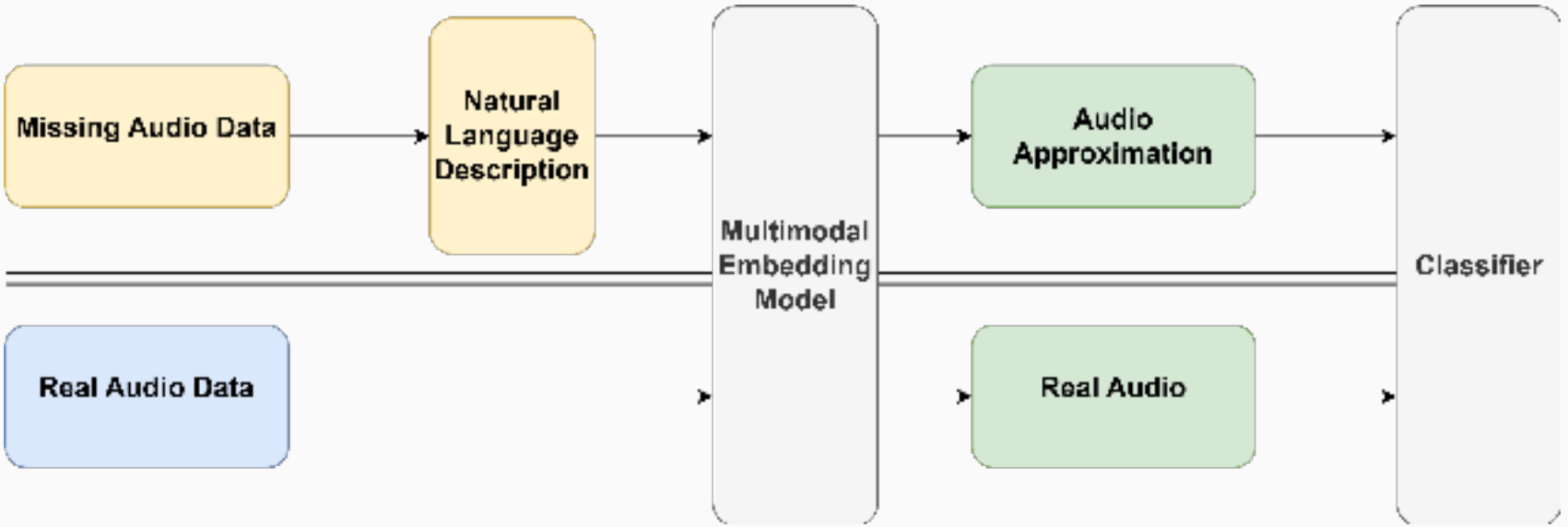
Text-only Training Requirements

1. Little to no genuine data
2. A multimodal embedding model
3. Desired data is well-described by language

Text-Only Training

Text-only Training:

Training machine learning models with descriptions of cross-modal data



Top: Training. **Bottom:** Inference.

