Generative Audio Language Modeling with Continuous-valued Tokens and Masked Next-Token Prediction

Shu-wen Yang [1 2] **Byeonggeun Kim** [* 2] **Kuan-Po Huang** [* 1 2] **Qingming Tang** [2] **Huy Phan** [2] **Bo-Ru Lu** [2] **Harsha Sundar** [2] **Shalini Ghosh** [2] **Hung-yi Lee** [1] **Chieh-Chi Kao** [2] **Chao Wang** [2]

Abstract

Autoregressive next-token prediction with the Transformer decoder has become a de facto standard in large language models (LLMs), achieving remarkable success in Natural Language Processing (NLP) at scale. Extending this paradigm to audio poses unique challenges due to its inherently continuous nature. We research audio generation with a causal language model (LM) without discrete tokens. We leverage token-wise diffusion to model the continuous distribution of the next continuous-valued token. Our approach delivers significant improvements over previous discrete solution, AudioGen, achieving 20% and 40% relative gains on AudioCaps in Frechet Audio Distance (FAD) and Kullback-Leibler (KL) divergence, respectively. Additionally, we propose a novel masked next-token prediction task that incorporates masked prediction into the causal LM framework. On AudioCaps, the innovation yields 41% and 33% relative FAD improvements over AudioGen Base (285M) and AudioGen Large (1B) models, respectively, and is on par with the state-of-the-art (SOTA) diffusion models. Furthermore, we achieve these results with significantly fewer parameters—193M for our Base and 462M for our Large models.

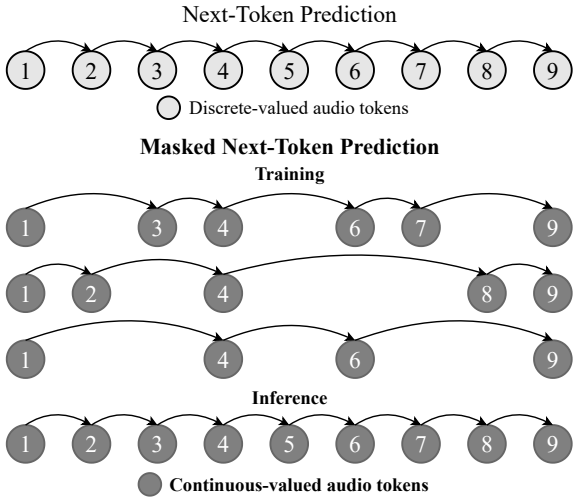


Figure 1. **Causal Language Modeling on Continuous-valued Audio Tokens with Masked Next-Token Prediction.** The audio tokens are low-dimensional continuous latent. We use a standard Transformer decoder for the audio language modeling. Our masked next-token prediction learns to predict any future token given any subset of the past tokens, which turns out benefit the standard next-token prediction and rival the bidirectional diffusion models and masked generative models. The figure illustrates the prediction order in which all tokens are conditioned on all previously predicted tokens.

1. Introduction

Large Language Models (LLMs) has revolutionized artificial intelligence (Wang et al., 2018; Žagar & Robnik-Šikonja, 2022; Hendrycks et al., 2021). They show remarkable emergent capabilities and human-level understanding and reasoning abilities (Achiam et al., 2023; Team et al., 2023) after scaling up to thousands of billions of parameters, such as the GPT series (Brown et al., 2020; Achiam et al., 2023). The essence behind LLMs is simple, *scaling up the model by more parameters and data with a Transformer decoder (Vaswani, 2017) and next-token prediction*. Given the success of this paradigm, plenty of efforts have been put into developing better theories and tools for this standardized model, including the architecture improvement (Chowdhery et al., 2023; Du et al., 2022) scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022), and efficient inference and streaming infrastructure (Kwon et al., 2023; Pope et al., 2023).

