ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10890066

# SoloAudio: Target Sound Extraction with Language-oriented Audio Diffusion Transformer

Helin Wang†, Jiarui Hai†, Yen-Ju Lu, Karan Thakkar, Mounya Elhilali, and Najim Dehak
Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, USA
Email: hwang258@jhu.edu, jhai2@jhu.edu

***Abstract*—In this paper, we introduce SoloAudio, a novel diffusion-based generative model for target sound extraction (TSE). Our approach trains latent diffusion models on audio, replacing the previous U-Net backbone with a skip-connected Transformer that operates on latent features. SoloAudio supports both audio-oriented and language-oriented TSE by utilizing a CLAP model as the feature extractor for target sounds. Furthermore, SoloAudio leverages synthetic audio generated by state-of-the-art text-to-audio models for training, demonstrating strong generalization to out-of-domain data and unseen sound events. We evaluate this approach on the FSD Kaggle 2018 mixture dataset and real data from AudioSet, where SoloAudio achieves the state-of-the-art results on both in-domain and out-of-domain data, and exhibits impressive zero-shot and few-shot capabilities. Source code[1] and demos[2] are released.**

***Index Terms*—target sound extraction, transformer, language-oriented, text-to-audio, zero-shot, few-shot.**

## I. Introduction

Human beings possess the remarkable ability to focus on a specific sound within a complex acoustic scene composed of various overlapping sound events [1], [2]. Recent works that aim to replicate this human capability computationally have framed the task as target sound extraction (TSE) [1], [3]–[5]. The objective of TSE is to extract sounds of interest from mixtures of overlapping audio, guided by clues that provide information about the target sound class. These clues can take the form of one-hot labels [3], [6], audio clips [7], or images [8], [9].

Most prior methods are based on discriminative models, which aim to minimize the difference between the estimated and target audio [4], [10]. While these models often produce good separation in non-overlapping regions, they tend to suffer significant performance degradation in overlapping areas. This is especially problematic in real-world scenarios where sound overlaps are common, making it a critical issue to address in TSE. With the advent of denoising diffusion probabilistic models (DDPMs) [11], [12], generative models have recently been applied successfully to TSE and source separation tasks [13]–[16]. DPM-TSE [14], a generative approach based on DDPM, achieves both cleaner target renderings and improved separability from unwanted sounds compared to discriminative models. However, DPM-TSE operates on log-mel spectrograms, where the diffusion process is applied, inherently limiting the reconstruction quality. Additionally, DPM-TSE relies solely on in-domain one-hot labels, which restricts its ability to generalize to out-of-domain data and unseen sound events.

Another challenge in the TSE task is the scarcity of training data, particularly clean, single-label audio, which is often used as the ground truth for target sounds. AudioSep [17] trains open-domain audio source separation models using natural language queries, with mixtures of large-scale multi-label audios. However, it struggles to produce clean sound isolations for a single target sound, which is critical for real-world applications.

To address these issues, we propose SoloAudio, an audio- and/or language-oriented diffusion Transformer model for TSE. Our main contributions are summarized as follows:
(i) We introduce a novel Transformer backbone with skip connections, applying the diffusion process in the latent space of an audio variational autoencoder (VAE). SoloAudio supports both audio clues and text clues, by utilizing a CLAP model [18].
(ii) We leverage synthetic audio from a text-to-audio (T2A) generation model [19] as additional training data. Thanks to advancements in T2A, we can generate high-quality, clean audio to improve the training of TSE models.
(iii) Experimental results on mixtures from the FSD Kaggle 2018 dataset [20] demonstrate that SoloAudio significantly outperforms state-of-the-art methods. Moreover, SoloAudio exhibits strong zero-shot and few-shot capabilities on out-of-domain data and unseen sound events.
(iv) Subjective evaluations on real-world data consistently demonstrate a clear preference among listeners for the audio extracted by SoloAudio, highlighting its superior ability to isolate target sounds while effectively eliminating irrelevant noise.

