

Supplementary Materials: Efficient and High-Quality Text-to-Audio Generation with Minimal Inference Steps

Huadai Liu
Zhejiang University
Shanghai Artificial Intelligence
Laboratory
liuhuadai@zju.edu.cn

Rongjie Huang
Zhejiang University
rongjiehuang@zju.edu.cn

Yang Liu
Zhejiang University
22160155@zju.edu.cn

Hengyuan Cao
Zhejiang University
caohy@zju.edu.cn

Jialei Wang
Zhejiang University
3220101016@zju.edu.cn

Xize Cheng
Zhejiang University
xizecheng@zju.edu.cn

Siqi Zheng
Alibaba Group
zsq174630@alibaba-inc.com

Zhou Zhao[†]
Zhejiang University
Shanghai Artificial Intelligence
Laboratory
zhaozhou@zju.edu.cn

1 DATASET DESCRIPTIONS

As shown in Table 2, we collect a large-scale audio-text dataset consisting of 0.92 million of audio samples with a total duration of approximately 3.7k hours. For text-to-music generation, we use the LP-MusicCaps for training. For Clotho dataset, we only use its evaluation set for zero-shot testing and do not use for training. For text-to-audio generation, we filter 95% of the samples that contain speech and music to build a more balanced dataset, as speech and music are the dominant classes in AudioSet.

2 ARCHITECTURE

We list the model hyper-parameters of AudioLCM in Table 1.

Hyperparameter		AudioLCM
Spectrogram Autoencoders	Input/Output Channels	80
	Hidden Channels	20
	Residual Blocks	2
	Spectrogram Size	80 × 624
	Channel Mult	[1, 2, 4]
Transformer Backbone	Input shape	(20, T)
	Condition Embed Dim	1024
	Feed-forward Hidden Size	576
	Transformer Heads	8
	Transformer Blocks	8
	Sampling Steps	2
CLAP Text Encoder	Transformer Embed Channels	768
	Output Project Channels	1024
	Token Length	77
Total Number of Non-trainable Parameters		984M
Total Number of Trainable Parameters		159M

Table 1: Hyperparameters of AudioLCM models.

3 EVALUATION

3.1 Subjective evaluation

To assess the generation quality, we conduct MOS (Mean Opinion Score) tests regarding audio quality and text-audio faithfulness, respectively scoring MOS-Q and MOS-F.

For audio quality, the raters were explicitly instructed to “focus on examining the audio quality and naturalness.” The testers were presented with audio samples and asked to rate their subjective score (MOS-P) on a 20-100 Likert scale.

For text-audio faithfulness, human raters were shown the audio and its caption and asked to respond to the question, “Does the natural language description align with the audio faithfully?” They had to choose one of the options - “completely,” “mostly,” or “somewhat” on a 20-100 Likert scale.

Our crowd-sourced subjective evaluation tests were conducted via Amazon Mechanical Turk where participants were paid \$8 hourly. A small subset of the generated audio samples used in the test can be found at <https://Echo-Audio.github.io/>.

3.2 Objective evaluation

Fréchet Audio Distance (FAD) [7] is adapted from the Fréchet Inception Distance (FID) to the audio domain, it is a reference-free perceptual metric that measures the distance between the generated and ground truth audio distributions. FAD is used to evaluate the quality of generated audio.

KL divergence is measured at a paired sample level between the generated audio and the ground truth audio, it is computed using the label distribution and is averaged as the final result.

CLAP score: adapted from the CLIP score [5, 14] to the audio domain and is a reference-free evaluation metric to measure audio-text alignment for this work that closely correlates with human perception.

4 DETAILED FORMULATION OF DDPM

As a blossoming class of generative models, denoising diffusion probabilistic models (DDPMs) [6, 16] has emerged to prove its capability to achieve leading performances in both image and audio syntheses [9, 17]. We define the data distribution as $q(\mathbf{x}_0)$. The diffusion process is defined by a fixed Markov chain from data \mathbf{x}_0 to the latent variable \mathbf{x}_T :

$$q(\mathbf{x}_1, \dots, \mathbf{x}_T | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}), \quad (1)$$

For a small positive constant β_t , a small Gaussian noise is added from \mathbf{x}_{t-1} to the distribution of \mathbf{x}_t under the function of $q(\mathbf{x}_t | \mathbf{x}_{t-1})$.