[*]Equal contribution [1]Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan [2]Amazon AGI, Bellevue, United States. Correspondence to: Shu-wen Yang <leo19941227@gmail.com>, Chieh-Chi Kao <chiehchi@amazon.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

We study text-to-audio (TTA) with the LLM framework. TTA is critical due to its various applications in virtual (augmented) reality, media creation, video editing, and game development (Liu et al., 2023a). Specifically, given a natural language prompt, the model generates a relevant sound. The causal language modeling approach is important to audio for its fundamental role in multi-modal LLMs (Wu et al., 2024; Yin et al., 2023; Liu et al., 2024b; Zhou et al., 2024) and the streaming capability for real-time user interaction (Wang et al., 2024a; Zhang et al., 2024). Hence, we ask *whether a causal language model can achieve high-fidelity audio generation.* Existing works applied a Transformer decoder with next-token prediction on discrete audio tokens for TTA (Kreuk et al., 2023; Yang et al., 2023a). Despite scaling to billions of parameters, these models lag behind the SOTA systems, which typically rely on diffusion techniques (Majumder et al., 2024; Liu et al., 2024c). As a result, to build an LLM that can speak or produce high-fidelity sound, the prevailing approach relies on external vocoders (Huang et al., 2024; Liu et al., 2023b; Wang et al., 2024b). However, the pipeline approach overlooks the possibility of directly generating audio with a scalable and streamable LLM, allowing us to enjoy the aforementioned resources like scaling, efficient inference and streaming infrastructure.

In this work, we enable audio generation for causal language models. We address the bottlenecks that block the decoder-only solution by integrating strengths from SOTA diffusion models (Majumder et al., 2024; Liu et al., 2024c), including token types, loss functions, and learning tasks. We reinterpret the variational auto-encoder (VAE) latents in latent diffusion models (LDMs) (Rombach et al., 2022) as *continuous-valued tokens*[1], replacing discrete acoustic tokens used in audio language modeling. To model the continuous distribution of the next token, we replace the cross-entropy loss by the token-wise diffusion loss (Li et al., 2024b), leaving the backbone Transformer decoder intact, allowing it to fully leverage the benefits and resources of LLMs. Using continuous-valued **audio** tokens and **next-token prediction (NTP)**, our 193M **AudioNTP Base** achieves 20% and 40% relative improvements over 285M AudioGen Base (Kreuk et al., 2023) on AudioCaps (Kim et al., 2019) in Frechet Audio Distance (FAD) and Kullback-Leibler (KL) divergence, respectively.

To further match the performance of SOTA diffusion models, we incorporate masked language modeling (MLM) and propose a novel learning task. MLM learns contextualized dependencies (Kenton & Toutanova, 2019; Vyas et al., 2023; Liu et al., 2024a; Li et al., 2024b; Chang et al., 2022) and develops understanding capabilities (Li et al., 2024a; 2023; Wei et al., 2023), both of which have been shown to benefit generation across various scenarios. Concretely, we randomly drop tokens as the masking, forming a shorter sequence. Next-token prediction is then performed on this shorter sequence as the masked prediction, a task we term **masked next-token prediction (MNTP)**. Compared to next-token prediction, MNTP predicts a random future token conditioned on a random subset of past tokens (Figure 1), a task that ultimately benefits next-token prediction. On AudioCaps, our **AudioMNTP Base** significantly outperforms AudioNTP Base, achieving a 41% relative FAD improvement over AudioGen Base. Scaling up to 462M **AudioMNTP Large** yields a 33% relative FAD improvement over 1B AudioGen Large, matching SOTA diffusion models (Liu et al., 2024c; Majumder et al., 2024) while remaining streamable and compatible with LLMs. Our contributions are:

- We propose the use of continuous-valued tokens in generative audio language modeling, demonstrating their superiority over discrete tokens.
- We introduce a novel learning task, masked next-token prediction (MNTP). Training with MNTP enhances the model's decoding performance on the regular next-token prediction.
- By integrating continuous-valued tokens and MNTP, we achieve SOTA-level audio generation within the causal LM framework, establishing a pathway for directly generating high-fidelity audio with LLMs.

2. Related Work

Text-guided Audio Generation. The most relevant works include AudioGen (Kreuk et al., 2023) and UniAudio (Yang et al., 2023a). AudioGen learns a causal LM on discrete tokens (Zeghidour et al., 2021; Défossez et al., 2023) with low downsampling rate to preserve audio fidelity, whereas UniAudio employs a multi-scale Transformer (Yu et al., 2023) to reduce RVQ token sequence length, thereby saving computation. On the other hand, LDMs (Rombach et al., 2022) have outperformed language modeling approaches in audio generation (Yang et al., 2023b; Liu et al., 2024c; Majumder et al., 2024) by modeling the joint probability distribution of continuous-valued tokens through diffusion processes. Among them, we adopt a causal LM similar to AudioGen but apply it to continuous-valued tokens, as in LDMs.