## II. Methodology

### A. Denoising Diffusion Probabilistic Model (DDPM)

DDPMs consist of a forward and backward process. The forward process incrementally adds Gaussian noise to the data, following a variance schedule $\beta_1, \ldots, \beta_T$.

$$q\left(x_t \mid x_{t-1}\right) := \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t \mathbf{I}\right) \tag{1}$$

The forward process enables sampling $x_t$ at any timestep $t$ in a closed form ($x_0$ is the clean signal):

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{2}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Following [14], we use a modified diffusion scheduler and $v$ prediction to improve the purity and overall performance of sound extraction. Additionally, we implement a diffusion noise schedule by keeping $\sqrt{\bar{\alpha}_1}$ unchanged, changing $\sqrt{\bar{\alpha}_T}$ to zero, and linearly rescaling $\sqrt{\bar{\alpha}_t}$ for intermediate $t \in [2, \ldots, T-1]$ respectively. This adjustment resolves the mismatch between training and inference and prevents the introduction of additional noise during sampling. A neural network is applied to predict velocity $v_t$:

$$v_t = \sqrt{\bar{\alpha}_t}\epsilon - \sqrt{1-\bar{\alpha}_t}x_0, \tag{3}$$

In the reverse process of diffusion models, the model gradually reconstructs the original data from a random Gaussian noise.

$$p_\theta\left(x_{t-1} \mid x_t\right) := \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t, \tilde{\beta}_t \mathbf{I}\right), \tag{4}$$

† Indicates equal contribution.
[1]https://github.com/WangHelin1997/SoloAudio
[2]https://wanghelin1997.github.io/SoloAudio-Demo

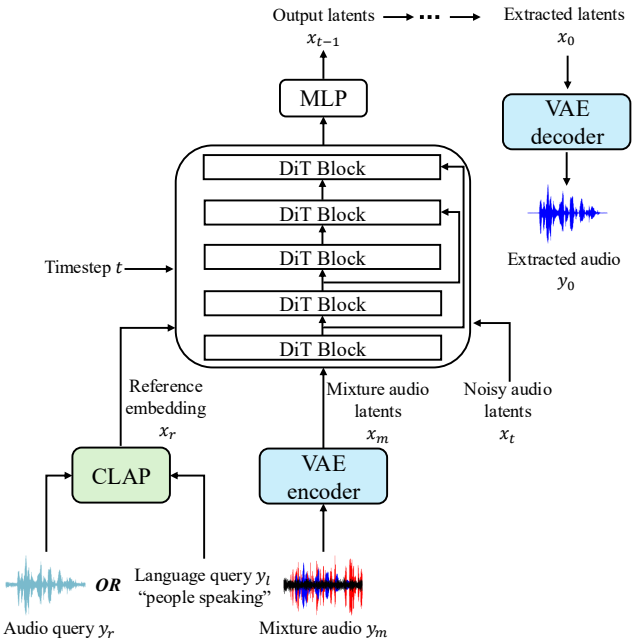



Fig. 1. Diagram of SoloAudio model.

where variance $\tilde{\beta}_t$ can be calculated from the forward process posteriors:

$$\tilde{\beta}_t := \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{5}$$

According to [14],

$$x_0 := \sqrt{\bar{\alpha}_t}x_t - \sqrt{1-\bar{\alpha}_t}v_t \tag{6}$$

$$\tilde{\mu}_t = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}\,(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}x_t \tag{7}$$

### B. SoloAudio

As shown in Fig. 1, our proposed SoloAudio model consists of several key components: a VAE encoder, a VAE decoder, a CLAP model, and a DiT-like model [21], [22].

Given a 1-D mixture audio signal $y_m$, the VAE encoder is applied to extract audio latents $x_m \in \mathbb{R}^{N\times C}$, where $N$ represents the number of feature frames, and $C$ denotes the dimension of the latent channels. We leverage the VAE latent space for the diffusion process due to its superior reconstruction quality compared to the mel spectrogram space [23]. The VAE model employs a fully-convolutional architecture, following the DAC encoder and decoder structure [24], but with a VAE bottleneck rather than vector quantization.

The CLAP model, which bridges language and audio spaces and enables zero-shot predictions [25], is used to extract the reference embedding $x_r$ from either an audio query $y_r$ or a language query $y_l$. For the noisy audio latents $x_t \in \mathbb{R}^{N\times C}$ at timestep $t$, we concatenate $x_t$ and $x_m$ on the channel dimension as the input to the DiT.

