Synthio: Augmenting Small-Scale Audio Classification Datasets with Synthetic Data

Sreyan Ghosh[♦♠*], Sonal Kumar[♠*], Zhifeng Kong[♦], Rafael Valle[♦], Bryan Catanzaro[♦] Dinesh Manocha[♠]
[♦]NVIDIA, CA, USA, [♠]University of Maryland, College Park, USA

Modular framework for Synthetic Data Generation

T2A model trained on ~8M data points

Home Page

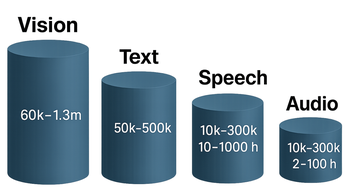
Audio Understanding Suffers from the Lack of Enough Data

(1) Audio (sounds & music) is low-resource compared to text, vision, and speech.

(2) Clean, diverse audio is hard to scrape or collect.

(3) Models trained on benchmarks generalize poorly to audio beyond the specific settings.

(4) Traditional augmentations alter data statistically—not semantically—limiting diversity and realism. They focus on surface-level patterns in the data rather than capturing the fundamental mechanisms that drive real-world data generation.



Problems with Generating Synthetic Data for Audio

Synthetic data can help—but comes with hurdles:

- Lack of strong open-source text-to-audio (T2A) models to generate **diverse** data at scale.
- Shortage of large-scale training data for training T2A models.
- T2A models trained on internet scale data *struggle* to generate audios **acoustically consistent** with the downstream dataset.
- T2A models struggle with prompt-following and fine-grained control, making tailored generation hard.

Figure 1. **(Left:)** Acoustic characteristics for the label ***bus*** vary by region or dataset (e.g., loud engines vs. quiet ones). *This hurts consistency.* **(Right:)** T2A models spuriously associate the label ***airport*** with announcements, which vary across regions. *This hurts diversity.*

Synthio: A Novel Synthetic Data Generation Pipeline for Audio

We propose Synthio, a new pipeline for generating high-quality synthetic audio. The primary aim is to generate synthetic data to expand small-scale audio classification datasets. It has 3 key parts:

(1) A T2A model trained on large-scale data: We train a DiT-based T2A model ≈8M cleaned audio-caption pairs (using Stable Audio Open), ensuring no leakage with benchmark datasets.

(2) Aligning the Text-to-Audio Models with Preference Optimization: We align T2A outputs with dataset-specific acoustic traits using **preference optimization**—so the synthetic audio "sounds like" examples in the downstream dataset.

(3) Generating Diverse Synthetic Augmentations: We propose **MixCap**, where we prompt the LLM to remix acoustic elements and imagine new soundscapes. Additionally, a self-reflection module filters and improves captions.

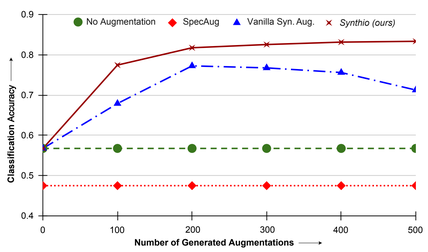
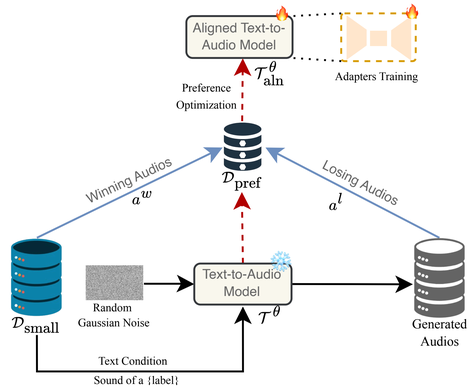


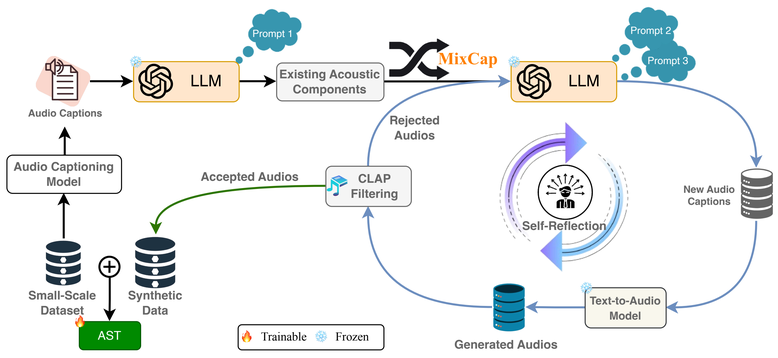
Figure 2. Synthio outperforms other methods on ESC-50 (100 samples) by generating consistent and diverse synthetic augmentations.