Language Modeling on Continuous-valued Tokens. Language modeling on continuous data has traditionally required quantizing data into discrete tokens (Chen et al., 2020b; Esser et al., 2021; Borsos et al., 2023; Kreuk et al., 2023). Recent works in image and speech generation (Li

[1]Specifically, we use a VAE to encode the Mel-spectrogram into a 2-D feature map, and serialize it into a 1-D sequence of low-dimensional latents.

et al., 2024b; Tschannen et al., 2025; Meng et al., 2024) demonstrates that continuous-valued tokens prevent information loss and achieve higher generation quality. These approaches model continuous next-token distributions using diffusion loss (Li et al., 2024b) or Gaussian mixture models (Tschannen et al., 2025; Meng et al., 2024). In our work, we employ diffusion loss for sound.

Masked Prediction in Generative Models. Masked prediction has been shown effective in various generative models. In text generation, masked infilling serves as a pretext task for bidirectional models (Raffel et al., 2020; Lewis et al., 2020), and iterative MLM has been applied to image generation (Li et al., 2023; Chang et al., 2022). Similarly, masked infilling benefits audio (Vyas et al., 2023) and speech (Liu et al., 2024a) generation in the flow-matching (Lipman et al., 2023) framework. Recently, several methods (Fried et al., 2023; Aghajanyan et al., 2022; Peng et al., 2024; Bavarian et al., 2022) reorder masked tokens to the sequence end, enabling decoder-only models to leverage both past and future context for the editing tasks. In contrast, we integrate masked prediction without future context, strictly preserving causality and improving unidirectional decoding.

Relation to MAR. Our work is based on MAR (Li et al., 2024b), which introduces continuous-valued tokens in the context of masked generative modeling (MGM) for image generation. In MGM, a bidirectional model iteratively performs MLM from an all-zero input, predicting all masked positions but retaining only a subset, until all positions are retained. Our approach differs from MAR by using a standard Transformer decoder with the classic next-token prediction, making it compatible with LLMs. We show that, in the audio domain using continuous-valued tokens, the unidirectional MNTP outperforms naive next-token prediction and is comparable to the bidirectional MAR. Furthermore, in the left-to-right causal inference scenario, MNTP outperforms MAR.

3. Method

Our approach includes two main proposals: continuous-valued tokens in Section 3.1 and masked next-token prediction in Section 3.2.

3.1. AudioNTP: Continuous-valued Tokens

Common belief holds that next-token prediction is tied to a fixed size dictionary and cross-entropy loss. However, the task by definition is *predicting the next token given the previous tokens*, independent of the exact realization of the *token* and *how to model the next-token distribution*. Following the paradigm shift in CV (Tschannen et al., 2025; Li et al., 2024b), we investigate the use of continuous-valued tokens in audio language modeling. Figure 2 illustrates the framework of AudioNTP.

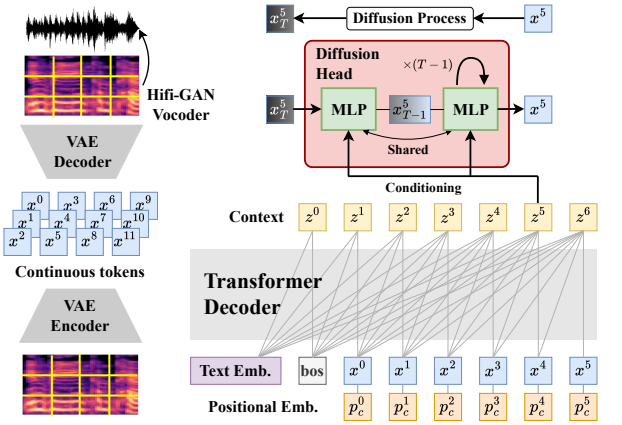


Figure 2. The framework of the continuous-valued audio token proposal. The waveform is transformed into the Mel-spectrogram, subsequently encoded into continuous-valued tokens. The Transformer decoder learns the next token prediction on these tokens via the token-wise diffusion loss with a small MLP diffusion head. This framework is termed **AudioNTP**.