The DiT block, detailed in Fig. 2, includes an adaptive layer norm block, a multi-head self-attention (MHSA) block, and a multi-layer perceptron (MLP) block. The timestep $t$ and reference embedding $x_r$ serve as conditional information to regress the dimension-wise scale and shift parameters, which are incorporated into each block.

The primary distinction between our network architecture and DiT lies in the use of long skip connections in SoloAudio, bridging shallow and deep DiT blocks as in [26]. These skip connections

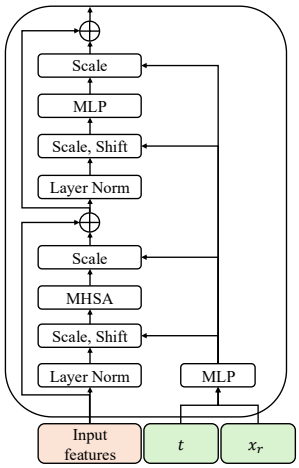



Fig. 2. Diagram of the DiT block.

create shortcuts for low-level features, streamlining the training of the entire $v$-prediction network. Furthermore, we incorporate rotary positional embeddings (RoPE) [27] for enhanced position encoding of audio latents.

### C. Inference

During inference, we obtain the output latents $x_t \in \mathbb{R}^{N\times C}$ by feeding $x_t$, $y_m$, $y_r$ (or $y_l$) and $t$ into the DiT model. After $T$ denoising sampling steps, the clean target latents $x_0 \in \mathbb{R}^{N\times C}$ could be estimated.

We apply classifier-free guidance (CFG) to steer the sampling process. This involves training the model in two modes: conditioned and unconditioned, enabling it to learn both how to generate general outputs and how to generate outputs that match specific conditioning inputs. The CFG technique adjusts the model's output $v_t$ during sampling, which can be expressed as:

$$v_t' = v_t^{uncond} + \gamma(v_t^{cond} - v_t^{uncond}) \tag{8}$$

where $\gamma$ represents the guidance scale, $v_t^{uncond}$ is the prediction of the unconditioned sampling and $v_t^{cond}$ is the prediction of the conditioned sampling.

## III. Experiments

### A. DataSet

*1) Synthetic data (FSD-Mix):* Following [3], [6], we created datasets of simulated mixtures using the Freesound Dataset Kaggle 2018 corpus[3] (FSD) [20]. The audio clips in the FSD vary in length, from 0.3 to 30 seconds. We generated 10-second audio mixtures, each consisting of one target sound and 1-3 interfering sounds, randomly selected from the FSD. The signal-to-noise ratio (SNR) of the interfering sounds is randomly set within a range of $-10$ to 10 dB. These sounds were superimposed at random time points over a 10-second background noise, sourced from the DCASE 2019 Challenge's acoustic scene classification task[4] [28]. The SNR for the background noise was randomly set between $-5$ and 10 dB. All audio clips were resampled to 24 kHz. Each training audio file was simulated for 3 mixtures, resulting in $28,419$ samples for the training set, 160 for the validation set, and $1,440$ for the test set. The corpus

[3]https://www.kaggle.com/c/freesound-audio-tagging
[4]https://dcase.community/challenge2019/task-acoustic-scene-classification

TABLE I
RESULTS ON THE FSD-MIX DATASET.

| Method | Audio-oriented | | | | Language-oriented | | | |
|---|---|---|---|---|---|---|---|---|
| | FD ↓ | KL ↓ | CLAP-audio ↑ | ViSQOL ↑ | FD ↓ | KL ↓ | CLAP-audio ↑ | ViSQOL ↑ |
| DPM-TSE | 29.262 | 1.661 | 0.623 | 2.180 | 27.121 | 1.610 | 0.640 | 2.201 |
| SoloAudio | **5.875** | **1.108** | **0.772** | **2.411** | **4.986** | **0.976** | **0.801** | **2.498** |
| *w/o skip connection* | 9.128 | 1.304 | 0.738 | 2.369 | 8.481 | 1.170 | 0.763 | 2.457 |

contains 41 sound event categories, ranging from human-produced sounds to musical instruments and object noises.