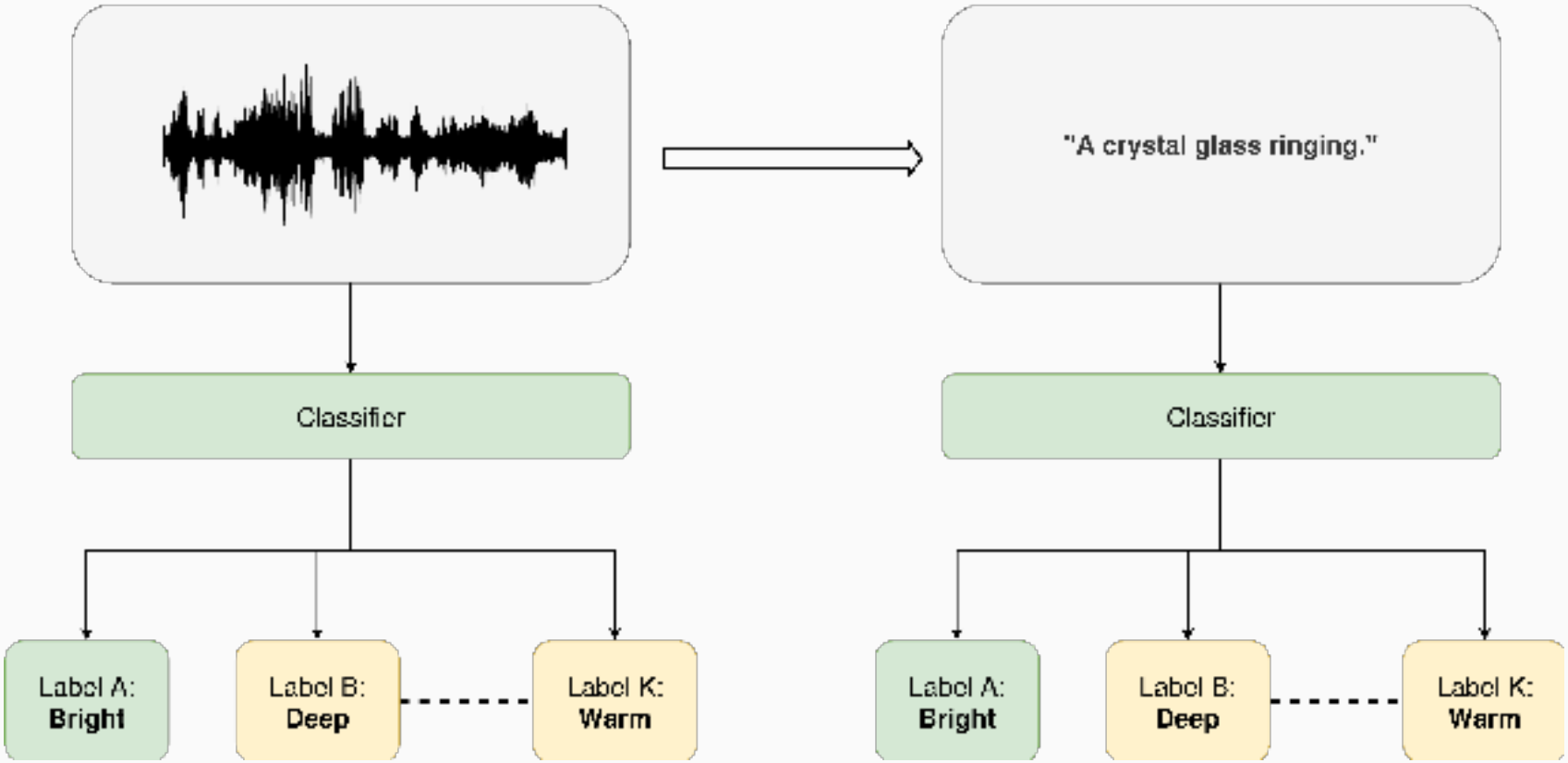
Case Study: Audio-Timbre Classification 1

Audio-Timbre Classification

The process of identifying the *subjective description* that best matches the timbre of an audio recording.

In other words: How would a human describe this sound?

Audio Timbre Classification



Datasets

Text Datasets Preliminary: Audio Test Sets

The 8 audio-timbre test datasets inform training set creation.

Class:	Booming	Bright	Deep	Hard	Reverb	Rough	Sharp	Warm
Count	102	210	218	206	220	212	165	184
Prop. Positive	.30	.62	.61	.67	.53	.55	.60	.42

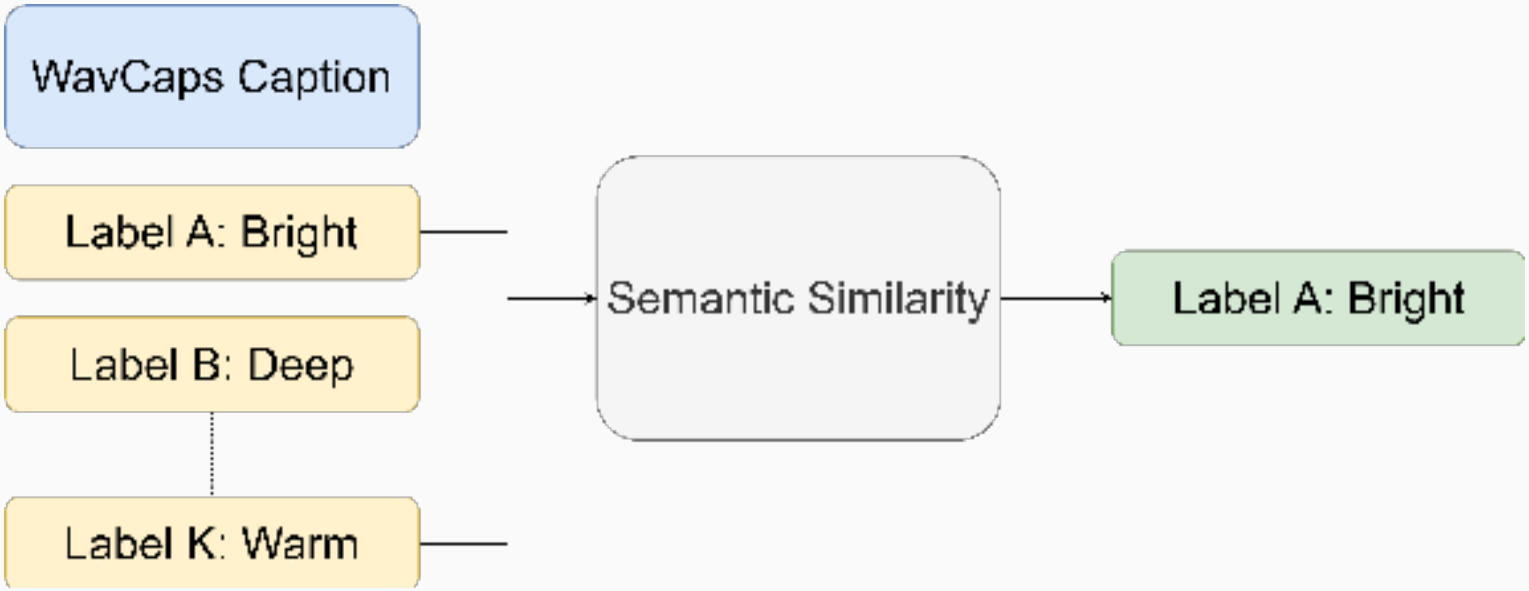
We simplify the AudioCommons [6] timbre datasets. Each word has a different small dataset of human ratings.

