IA321 - Summary Report
AudioLDM: Text-to-Audio Generation with Latent Diffusion Models

Pin Jin, 14 janvier 2025

I. Brief introduction

This paper introduces "AudioLDM", a novel system for text-to-audio (TTA) generation, motivated by the need for high-quality, computationally efficient, and versatile TTA systems. Traditional methods often rely on discrete representations of audio signals or paired audio-text datasets, which suffer from quality and scalability issues. Inspired by advancements in latent diffusion models (LDM) in image generation and the Contrastive Language-Audio Pretraining (CLAP) model for cross-modal embeddings, AudioLDM leverages continuous latent representations learned from audio signals without requiring paired datasets for training. The system achieves state-of-the-art TTA performance with enhanced sample quality, reduced computational costs, and the ability to perform zero-shot text-guided audio manipulations like style transfer and inpainting. The official implementation and demos are available at https://audioldm.github.io.

II. Text-Conditional Audio Generation

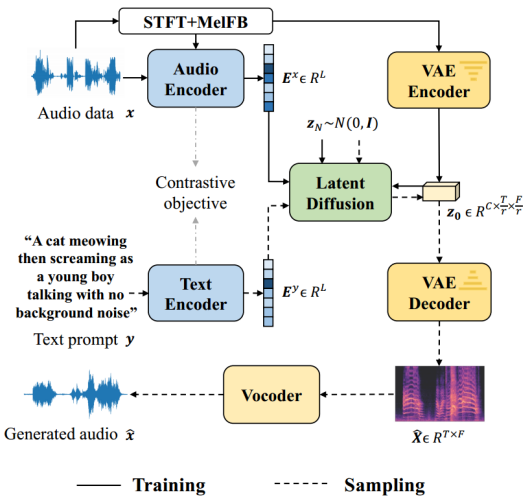

Fig. 1. Training and sampling process of AudioLDM

A. Contrastive Language-Audio Pretraining (CLAP)

Inspired by Contrastive Language-Image Pretraining (CLIP) [1], authors leverage Contrastive Language-Audio Pretraining [2] to enhance Text-to-Audio (TTA) generation. They employ a RoBERTa-based encoder for text [3] and an HTSAT-based encoder for audio [4] to generate embeddings E^y and E^x respectively within a shared embedding space of dimension L. Training by the symmetric cross-entropy loss, the CLAP model effectively aligns audio and text embeddings, demonstrating strong generalization capabilities across various downstream tasks, including zero-shot audio classification, and facilitates the integration of cross-modal information for previously unseen language or audio samples.

B. Conditional Latent Diffusion Model (LDM)

The system generates an audio sample $\hat{x}$ from a text description y by estimating the true conditional data distribution $q(z_0 \mid E^y)$ with a model distribution $p_\theta(z_0 \mid E^y)$. Here, z_0 represents the prior of an audio sample in the compressed mel-spectrogram space, and E^y is the text embedding obtained from the pretrained CLAP text encoder. The diffusion process [5] consists of a forward process that transforms the data distribution into a standard Gaussian distribution using a predefined noise schedule, and a reverse process that generates data samples from noise based on an inference noise schedule. The model is optimized using a reweighted noise estimation objective: $L_n(\theta) = \mathbb{E}_{z_0,\epsilon,n} \|\epsilon - \epsilon_\theta(z_n, n, E^x)\|_2^2$. By leveraging the shared cross-modal embedding space provided by CLAP, the LDM effectively generates audio priors conditioned on text embeddings without requiring text supervision during training, thereby facilitating robust TTA generation.

C. Conditioning Augmentation

To overcome the problem of comparatively limited language-audio datasets, the authors propose a data augmentation strategy that operates solely on audio embeddings E^x during the training of LDMs. They implement a mixup augmentation on audio samples x_1 and x_2 using the equation: $x_{1,2} = \lambda x_1 + (1-\lambda)x_2$, where λ is a scaling factor sampled from a Beta distribution $\mathcal{B}(5,5)$ [6]. It is also applied in section II-E to improve the reconstruction effect. This approach increases the number of training data pairs (z_0, E^x), enhancing the robustness of LDMs to CLAP embeddings without the need to augment language-audio pairs.

D. Classifier-Free Guidance (CFG)

CFG is presented as a state-of-the-art technique that guides diffusion models by modifying the noise estimation during the sampling process [7]. During training, the conditioning information E^x is randomly discarded with a fixed probability (e.g., 10%), allowing the model to learn both conditional $\epsilon_\theta(z_n, n, E^x)$ and unconditional $\epsilon_\theta(z_n, n)$ denoising functions. In the generation phase, the text embedding E^y is used as a condition, and the noise estimation is adjusted using

a guidance scale w as follows: $\hat{\epsilon}_\theta(z_n, n, E^y) = w\epsilon_\theta(z_n, n) + (1-w)\epsilon_\theta(z_n, n, E^y)$. This approach allows for more precise control over the generation process by leveraging detailed text descriptions.