*2) Synthetic data (TangoSyn-Mix):* A recently released variant of Tango [19], which has demonstrated state-of-the-art performance in text-to-audio generation, was used to synthesize data from text descriptions. Specifically, we used 300 categories from VGG-Sound [29] and manually assessed the quality of the generated audio by listening to three samples from each category as Tango might fail to actually generate some sound categories. After initial filtering, 227 categories were retained. For each category, we generated 24 samples using different random seeds and text augmentations. The TangoSyn-Mix dataset was created following the same simulated process as the FSD-Mix dataset, resulting in a training set with a total of $95,340$ audio files. Compared to FSD-Mix, 22 categories overlap with TangoSyn-Mix, while the remaining 19 categories are excluded from TangoSyn-Mix and reserved for evaluating the few-shot and zero-shot capabilities of the models.

*3) Real evaluation data (AudioSet):* The AudioSet evaluation set was used for real-world TSE evaluation [30]. We selected audio from 41 FSD categories and randomly chose 5 samples per category. After listening to these samples, we manually selected 2 samples per category to ensure the presence of the category-related sound, resulting in a total of 82 selected audio samples.

We open-source the training and evaluation data used in our experiments.

### B. Experimental Setups

We conducted experiments using a 24kHz audio sample rate for both the waveform VAE and the SoloAudio model. The waveform latent representation operates at 50Hz and contains 128 channels. The VAE was trained on AudioSet to handle a wide range of general audio classes. SoloAudio's DiT follows DiT-B[5], which is composed of 12 DiT blocks, each with 768 channels and 12 attention heads.

The CLAP embedding has a dimension of 512. We augment the text using the following formats: "[CLS]", "*An audio clip of* [CLS]", or "*The sound of* [CLS]", where [CLS] is the target sound category. Our model was trained using the AdamW optimizer with a learning rate of $0.0001$, weight decay of $0.0001$, a batch size of 128, and for 100 epochs. The diffusion and inference steps for SoloAudio are set to 1000 and 50, respectively, with the variance $\beta$ ranging from $0.00085$ to $0.012$. The model was trained on one NVIDIA A100-80GB GPU for two days. We allocated $10\%$ of the data for unconditioned training and $90\%$ for conditioned training. During sampling, the default guidance scale $\gamma$ is set to $2.5$ for audio-oriented TSE and $3.0$ for language-oriented TSE based on our ablation studies. For the few-shot experiments, we fine-tuned the model using the AdamW optimizer with a learning rate of $0.00001$, a weight decay of $0.0001$, and a batch size of 32 over 20 epochs.

[5]https://github.com/facebookresearch/DiT/blob/main/models.py

### C. Baselines

We compare SoloAudio with three modern TSE models: WaveFormer[6] [10], AudioSep[7] [17], and DPM-TSE[8] [14]. WaveFormer operates in the waveform domain, AudioSep works on the STFT representation, and DPM-TSE uses the mel-spectrogram. Both WaveFormer and AudioSep support text-oriented TSE, and due to limited computational resources, we directly use their official checkpoints for the real-world TSE evaluation. WaveFormer was trained on the FSD mixture dataset, while AudioSep was trained on large-scale audio mixtures. DPM-TSE is originally designed to use one-hot labels; we retrained it by substituting the one-hot embeddings with CLAP embeddings for both audio-oriented and language-oriented TSE.

### D. Metrics

Following [14], we introduce perceptual evaluations and subjective assessment to evaluate TSE models.