Preprocessing

Each dataset is binarized from a scale of 0-100 to either 0 if the word does *not* describe the audio, or 1 if the word *does* describe the audio

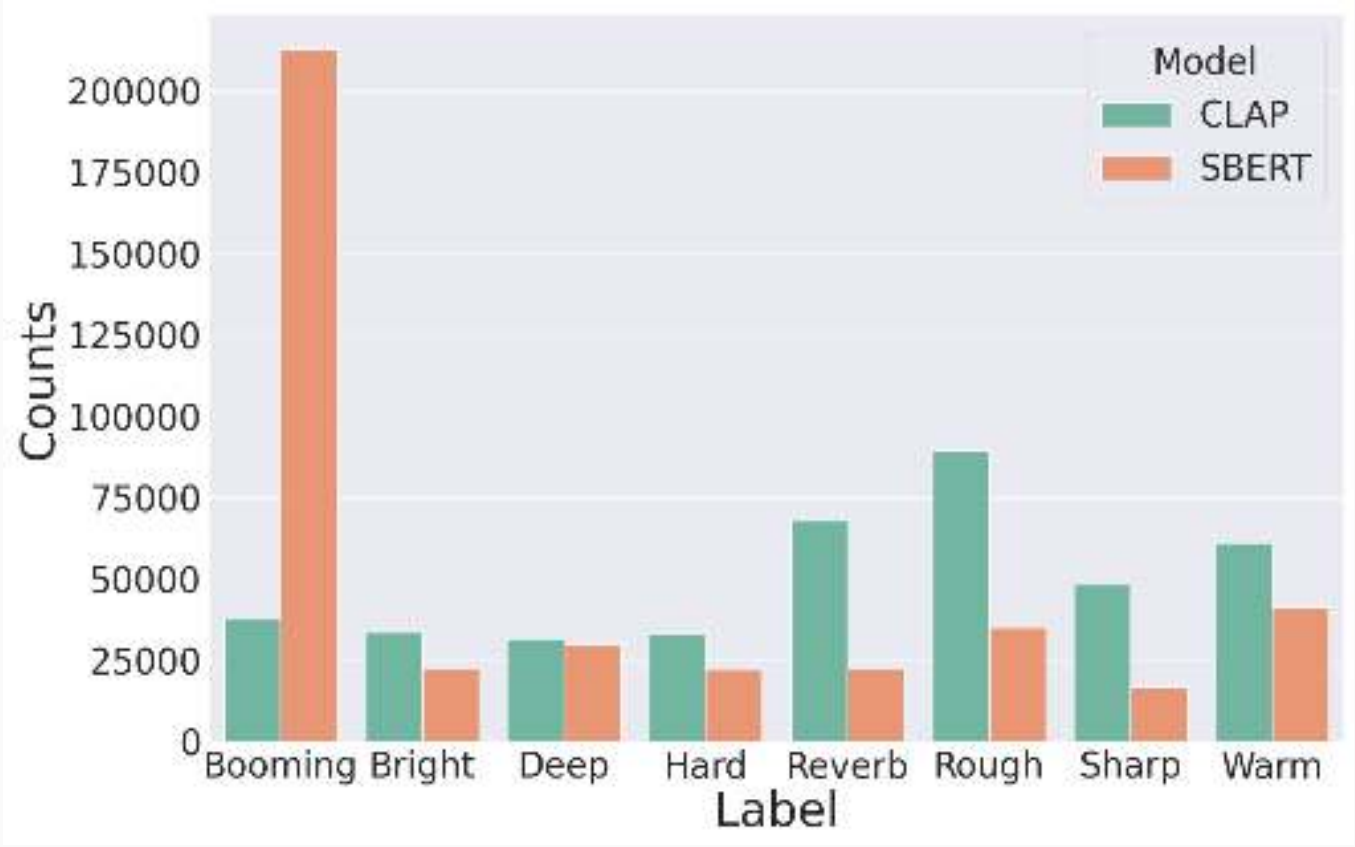
Text Dataset 1-2: Repurpose WavCaps

Dataset 1-2
Label WavCaps [5] captions with CLAP [8] or SBERT [7]



We choose the ALL-MINILM-L12-V2 variant of SBERT. It is an all-purpose text semantic similarity model.