The whole process gradually converts data \mathbf{x}_0 to whitened latents \mathbf{x}_T according to the fixed noise schedule β_1, \dots, β_T , where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (2)$$

Efficient training is optimizing a random term of t with stochastic gradient descent:

$$\mathcal{L}_\theta = \left\| \epsilon_\theta \left(\alpha_t \mathbf{x}_0 + \sqrt{1 - \alpha_t^2} \epsilon \right) - \epsilon \right\|_2^2 \quad (3)$$

Unlike the diffusion process, the reverse process is to recover samples from Gaussian noises. The reverse process is a Markov chain from \mathbf{x}_T to \mathbf{x}_0 parameterized by shared θ :

$$p_\theta(\mathbf{x}_0, \dots, \mathbf{x}_{T-1} | \mathbf{x}_T) = \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t), \quad (4)$$

where each iteration eliminates the Gaussian noise added in the diffusion process:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)^2 \mathbf{I}) \quad (5)$$

5 IMPLEMENTATION DETAILS

5.1 Ablation Studies on Novel Transformer Backbone

We further investigate the efficacy of the LLaMA architectures on the AudioLCM model applied to the audiocaps dataset, as depicted in Table 3. Remarkably, all incorporated designs within the transformer architecture exhibit enhancements in the performance of our AudioLCM system, with particular prominence observed in the case of Rotary Embeddings.

5.2 Analyses about Scalable Transformer

During the training phase of the original diffusion transformer, we encountered instability issues, particularly evident when utilizing low-bit training models and scaling up parameters. We investigate the performance of a novel transformer backbone designed to scale up the trainable parameters, as showcased in Table 4. The outcomes suggest a notable improvement in AudioLCM's performance following parameter scaling-up, underscoring the potential of scaled-up transformers for augmenting system performance.

6 POTENTIAL NEGATIVE SOCIETAL IMPACTS

This paper aims to achieve efficient and high-quality text-to-audio generation, which makes generative models practically applicable to text-to-audio generation deployment. A negative impact is the risk of misinformation. To alleviate it, we can train an additional classifier to discriminate the fakes. We believe the benefits outweigh the downsides.

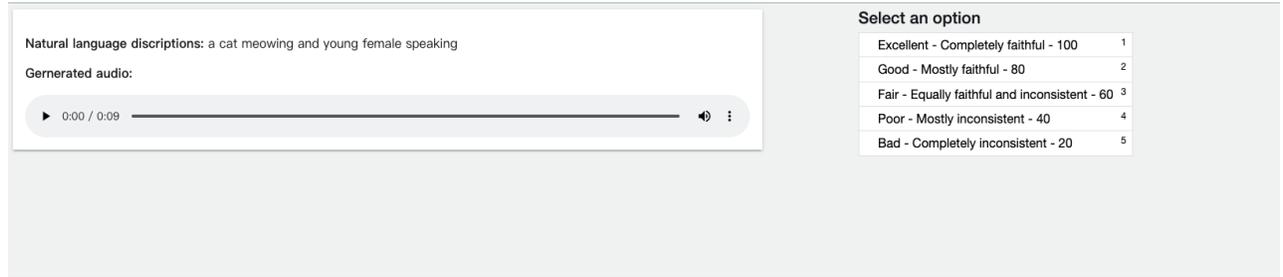
AudioLCM lowers the requirements for fast and high-quality text-to-audio synthesis, which may cause unemployment for people with related occupations, such as sound engineers and radio hosts. In addition, there is the potential for harm from non-consensual voice cloning or the generation of fake media, and the voices in the recordings might be overused than they expect.

7 LIMITATIONS

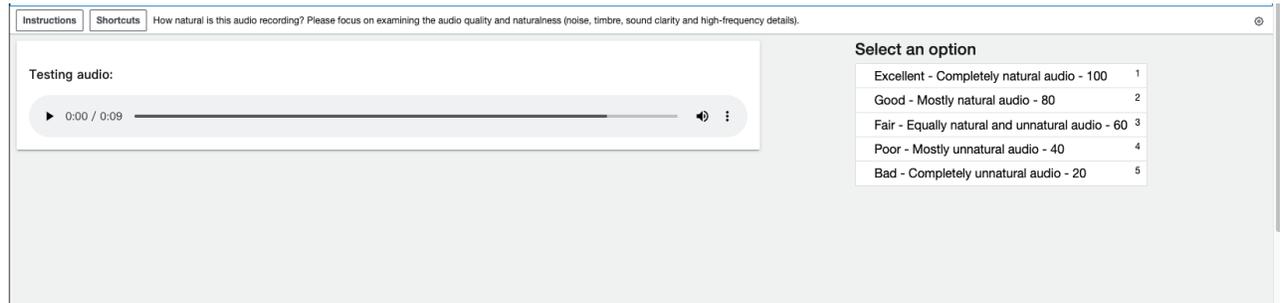
AudioLCM adopts multi-step consistency sampling which evaluate AudioLCM multiple times by alternating denoising and noise injection steps for improved sample quality. However, this enhancement disappeared when sampling steps is larger than ten steps. It is because the discretisation errors accumulated during the sampling phase. One of our future directions is to solve this problem.