Training. Given an input text prompt w and a length-s audio waveform $a = \{a^1, ..., a^s\}$, we tokenize it into a sequence of low-dimensional latents[2]: $x = \{x^1, ..., x^n\}$, where $x^i \in R^h$ and $h \in Z$ represents the latent dimension. We refer to this length-n, 1-D latent sequence x as the *continuous-valued tokens*. We train the language model with maximum likelihood on x over an audio corpus $D = \{a_j, w_j\}$. The loss function is then $\mathbb{E}_{(a,w)\sim D}\left[-log\ p(a \mid w)\right]$, where

$$p(a \mid w) = p(x^1, ..., x^n \mid w) = \prod_{i=1}^{n} p(x^i \mid x^1, ..., x^{i-1}, w) \tag{1}$$

To learn the next-token distribution $p(x^i \mid x^1, ..., x^{i-1}, w)$, our model consists of a Transformer decoder[3] C_θ and a MLP diffusion head M_ϕ, where θ and ϕ denote the parameters. Firstly, C_θ encodes the input sequence $\{w, \beta, x^1, ...x^{i-1}\}$ into a sequence of context vectors $z = \{z^1, ..., z^i\}$, where w is placed at the front as the input prompt and β is the BOS token: $z^i = C_\theta(w, \beta, x^1, ..., x^{i-1})$. A small multi-layer

[2] See Appendix F.1 for the detailed tokenization pipeline.

[3] Compared to MAR (Li et al., 2024b) which relies on a bidirectional Transformer to achieve competitive results, we stick to the causality to align with the LLM setting. Their use of continuous-valued tokens on causal LM performs poorly, while our approach achieves results comparable to the bidirectional counterpart with the masked next-token prediction.

perceptron (MLP) M_ϕ conditions on z^i and models the next-token distribution $p(x^i \mid z^i) = p(x^i \mid x^1, ..., x^{i-1}, w)$ with the following diffusion objective (Li et al., 2024b; Rombach et al., 2022).

$$\arg\min_{\theta,\phi} \mathbb{E}_{\varepsilon,t}\left[\left\|\varepsilon - M_\phi(x_t^i,\ z^i,\ t)\right\|^2\right]. \quad (2)$$

$\varepsilon \in \mathbb{R}^d$ is a Gaussian noise sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The diffused token x_t^i is obtained by $x_t^i = \sqrt{\bar{\alpha}_t}x^i + \sqrt{1-\bar{\alpha}_t}\varepsilon$ with the noise schedule $\bar{\alpha}_t$ (Ho et al., 2020; Nichol & Dhariwal, 2021). $t \in \{0, ..., T\}$ is a time step sampled from the noise schedule. We train C_θ and M_ϕ jointly. By backpropagating the gradient from the small diffusion head M_ϕ to the Transformer decoder C_θ, the decoder represents the next-token distribution into the context vector z^i. This facilitates accurate diffusion modeling even with a lightweight M_ϕ.

Inference. At position i, C_θ infers the context vector z^i by a single pass given the previously sampled tokens $z^i = C_\theta(y, \tilde{x}^1, ..., \tilde{x}^{i-1})$. Conditioning on z^i, M_ϕ iteratively denoises a Gaussian noise into a clean token $\tilde{x}^i$, known as the sampling process. Sampling is done via a reverse diffusion procedure (Ho et al., 2020):

$$\tilde{x}_{t-1}^i = \frac{1}{\sqrt{\alpha_t}}\left(\tilde{x}_t^i - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} M_\phi(\tilde{x}_t^i,\ z^i,\ t)\right) + \sigma_t\delta. \quad (3)$$

δ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and σ_t is the noise level at time step t. The decoding de-noising procedure starts from the timestamp T to timestamp 0. That is, $\tilde{x}_T^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\tilde{x}_0^i = \tilde{x}^i \sim p(x^i \mid z^i) = p(x^i \mid x^1, ..., x^{i-1}, w)$. The sampled tokens can then be de-tokenized into the sampled waveform[2].