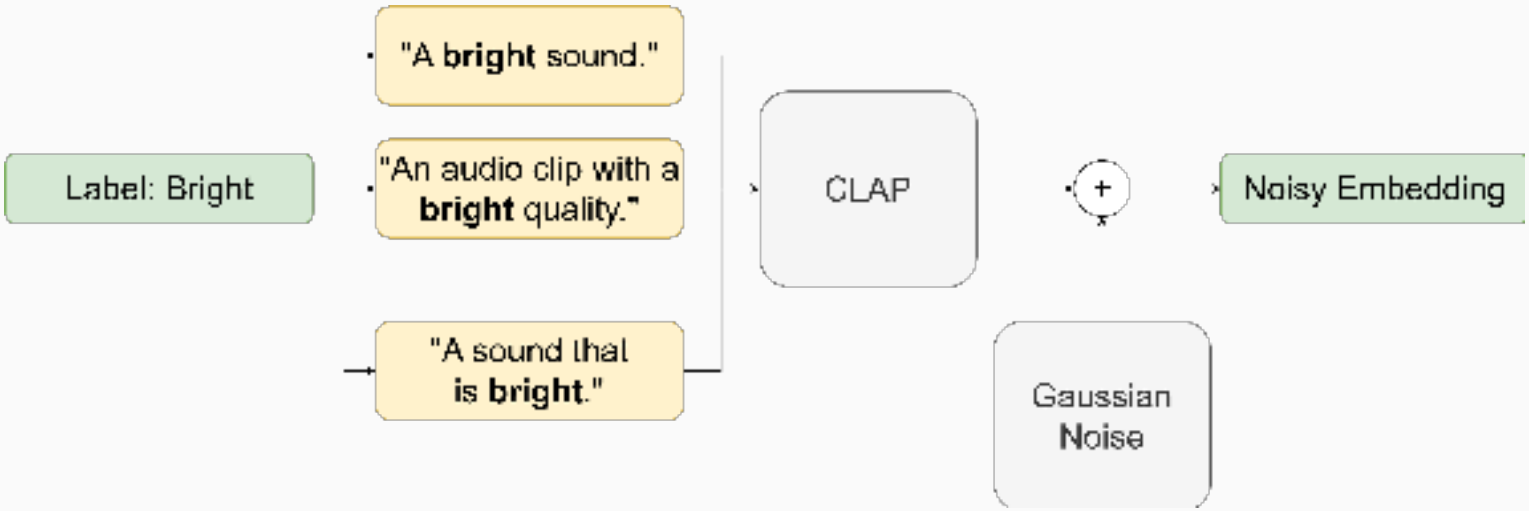
Text Dataset 1-2: WavCaps Distributions



The chosen semantic similarity model biases the dataset distribution.

Text Dataset 3: Label-Centered Noise

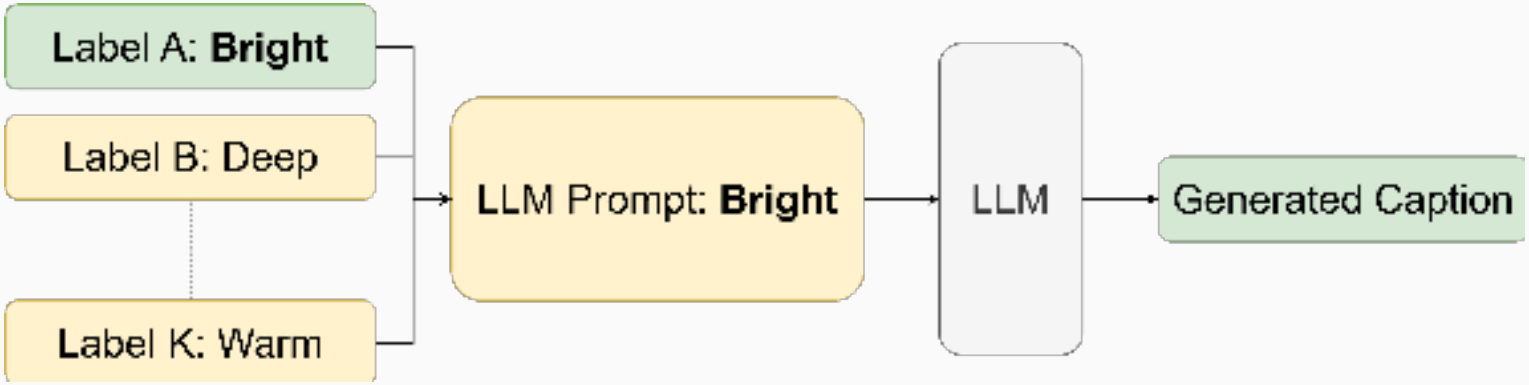
Dataset 1-2	Dataset 3
Label WavCaps [5] captions with CLAP [8] or SBERT [7]	Add noise to class label embeddings