Dataset	Hours	Type	Source
Audiocaps	109hrs	caption	[8]
WavCaps	2056hrs	caption	[11]
WavText5K	25hrs	caption	[1]
MACS	48hrs	caption	[10]
Clothv2	152hrs	caption	[3]
Audiostock	44hrs	caption	https://audiostock.net
epidemic sound	220hrs	caption	https://www.epidemicsound.com
Adobe Audition Sound Effects	26hrs	caption	https://www.adobe.com/products/audition/offers/AdobeAuditionDLCSEFX.html
LP-MusicCaps	5673hrs	caption	[2]
FSD50K	108hrs	label	https://annotator.freesound.org/fsd
ODEON_Sound_Effects	20hrs	label	https://www.paramountmotion.com/odeon-sound-effects
UrbanSound8K	9hrs	label	[15]
ESC-50	3hrs	label	[13]
filteraudioset	945hrs	multi label	[4]
TUT	13hrs	label	[12]

Table 2: Statistics for the Datasets used in the paper.



(a) Screenshot of MOS-F testing.



(b) Screenshot of MOS-Q testing.

Figure 1: Screenshots of subjective evaluations.

Method	FAD	KL
AudioLCM	1.67	1.37
w/o RoPE	1.80	1.45
w/o RMSNorm	1.74	1.41
w/o SwiGLU	1.78	1.40

Table 3: Comparison of audio quality in the ablation study with LLaMA designs. RoPE denotes Rotary Embeddings.

Model	Parameters	FAD	KL
AudioLCM-B	159M	1.67	1.37
AudioLCM-M	561M	1.62	1.33
AudioLCM-L	996M	1.53	1.30
AudioLCM-XL	2.4B	1.50	1.28

Table 4: AudioLCM-B, AudioLCM-M, AudioLCM-L, and AudioLCM-XL respectively represent the base, medium, large, and extra large models of AudioLCM. The presented figures only account for trainable parameters, i.e., those within the transformer architecture, evaluated on AudioCaps.

349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406

407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464

REFERENCES

- [1] Soham Deshmukh, Benjamin Elizalde, and Huaming Wang. 2022. Audio retrieval with wavtext5k and clap training. *arXiv preprint arXiv:2209.14275* (2022).
- [2] SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. LP-MusicCaps: LLM-Based Pseudo Music Captioning. arXiv:2307.16372 [cs.SD]
- [3] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. [n. d.]. Clotho: An Audio Captioning Dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020)*. 736–740.
- [4] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. [n. d.]. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2017)*. 776–780.
- [5] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*.
- [6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. [n. d.]. Denoising Diffusion Probabilistic Models. In *Proc. of NeurIPS* (2020).
- [7] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2018. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. *arXiv preprint arXiv:1812.08466* (2018).
- [8] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. [n. d.]. AudioCaps: Generating Captions for Audios in the Wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (2019). 119–132.
- [9] Huadai Liu, Rongjie Huang, Xuan Lin, Wenqiang Xu, Maozong Zheng, Hong Chen, Jinzheng He, and Zhou Zhao. 2023. Vit-tts: visual text-to-speech with scalable diffusion transformer. *arXiv preprint arXiv:2305.12708* (2023).
- [10] Irene Martin-Morató and Annamaria Mesaros. [n. d.]. What Is the Ground Truth? Reliability of Multi-Annotator Data for Audio Tagging. In *2021 29th European Signal Processing Conference (EUSIPCO) (2021)*. 76–80.
- [11] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. [n. d.]. WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research. arXiv:2303.17395 [eess.AS]
- [12] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *2016 24th european signal processing conference (EUSIPCO)*. 1128–1132. <https://doi.org/10.1109/EUSIPCO.2016.7760424>
- [13] Karol J Piczak. [n. d.]. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia* (2015). 1015–1018.
- [14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 [cs.CV]
- [15] J. Salamon, C. Jacoby, and J. P. Bello. [n. d.]. A Dataset and Taxonomy for Urban Sound Research. In *22nd ACM International Conference on Multimedia (ACM-MM'14)* (Orlando, FL, USA, 2014-11). 1041–1044.
- [16] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising Diffusion Implicit Models. In *Proc. of ICLR*.
- [17] Jiahao Xun, Shengyu Zhang, Zhou Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiuqiang He, Xiaofei He, Tat-Seng Chua, and Fei Wu. 2021. Why do we click: visual impression-aware news recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3881–3890.