E. Decoder

They employ a Variational Autoencoder (VAE) to compress the mel-spectrogram $X \in R^{T\times F}$ into a compact latent space $z \in R^{C\times\frac{T}{r}\times\frac{F}{r}}$, where r represents the compression level. The VAE architecture consists of an encoder and a decoder, both utilizing stacked convolutional modules. The training objective integrates a reconstruction loss, an adversarial loss, and a Gaussian constraint loss to ensure high-quality reconstructions. During the sampling phase, the decoder reconstructs the mel-spectrogram $\hat{X}$ from the audio prior $\hat{z}_0$ generated by the LDMs.

III. Text-Guided Audio Manipulation

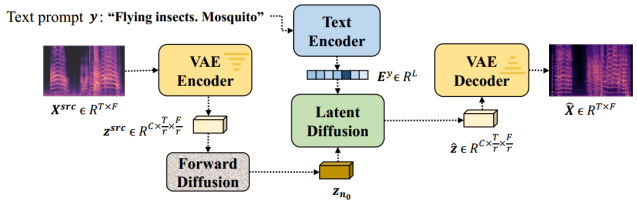

Fig. 2. Audio style transfer with AudioLDM

A. Style Transfer

Given a source audio sample x^{src}, its noisy latent representation z_{n_0} can be computed using a predefined time step $n_0 \leq N$ following the forward diffusion process. By initiating the reverse diffusion process from z_{n_0} with a pretrained AudioLDM model, the system manipulates the audio based on a text input y, controlled by the time step n_0: $p_\theta(z_{0:n_0} \mid E^y) = p(z_{n_0}) \prod_{n=1}^{n_0} p_\theta(z_{n-1} \mid z_n, E^y)$. A larger n_0 results in more significant manipulation, resembling standard TTA generation.

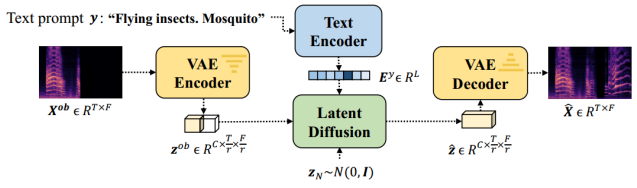

Fig. 3. Audio inpainting with AudioLDM

B. Inpainting and Super-Resolution

Authors address the generation of missing audio segments given observed parts x^{ob}. This is achieved by integrating the observed latent representation z^{ob} into the generated latent space z during the reverse diffusion process. Specifically, at each inference step, the generated latent z_{n-1} is modified using an observation mask m as follows: $z'_{n-1} = (1-m) \odot z_{n-1} + m \odot z^{\text{ob}}_{n-1}$, where m indicates the observed regions in the latent space. The convolutional structure of the VAE ensures spatial correspondence in the mel-spectrogram, allowing precise retention of observed time-frequency bins.

IV. Experiment

A. Datasets and Metrics

The experiments leverage four primary datasets to train the Text-to-Audio models: AudioSet (AS), AudioCaps (AC), FreeSound (FS), and BBC Sound Effect library (SFX). The evaluation is conducted on both AC and AS, using text descriptions derived from captions and labels for broader sound categories.

Objective evaluation metrics include Frechet Distance (FD), Inception Score (IS), Kullback–Leibler (KL) divergence, and Frechet Audio Distance (FAD), complemented by subjective assessments from audio professionals rating overall quality (OVL) and relevance (REL).

Baseline comparisons are made against DiffSound and AudioGen, and the authors train two versions of their proposed models, AudioLDM-S and AudioLDM-L, primarily on the AC dataset, with an extended model named AudioLDM-X-Full on all 4 datasets, exploring the effect of the scale of training data.

B. Results analysis

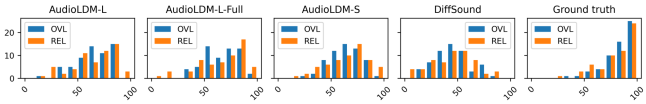

Fig. 4. The histogram of the human evaluation result.