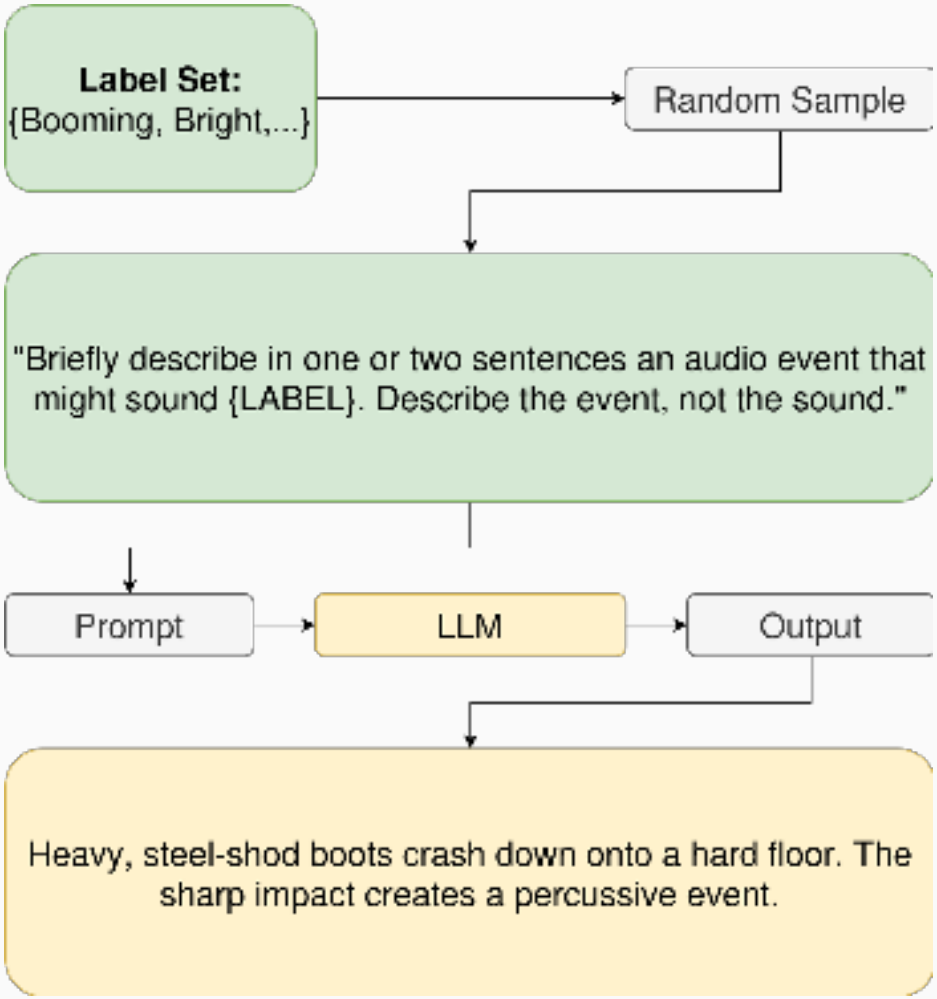
Sentences near the anchor prompts should share semantic meaning.

Text Dataset 4: LLM-Generated Captions

Dataset 1-2	**Dataset 3**	**Dataset 4**
Label WavCaps [5] captions with CLAP [8] or SBERT [7]	Add noise to class label embeddings	Generate labeled audio captions with an LLM