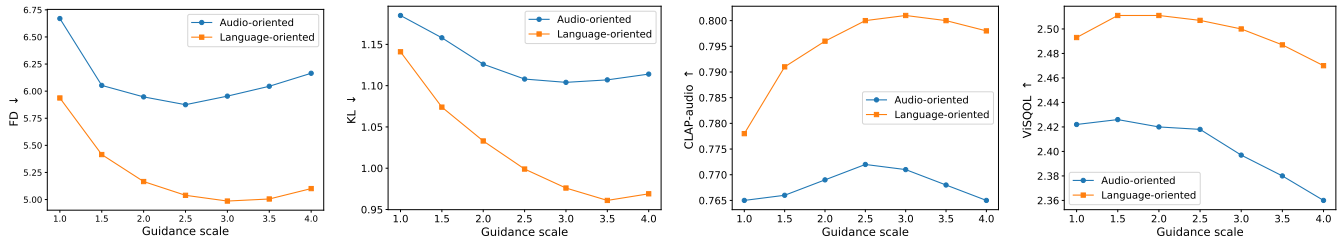
*1) Objective metrics:* We use five automatic evaluation functions: (i) **ViSQOL** [31] is an algorithm to assess the quality of audio signals by approximating human perceptual responses based on five-scaled mean opinion scores. (ii) **Frechet Distance (FD)** [32] in audio indicates the similarity between generated samples and target samples. (iii) **Kullback–Leibler (KL) divergence** is measured at a paired sample level and averaged as the final result. FD and KL are built upon a state-of-the-art audio classifier PANNs [33]. (iv) **CLAP-audio** is calculated using CLAP features between generated samples and target samples. (v) **CLAP-text** is calculated using CLAP features between generated samples and target text.

*2) Subjective metrics:* Following [14], we recruited 12 participants with recording or music production experiences to evaluate the listening perceptual quality of audios predicted by different TSE models. We evaluated the performance of language-oriented TSE in real-world application scenarios using the real evaluation data described in Section III-A3. Each subject was asked to evaluate 41 audio pairs for each model. Each audio pair included the original mixture, a description of the target sound, and the model's prediction for the extracted sound. For each audio pair, subjects were asked to respond to two questions:

**(i) Extraction: Does the generated audio contain the target sound as described in the text?** Ratings ranged from 1 to 5, where 1 indicated that the target sound could not be heard at all in the generated audio, and 5 indicated that the generated audio fully captured the target sound from the mixture as described.

**(ii) Purity: Does the generated audio only contain the sound corresponding to the text description?** Ratings ranged from 1 to 5, where 1 indicated that the generated audio contained many unrelated sounds, and 5 indicated that it contained only the target sound with no detectable unrelated sounds.

[6]https://github.com/vb000/Waveformer
[7]https://github.com/Audio-AGI/AudioSep
[8]https://github.com/haidog-yaqub/DPMTSE




(a) FD ↓ (b) KL ↓ (c) CLAP-audio ↑ (d) ViSQOL ↑

Fig. 3. Influence of the guidance scale

TABLE II
RESULTS ON THE FSD-MIX DATASET. WE TEST BOTH 22 SEEN LABELS (S) AND 19 UNSEEN LABELS (UNS) FROM THE SYNVGG-MIX TRAINING DATA.

| Method | Audio-oriented | | | | | | | | Language-oriented | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FD ↓ | | KL ↓ | | CLAP-audio ↑ | | ViSQOL ↑ | | FD ↓ | | KL ↓ | | CLAP-audio ↑ | | ViSQOL ↑ | |
| | S | UNS | S | UNS | S | UNS | S | UNS | S | UNS | S | UNS | S | UNS | S | UNS |
| FSD | 7.015 | 8.171 | 1.015 | 1.198 | 0.787 | 0.757 | 2.453 | 2.370 | 6.406 | 7.192 | 0.882 | 1.069 | 0.812 | 0.790 | 2.523 | 2.473 |
| FSD+TangoSyn | 5.317 | 6.158 | 0.879 | 1.132 | 0.806 | 0.771 | 2.498 | 2.398 | 4.779 | 5.048 | 0.676 | 0.818 | 0.844 | 0.827 | 2.630 | 2.583 |
| zero-shot | 22.673 | 20.303 | 1.983 | 2.220 | 0.594 | 0.576 | 2.003 | 2.050 | 39.884 | 38.009 | 1.685 | 2.333 | 0.641 | 0.570 | 2.146 | 2.014 |
| one-shot | 14.144 | 13.121 | 1.610 | 1.890 | 0.644 | 0.619 | 2.161 | 2.124 | 9.461 | 12.368 | 1.338 | 1.975 | 0.718 | 0.623 | 2.319 | 2.114 |
| 10-shot | 8.361 | 9.894 | 1.359 | 1.671 | 0.734 | 0.677 | 2.383 | 2.230 | 7.852 | 10.388 | 1.129 | 1.666 | 0.758 | 0.671 | 2.437 | 2.220 |