1) Main Comparaison: The main evaluation results on the AC test set are shown in Figure 5. Given the single training dataset AC, AudioLDMS can achieve better generation results than the baseline models on both objective and subjective evaluations, even with smaller model size. By expanding model capacity with AudioLDM-L, we further improve the overall results. Then, by incorporating AS and the two other datasets into training, our model AudioLDM-L-Full achieves the best quality, with an FD of 23.31. This success is largely attributed to using CLAP, which decouples learning the audio-text relationship from the generative process by aligning their embeddings beforehand. The human evaluation shows a trend similar to other evaluation metrics, with AduioLDM method achieving overall quality and relevance scores around 64, significantly higher than DiffSound's 45.00 OVL and 43.83 REL. Larger models yield better audio quality, and scaling up the training data further boosts both OVL and REL. Additionally, score distributions, as shown in Figure 4, indicate that AduioLDMs concentrate on higher ratings compared to DiffSound, and high scores for randomly selected real recordings confirm the reliability of the evaluations.

Model	Text Data	Use CLAP	Params	Duration (h)	FD ↓	IS ↑	KL ↓	FAD ↓	OVL ↑	REL ↑
Ground truth	-	-	-	-	-	-	-	-	$83.61_{\pm 1.1}$	$80.11_{\pm 1.2}$
DiffSound† (Yang et al., 2022)	✓	✗	400M	5420	47.68	4.01	2.52	7.75	$45.00_{\pm 2.6}$	$43.83_{\pm 2.3}$
AudioGen† (Kreuk et al., 2022)	✓	✗	285M	8067	-	-	2.09	3.13	-	-
AudioLDM-S-Full-RoBERTa	✓	✗	181M	145	32.13	4.02	3.25	5.89	-	-
AudioLDM-S	✗	✓	181M	145	29.48	6.90	1.97	2.43	$63.41_{\pm 1.4}$	$64.83_{\pm 0.9}$
AudioLDM-L	✗	✓	739M	145	27.12	7.51	1.86	2.08	$64.30_{\pm 1.6}$	$64.72_{\pm 1.6}$
AudioLDM-S-Full	✗	✓	181M	8886	23.47	7.57	1.98	2.32	-	-
AudioLDM-L-Full	✗	✓	739M	8886	**23.31**	**8.13**	**1.59**	**1.96**	$\mathbf{65.91}_{\pm 1.0}$	$\mathbf{65.97}_{\pm 1.6}$

Fig. 5. The comparison between AudioLDM and baseline TTA generation models

Model	Text	Audio	FD ↓	IS ↑	KL ↓
AudioLDM-S	✓	✓	31.26	6.35	2.01
AudioLDM-S	✗	✓	**29.48**	**6.90**	**1.97**
AudioLDM-S-Full	✓	✓	27.20	7.52	2.38
AudioLDM-S-Full	✗	✓	**23.47**	**7.57**	**1.98**
AudioLDM-L-Full	✓	✓	25.79	7.95	2.26
AudioLDM-L-Full	✗	✓	**23.31**	**8.13**	**1.59**

Fig. 6. The comparison between text embedding and audio embedding as conditioning information on the training of LDMs

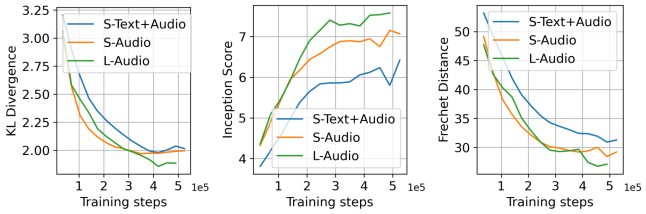

Fig. 7. The comparison in the training process

2) Conditioning Information: As shown in Figure 6, using audio embeddings E_x as conditioning information for training Latent Diffusion Models (LDMs) yields superior results compared to using text embeddings E_y. In experiments where both setups employed similar data augmentation strategies, models trained with E_x outperformed those trained with E_y. The primary reason is that text embeddings often struggle to capture the complexity and nuances of sounds due to the inherent ambiguity and subjectivity in human descriptions. In contrast, E_x is directly extracted from the audio signal, aligning closely with the most accurate text descriptions and providing more reliable and robust conditioning information to the LDMs. Figure 7 illustrates that training with audio embeddings consistently leads to higher sample quality throughout the training process. Additionally, while larger models may converge more slowly, they ultimately achieve better performance. This indicates that scaling up model size in conjunction with audio-based conditioning contributes to improved final results.