Text Dataset 4: LLM Prompt



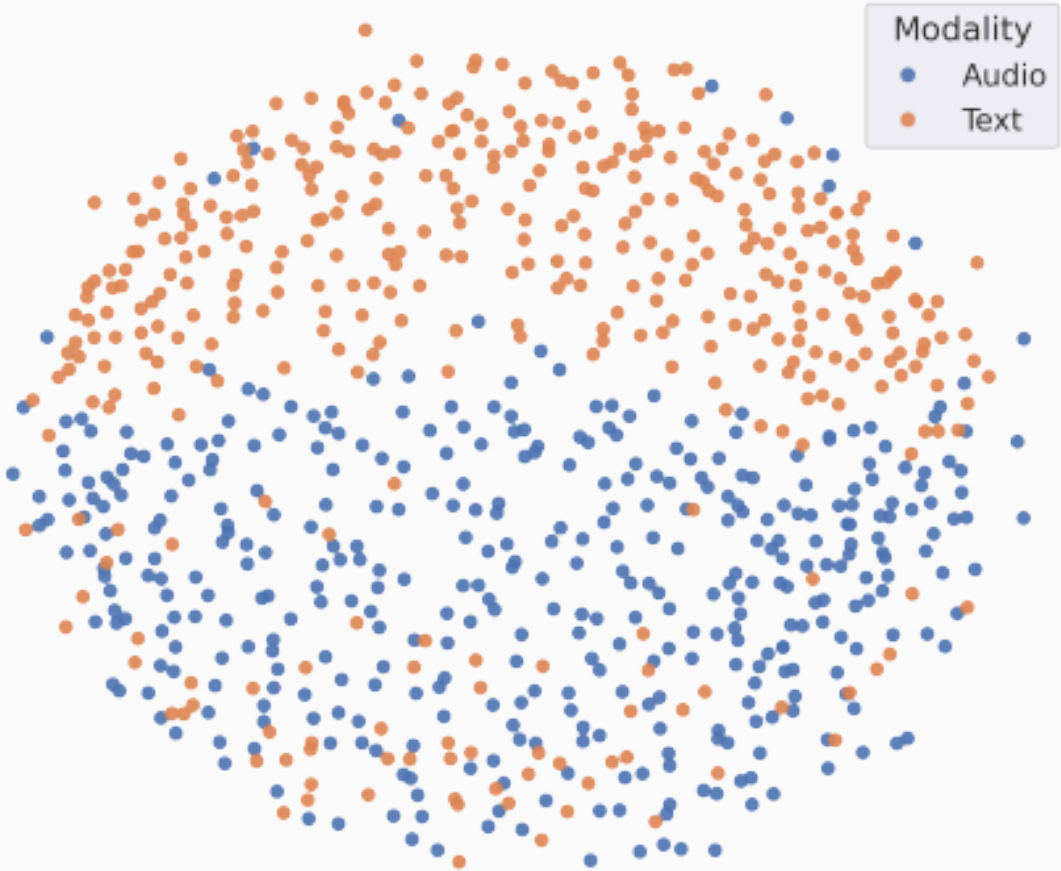
We chose LLAMA-3.1-8B-INSTRUCT as our LLM.

Text Dataset Recap

Dataset 1-2	**Dataset 3**	**Dataset 4**
Label WavCaps [5] captions with CLAP [8] or SBERT [7]	Add noise to class label embeddings	Generate labeled audio captions with an LLM

Augmentations

The Modality Gap in CLAP

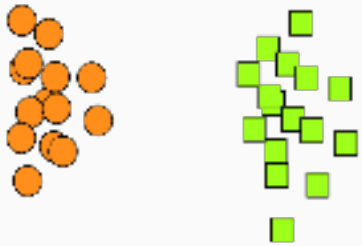


Random sample of 400 audio and text samples from Clotho [1]

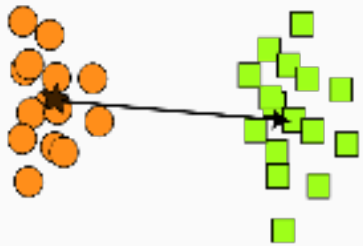
soundbendor lab

Augmentation Methods

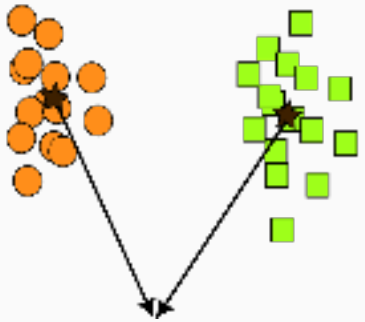
Clean	Nothing	Mean Shifting [4]	Connect-Collapse (C2) [9]
Noisy	Gaussian Noise [3]	Shift + Noise [2]	C2 + Corrupt (C3) [10]