1. Aligning the T2A Model with Preference Optimization



- We propose aligning the student (T2A model) with the teacher (human-labeled dataset) using preference optimization (DPO) to ensure generated audio sounds similar to desired dataset characteristics
- In **Step 1**, we create a preference dataset by generating multiple synthetic audios per caption and pairing them with gold-standard audio, labeling them as "winner" or "loser".
- In **Step 2**, we train the T2A model using DPO so it prefers generating audio that matches the labeled "winner" characteristics.
- Standard fine-tuning (ERM) encourages "similar" audios and overfits on small-scale datasets. Our approach encourages diverse, dataset-aligned augmentations that "sound similar" and outperform ERM.

2. Generating Diverse Synthetic Augmentations



- We propose **language-guided audio imagination**, where we use LLMs to generate meaningful and controllable captions for synthetic audio generation.
- Our method, **MixCap**, extracts acoustic components (e.g., background sounds, events) using a captioning model and prompts the LLM to remix these with ground-truth labels for richer, more diverse captions.
- We filter generated audio-caption pairs using **CLAP-based similarity**: only audios closely matching the target label are accepted.
- For rejected samples, we use **LLM-based self-reflection** to revise captions and re-generate audio, iterating until quality thresholds are met or no rejected samples remain.
- The final accepted set of audios and captions is used to **expand the small-scale dataset and fine-tune an audio classification model** for downstream tasks, helping it learn both global structure and compositional acoustic detail.

Main Results

n	Method	ESC-50	USD8K	GTZAN	Medley	TUT	NSynth	VS	MSDB	DCASE	FSD50K
50	Gold-only (No Aug.)	22.25	55.09	47.05	47.23	37.60	33.32	77.49	56.85	12.09	7.16
	Random Noise	18.50	57.42	45.20	46.55	35.86	32.42	76.41	52.55	13.21	8.06
	Pitch Shifting	20.55	59.32	46.80	48.17	37.22	34.34	78.17	54.50	12.93	10.04
	SpecAugment	19.50	58.36	46.00	47.18	36.73	27.32	77.27	53.25	12.81	7.93
	Audiomentations	20.35	60.13	47.25	48.30	38.24	28.15	79.12	54.51	13.28	10.17
	Retrieval	19.20	37.14	42.55	43.65	35.80	31.27	71.42	51.35	10.53	7.28
	Vanilla Syn. Aug.	40.75	63.54	55.35	47.23	41.50	33.17	78.37	54.10	15.89	10.63
	+ LLM Caps.	36.80	65.84	63.74	55.36	40.90	38.17	78.77	57.05	13.07	10.70
	Synthio *(ours)*	**49.50**$_{\uparrow 122\%}$	**76.12**$_{\uparrow 38\%}$	**68.20**$_{\uparrow 44\%}$	**60.58**$_{\uparrow 28\%}$	**43.84**$_{\uparrow 17\%}$	**40.83**$_{\uparrow 22\%}$	**80.67**$_{\uparrow 4\%}$	**60.15**$_{\uparrow 5\%}$	**17.23**$_{\uparrow 42\%}$	**13.91**$_{\uparrow 94\%}$
100	Gold-only (No Aug.)	56.75	72.89	64.15	57.81	47.14	39.11	84.32	65.60	12.50	10.53
	Random Noise	58.50	71.54	65.50	56.98	46.21	38.20	83.33	66.15	13.35	13.71
	Pitch Shifting	59.55	73.52	66.75	58.46	47.50	39.53	85.07	68.25	12.19	13.11
	SpecAugment	47.50	72.43	69.75	58.06	50.07	41.96	85.14	66.40	14.17	14.80
	Audiomentations	48.50	73.82	71.05	59.32	51.14	42.15	85.24	68.40	16.93	13.55
	Retrieval	52.45	68.24	61.55	54.83	45.39	37.84	83.27	58.55	10.93	10.05
	Vanilla Syn. Aug.	77.25	77.31	68.25	63.58	49.96	42.31	84.78	63.55	15.73	12.63
	+ LLM Caps.	67.05	79.73	67.90	65.79	48.63	41.83	84.83	65.95	16.32	13.25
	Synthio *(ours)*	**83.35**$_{\uparrow 47\%}$	**85.00**$_{\uparrow 17\%}$	**71.20**$_{\uparrow 11\%}$	**71.23**$_{\uparrow 23\%}$	**52.42**$_{\uparrow 11\%}$	**44.92**$_{\uparrow 15\%}$	**86.70**$_{\uparrow 3\%}$	**68.80**$_{\uparrow 5\%}$	**19.38**$_{\uparrow 55\%}$	**16.35**$_{\uparrow 55\%}$
200	Gold-only (No Aug.)	84.75	74.80	77.00	67.41	55.32	48.77	87.38	68.80	23.15	13.59
	Random Noise	83.55	75.15	75.50	66.71	54.42	47.83	86.45	65.45	24.82	15.32
	Pitch Shifting	84.90	74.48	78.55	67.74	55.44	48.12	87.47	69.80	23.11	17.51
	SpecAugment	85.10	76.46	76.25	65.70	55.72	54.80	87.42	69.25	27.36	17.93
	Audiomentations	85.25	75.80	77.30	67.00	55.21	53.15	86.08	70.50	26.29	18.36
	Retrieval	82.55	71.20	73.65	65.80	53.25	47.63	86.28	63.55	19.51	15.36
	Vanilla Syn. Aug.	85.40	77.96	77.10	78.97	55.51	55.20	86.49	72.95	28.55	19.04
	+ LLM Caps.	85.80	78.37	79.55	74.14	54.73	56.21	87.02	73.16	28.40	18.14
	Synthio *(ours)*	**86.10**$_{\uparrow 2\%}$	**82.81**$_{\uparrow 11\%}$	**82.05**$_{\uparrow 7\%}$	**79.40**$_{\uparrow 18\%}$	**56.83**$_{\uparrow 3\%}$	**57.10**$_{\uparrow 17\%}$	**87.52**$_{\uparrow 0.2\%}$	**80.40**$_{\uparrow 17\%}$	**32.81**$_{\uparrow 42\%}$	**20.85**$_{\uparrow 53\%}$