3) Text-Guided Audio Manipulation: The study demonstrates text-guided audio manipulation using AudioLDM for super-resolution (upsampling from 8 kHz to 16 kHz) and inpainting tasks. As shown in Figure 8, while AudioLDM

Task	Super-resolution		Inpainting
Dataset	AudioCaps	VCTK	AudioCaps
Unprocessed	2.76	2.15	10.86
Kuleshov et al. (2017)	-	1.32	-
Liu et al. (2022a)	-	**0.78**	-
AudioLDM-S	1.59	1.12	2.33
AudioLDM-L	**1.43**	0.98	**1.92**

Fig. 8. Performance comparisons on zero-shot super-resolution and inpainting, which are evaluated by LSD and FAD, respectively.

outperforms the AudioUNet baseline in super-resolution, its results fall short of NVSR, partly due to artifacts like white noise from training on diverse, noisy signals. Despite this, AudioLDM establishes a promising zero-shot approach for text-guided audio manipulation, suggesting potential for further improvements.

Setting	FD↓	IS↑	KL↓	OVL ↑	REL ↑
AudioLDM-S	**29.48**	**6.90**	**1.97**	63.41	64.83
w. Simple attn	33.12	6.15	2.09	-	-
w. Balance samp	34.05	6.21	2.16	-	-
w. Cond aug	31.88	6.25	2.02	**64.49**	**65.01**

Fig. 9. The ablation study on the attention mechanism, the balance sampling technique, and the conditioning augmentation algorithm.

DDIM steps	10	25	50	100	200
FD	55.84	42.84	35.71	30.17	**29.48**
IS	4.21	5.91	6.51	6.85	**6.90**
KL	2.47	2.12	2.01	**1.94**	1.97

Fig. 10. Effect of sampling steps of LDMs with a DDIM sampler

4) Ablation Study: The ablation study on AudioLDM-S, as shown in Figure 9, highlights that employing a complex attention mechanism is crucial, as simplifying it or using balanced sampling strategies reduces performance, while conditioning augmentation improves subjective evaluations. Figure 10 indicates that increasing DDIM sampling steps enhances audio quality, with diminishing returns after about 100 steps. Besides, a suitable guidance scale strikes a balance between conditional generation quality and sample diversity.

V. Code Testing (Optional)

Based on the public official code base of the paper "AudioLDM" (https://github.com/haoheliu/AudioLDM), I conducted tests on the code similar to those in the paper.

I put the test and generation code I wrote at the following address: https://github.com/SebastienJin/AudioLDM, which will be updated if new changes follow.

A. Dataset

AudioCaps (https://github.com/cdjkim/audiocaps) is a large-scale dataset paired with human-written textual descriptions, which were collected via crowdsourcing based on the AudioSet dataset. However, instead of providing the complete audio data directly, the database provides the YouTube video ID that the audio belongs to and the start time of the clip. Therefore, we need to first find a way to download the corresponding audio.

Most of the methods searched on github (for example, https://github.com/MorenoLaQuatra/audiocaps-download) require downloading and installing ffmpeg for video and audio conversion and streaming, but its installation requires sudo privileges, which makes it impossible to download it quickly on the server.

However, it is discovered that a team of Japanese researchers provide all the audio they download at the time based on the original database, which can be found at the following address: https://github.com/sarulab-speech/ml-audiocaps. Due to the disappearance of some of the youtube videos, the size of the test dataset is reduced from the original 4875 to 4420 audio.

B. Model Evaluation

Due to computing resource and time constraints, I cannot train the original model from scratch, so in the relevant tests, I will directly use the model checkpoints published by the authors in the code base. As described, we have totally 6 checkpoints: audioldm-m-text-ft, audioldm-s-text-ft, audioldm-m-full, audioldm-s-full, audioldm-l-full, and audioldm-s-full-v2.

In fact, since the source code only provides the model, we also need to write a matched set of pipelines for the computation of the various metrics of the model on the test set. We use the VGGish model to capture audio features for both real and generated audio, and compute the corresponding Frechet Distance, Inception Score, and Kullback-Leibler divergence based on the features.

C. Pipline Pseudocode

for all $audio \in TestDataset$ **do**
  if exists($audio$) **then**
    $real\_features \leftarrow VGG(real\_audio)$
    $generated\_audio \leftarrow AudioLDM(Prompt)$
    $gen\_features \leftarrow VGG(generated\_audio)$
    frechet_distance($real\_features, gen\_features$)
    inception_score($real\_features, gen\_features$)
    kl_divergence($real\_features, gen\_features$)
Compute averages for FD, IS, and KL

D. Problems Encountered

Generation is too slow! I ran the generation test following the code described in the original paper library, and it took about 25 minutes to generate an audio without using the GPU! This made it difficult to complete the test on the entire test set at all. It wouldn't reasonably be that long, but I've modified the code to no avail.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] 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.

[3] Z. Liu, W. Lin, Y. Shi, and J. Zhao, "A robustly optimized bert pre-training approach with post-training," in *China National Conference on Chinese Computational Linguistics*. Springer, 2021, pp. 471–484.

[4] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

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

[6] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3292–3306, 2021.

[7] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.