No augmentation



Mean shifting pushes one modality towards the other



C2 centers data at 0

Experiments

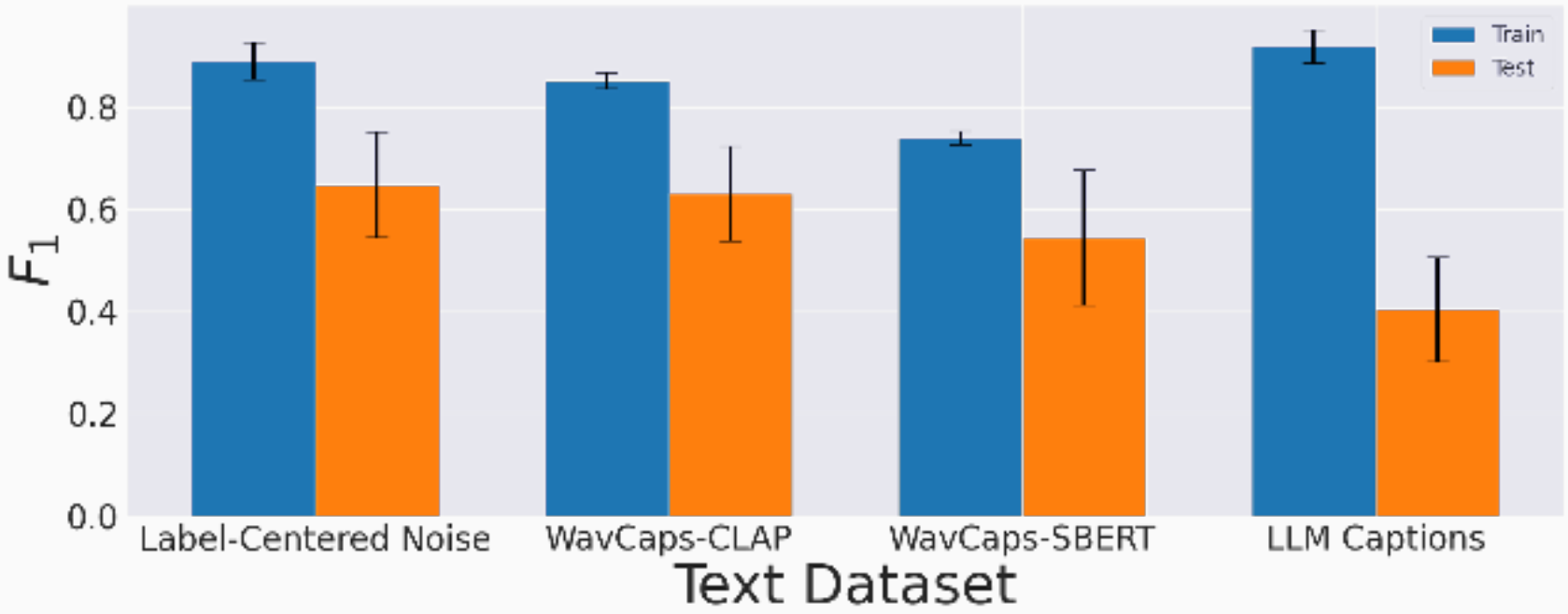
Experimental Design

- Select training set, augmentation method, and test set
- Train regularized linear classifier [10]
- Repeat 10 times

Class:	Booming	Bright	Deep	Hard	Reverb	Rough	Sharp	Warm
Count	102	210	218	206	220	212	165	184
Prop. Positive	.30	.62	.61	.67	.53	.55	.60	.42