TABLE III
RESULTS ON THE REAL AUDIOSET DATASET. WE REPORT EXTRACTION AND PURITY RESULTS WITH THEIR 95% CONFIDENCE INTERVALS.

| Method | CLAP-text ↑ | Extraction ↑ | Purity ↑ |
|---|---|---|---|
| AudioSep | 0.168 | $\mathbf{4.431} \pm 0.089$ | $2.487 \pm 0.142$ |
| WaveFormer | 0.097 | $3.492 \pm 0.109$ | $3.266 \pm 0.132$ |
| DPM-TSE | 0.096 | $2.437 \pm 0.123$ | $3.939 \pm 0.130$ |
| SoloAudio | **0.213** | $3.923 \pm 0.113$ | $\mathbf{4.263} \pm 0.109$ |

## IV. RESULTS

### A. Comparison with DPM-TSE

We compare SoloAudio with DPM-TSE using in-domain data, training and testing both models on the FSD-Mix dataset under identical conditions. As shown in Table I, SoloAudio significantly outperforms DPM-TSE across all metrics. Both audio-oriented and language-oriented TSE highlight the effectiveness of SoloAudio. Besides, we found that the language-oriented TSE performs better thant the audio-oriented TSE.

### B. Ablation Studies

Table I shows the impact of adding skip connections to the DiT model, resulting in a clear performance improvement. In addition, we examine the impact of the CFG guidance scale on model performance. As shown in Fig. 3, as the guidance scale increases, performance initially improves but then declines. We select optimal values of 2.5 for the audio-oriented TSE and 3.0 for the language-oriented TSE.

### C. Influence of Synthetic Data

We compare the results of SoloAudio on FSD-mix data with and without synthetic data. The FSD data contains 22 labels present in the TangoSyn data, leaving 19 labels unseen. Table II highlights the impact of synthetic data, showing that using TangoSyn clearly improves TSE performance on both seen and unseen data.

### D. Zero-shot and Few-shot TSE

To further evaluate the few-shot and zero-shot capabilities of the models, we utilized the SoloAudio model trained exclusively on TangoSyn data. For the zero-shot setting, we directly tested the model on the out-of-domain FSD-Mix test set, which contains unseen labels. In the few-shot setting, we fine-tuned the model using either 1 or 10 samples per category from the FSD-Mix training set and evaluated its performance on the FSD-Mix test set. Table II presents these results. Overall, SoloAudio demonstrates remarkable zero-shot capability on out-of-domain data with unseen labels. Moreover, fine-tuning with a small number of samples (1 or 10) leads to a significant performance improvement across all metrics.

### E. Performance on Real Data

Furthermore, we performed both objective and subjective evaluations on real data to compare SoloAudio with three state-of-the-art TSE models. Table III summarizes the performance of the models, with our proposed SoloAudio achieving the highest CLAP-text score, demonstrating strong alignment with the target sound prompt. In the listening test, SoloAudio records the highest Purity score and a strong Extraction score, highlighting its clear advantage in isolating and recovering target sounds with minimal interference. Although AudioSep achieves the highest Extraction score, its low Purity score indicates difficulties in removing unrelated noise. This issue could arise from training the model on multi-label audio samples, which may hinder its ability to accurately extract individual sounds.

## V. CONCLUSIONS

In this paper, we propose a generative method for TSE, built on a latent diffusion model with a skip-connected Transformer. We also explore the use of synthetic data generated by T2A, demonstrating its strong potential for training TSE models. In future work, we aim to (1) improve the sampling speed of SoloAudio, (2) investigate more effective T2A tools and audio-text alignment methods, (3) scale up training with larger datasets, and (4) explore the use of alternative target references, such as images and videos.

## REFERENCES


[1] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, Y. Ohishi, and S. Araki, "Soundbeam: Target sound extraction conditioned on sound-class labels and enrollment clues for increased performance and continuous learning," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 121–136, 2023.