Result Analysis

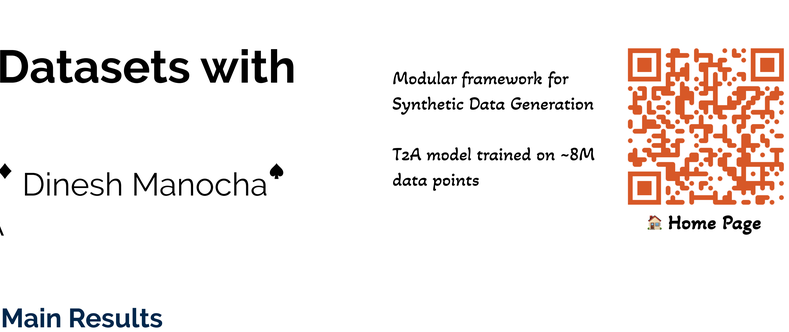


Figure 4. Comparison of spectral and pitch features between generated audios in $\mathcal{D}_{\text{syn}}$ and real audios in $\mathcal{D}_{\text{small}}$ (for n = 100). Synthio-generated audios closely replicate the features of real data.

#	Method	USD8K(↓)	NSynth(↓)
100	Vanilla Syn. Aug.	45.17	31.76
	Synthio *(ours)*	**35.09**	**22.97**
	w/ Template Captions	46.82	33.00
	w/ ERM	50.01	42.33
200	Vanilla Syn. Aug.	47.22	33.81
	Synthio *(ours)*	**34.58**	**23.03**
	w/ Template Captions	46.84	37.16
	w/ ERM	52.54	43.98

n	Method	GTZAN	VS	TUT	MSDB
100	Gold-only (No Aug.)	64.15	84.32	47.14	65.60
	Vanilla Syn. Aug.	29.05	34.13	21.69	35.60
	Synthio *(ours)*	**33.10**	**39.20**	**24.51**	**56.45**
	w/ Template Captions	24.50	30.99	21.73	40.40
	w/ ERM	25.65	32.76	24.40	42.85
	w/o DPO	17.60	21.57	20.39	30.20
200	Gold-only (No Aug.)	77.00	87.38	55.32	68.80
	Vanilla Syn. Aug.	32.35	41.96	24.23	39.25
	Synthio *(ours)*	**35.15**	**48.14**	**27.00**	**61.45**
	w/ Template Captions	29.90	35.53	23.61	41.20
	w/ ERM	28.10	36.29	25.71	46.70
	w/o DPO	19.85	26.85	21.40	36.75

Figure 5. **(Left)** CLAP similarity score between real audios and generated data. **(Right)** Performance comparison of Synthio with baselines on *synthetic-only* audio classification.

Future Work

- Adopt and modify the Synthio for more complex audio understanding tasks like AQA.
- We want to make Synthio computationally less expensive.
- We want to extend the applications of Synthio to spoken language.
- We want to explore the impact of controlled data generation on performance.