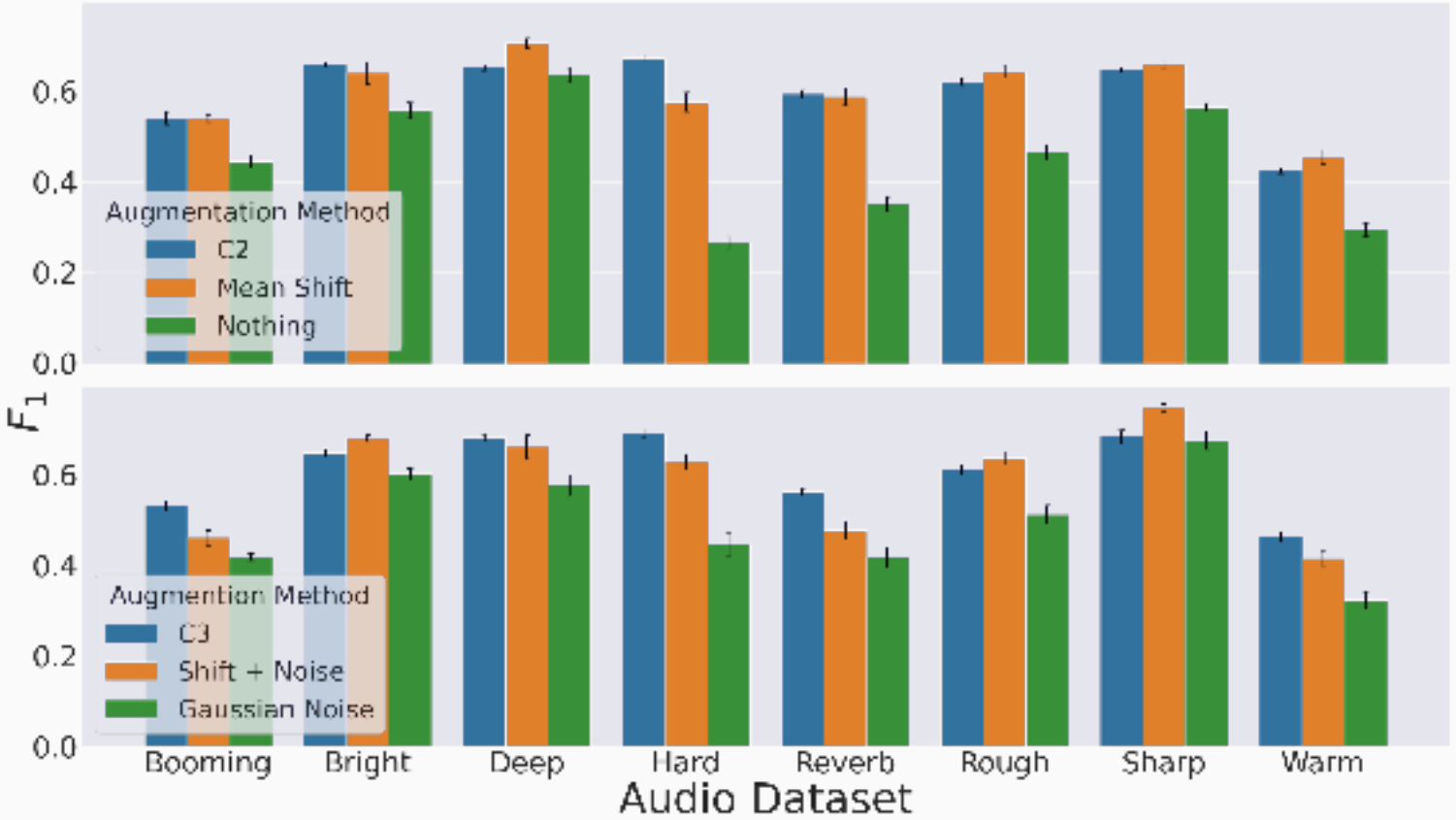
Results

Data Creation Methods Results



Differences between the text (training) and audio (test) F_1-scores

Augmentation Method Results



Top: Clean augmentation methods
Bottom: Noise-added augmentation methods

Baseline Results

Dataset	Booming	Bright	Deep	Hard	Reverb	Rough	Sharp	Warm
Average Model	0.52	0.66	0.68	0.64	0.56	0.63	0.68	0.44
CLAP Zero-Shot	0.58	0.76	**0.85**	0.25	0.57	0.62	0.75	0.47
Best Model	**0.67**	**0.79**	**0.85**	**0.81**	**0.66**	**0.78**	**0.80**	**0.62**
Audio Training	0.72	0.79	0.89	0.85	0.82	0.81	0.78	0.77

On average, models trained with ToT are competitive with CLAP
Our best ToT models are competitive with training on real audio

Conclusion

- Three data creation techniques:
  - WavCaps Captions, Label-Centered Noise, LLM-Generated Captions
- Six modality gap augmentation methods:
  - Nothing, Mean Shift, Connect-Collapse (C2)
  - Gaussian Noise, Shift+Noise, Connect-Collapse-Corrupt (C3)
- Text-only Training successfully trained an audio-timbre classifier

Dataset	Booming	Bright	Deep	Hard	Reverb	Rough	Sharp	Warm
Average Model	0.52	0.66	0.68	0.64	0.56	0.63	0.68	0.44
CLAP Zero-Shot	0.58	0.76	**0.85**	0.25	0.57	0.62	0.75	0.47
Best Model	**0.67**	**0.79**	**0.85**	**0.81**	**0.66**	**0.78**	**0.80**	**0.62**
Audio Training	0.72	0.79	0.89	0.85	0.82	0.81	0.78	0.77

References

- Drossos, K., Lipping, S., Virtanen, T.: Clotho: an Audio Captioning Dataset. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 736–740. IEEE, Barcelona, Spain (May 2020). https://doi.org/10.1109/ICASSP40776.2020.9052990, https://ieeexplore.ieee.org/document/9052990/
- Gu, S., Clark, C., Kembhavi, A.: I can't believe there's no images! : Learning Visual Tasks Using Only Language Supervision. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2672–2683 (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.00252, https://ieeexplore.ieee.org/document/10376589, iSSN: 2380-7504
- Kouzelis, T., Katsouros, V.: Weakly-Supervised Automated Audio Captioning via Text Only Training. In: Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023). pp. 81–85. Tampere, Finland (Sep 2023)
- Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. pp. 17612–17625. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (Nov 2022)
- Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M.D., Zou, Y., Wang, W.: WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. IEEE/ACM Transactions on Audio, Speech, and Language Processing **32**, 3339–3354 (2024). https://doi.org/10.1109/taslp.2024.3419446, https://ieeexplore.ieee.org/document/10572302/, publisher: Institute of Electrical and Electronics Engineers (IEEE)
- Pearce, A., Brookes, T., Mason, R.: Timbral Characterisation Tool v0.2 Development Dataset (Jan 2019). https://doi.org/10.5281/zenodo.2545496, https://zenodo.org/records/2545496
- Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3980–3990. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1410, https://www.aclweb.org/anthology/D19-1410
- Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE, Rhodes Island, Greece (Jun 2023). https://doi.org/10.1109/ICASSP49357.2023.10095969, https://ieeexplore.ieee.org/document/10095969/
- Zhang, Y., HaoChen, J.Z., Huang, S.C., Wang, K.C., Zou, J., Yeung, S.: Diagnosing and Rectifying Vision Models using Language (Sep 2022), https://openreview.net/forum?id=D-zfUK7BR6c
- Zhang, Y., Sui, E., Yeung, S.: Connect, Collapse, Corrupt: Learning Cross-Modal Tasks with Uni-Modal Data (Oct 2023), https://openreview.net/forum?id=ttXg3SKAg5

Thank You!