[2] D. Chong, H. Wang, P. Zhou, and Q. Zeng, "Masked spectrogram prediction for self-supervised audio pre-training," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5.

[3] H. Wang, D. Yang, C. Weng, J. Yu, and Y. Zou, "Improving target sound extraction with timestamp information," in *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, H. Ko and J. H. L. Hansen, Eds. ISCA, 2022, pp. 1526–1530.

[4] M. Delcroix, J. B. Vázquez, T. Ochiai, K. Kinoshita, and S. Araki, "Few-shot learning of new sound classes for target sound extraction," in *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 3500–3504.

[5] D. Kim, M. Baek, Y. Kim, and J. Chang, "Improving target sound extraction with timestamp knowledge distillation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024, pp. 1396–1400.

[6] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 1441–1445.

[7] B. Gfeller, D. Roblek, and M. Tagliasacchi, "One-shot conditional audio filtering of arbitrary sounds," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 501–505.

[8] C. Li, Y. Qian, Z. Chen, D. Wang, T. Yoshioka, S. Liu, Y. Qian, and M. Zeng, "Target sound extraction with variable cross-modality clues," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5.

[9] R. Gao and K. Grauman, "Co-separating sounds of visual objects," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 3878–3887.

[10] B. Veluri, J. Chan, M. Itani, T. Chen, T. Yoshioka, and S. Gollakota, "Real-time target sound extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[12] Q. Zhang, M. Tao, and Y. Chen, "gddim: Generalized denoising diffusion implicit models," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[13] H. Wang, J. Villalba, L. Moro-Velazquez, J. Hai, T. Thebaud, and N. Dehak, "Noise-robust speech separation with fast generative correction," *arXiv preprint arXiv:2406.07461*, 2024, unpublished.

[14] J. Hai, H. Wang, D. Yang, K. Thakkar, N. Dehak, and M. Elhilali, "DPM-TSE: A diffusion probabilistic model for target sound extraction," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*. IEEE, 2024, pp. 1196–1200.

[15] G. Mariani, I. Tallini, E. Postolache, M. Mancusi, L. Cosmo, and E. Rodolà, "Multi-source diffusion models for simultaneous music generation and separation," in *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[16] N. Kamo, M. Delcroix, and T. Nakatani, "Target speech extraction with conditional diffusion model," in *24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023*, N. Harte, J. Carson-Berndsen, and G. Jones, Eds. ISCA, 2023, pp. 176–180.

[17] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, "Separate anything you describe," *CoRR*, vol. abs/2308.05037, 2023, unpublished.

[18] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "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)*. IEEE, 2023, pp. 1–5.

[19] Z. Kong, S.-g. Lee, D. Ghosal, N. Majumder, A. Mehrish, R. Valle, S. Poria, and B. Catanzaro, "Improving text-to-audio models with synthetic captions," *arXiv preprint arXiv:2406.15487*, 2024.

[20] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra, "Freesound datasets: A platform for the creation of open audio datasets," in *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 486–493.

[21] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 4172–4182.

[22] J. Hai, Y. Xu, H. Zhang, C. Li, H. Wang, M. Elhilali, and D. Yu, "Ezaudio: Enhancing text-to-audio generation with efficient diffusion transformer," *CoRR*, vol. abs/2409.10819, 2024.

[23] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.

[24] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[25] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 2023, pp. 1–5.

[26] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 22 669–22 679.

[27] J. Su, M. H. M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, vol. 568, p. 127063, 2024.

[28] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *Scenes and Events 2018 Workshop (DCASE2018)*, 2018, p. 9.

[29] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 721–725.

[30] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 776–780.

[31] M. Chinen, F. S. Lim, J. Skoglund, N. Gureev, F. O'Gorman, and A. Hines, "Visqol v3: An open source production ready objective speech and audio metric," in *2020 twelfth international conference on quality of multimedia experience (QoMEX)*. IEEE, 2020, pp. 1–6.

[32] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. P. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," in *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 21 450–21 474.

[33] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 2880–2894, 2020.