3.2. AudioMNTP: Masked Next-Token Prediction

AudioNTP yields significant improvement over AudioGen. However, as shown in our results, the generation quality and diversity still lag behind LDMs (Liu et al., 2024c; Majumder et al., 2024). Instead of resorting to the bidirectional solution[3], we stick to the causality to maximally follow the LLM framework. Given the success of MLM in generative models, we seek to migrate the core idea of MLM into the causal next-token prediction. We hypothesize that the key mechanism behind MLM is *predicting the unseen token given the sparse context*[4], essentially independent of the model directionality. We devise a novel learning task tailored to the causal architectures termed **masked next-token prediction (MNTP)**. Figure 3 illustrates the framework.

[4]In MAE (He et al., 2022) and AudioMAE (Huang et al., 2022), over 70% tokens are masked.

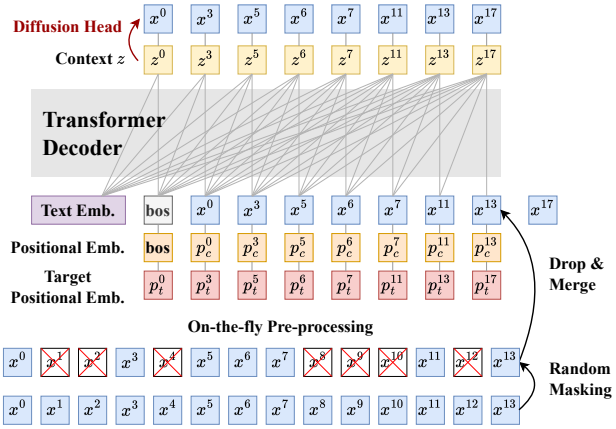


Figure 3. The framework of the masked next-token prediction (MNTP). The continuous-valued tokens are first masked (dropped) to form a shorter sequence. The Transformer decoder then learns the next-token prediction on the dropped sequence with the diffusion loss. This framework is termed **AudioMNTP**.

Training. Before feeding the token sequence $x = \{x^1, ..., x^n\}$ into C_θ, we first apply masking on it. We denote the mask as $v = \{v^1, ..., v^n\}$ where $v^i \in \{0, 1\}$ and $v^i = 0$ means the position i is masked. The masked sequence is then $\bar{x}_v = x \odot v$. Recent works show that dropping instead of masking yields similar performances while greatly reduce the training cost[5]. We drop the tokens to form a shorter sequence $x_v = \{\bar{x}_v^i \in \bar{x}_v \mid \bar{x}_v^i \neq 0\}$. We learn the next-token prediction on x_v as the masked prediction[6]. Upon a closer look, each position i learns to predict various future positions, depending on the random masking patterns. We can then view MNTP as a *multi-task learning with input dropout* (Hou et al., 2022; Zaremba et al., 2014; Sennrich et al., 2016), where **skip-token prediction** benefits next-token prediction during inference[7]. Since the model is unaware of the current masking pattern, it is confusing to naively ask the model predict various future positions indiscriminately. We introduce the **target positional embedding** p_t to C_θ, as shown by Figure 3. Originally, each token has its own learnable content positional embedding p_c. We learn an additional target positional embeddings representing the

[5]Due to the drastically reduced sequence length (Huang et al., 2022; He et al., 2022; Baevski et al., 2023).

[6]Appendix D.1 reviews the similarity and differences between MLM and MNTP.

[7]Predicting skip positions have been proven effective for audio representation learning (Oord et al., 2018; Chung & Glass, 2020). We verify its effectiveness in autoregressive generative modeling. Compared to text, continuous data exhibits high similarity between consecutive tokens, allowing the model to easily exploit local smoothness (Oord et al., 2018) for next-token prediction. We avoid this trivial prediction and facilitate a sparse context required by MLM via skip-token prediction.

position to be predicted and add both embeddings to the audio tokens. Consequently, masked next-token prediction can be viewed as a generalized form of causal language modeling. The model can ***predict any future timestamp given any subset of past information***, with *predicting the immediate next token* and *conditioning on all previous tokens* as special cases of the standard next-token prediction[8].

Inference. The inference procedure of AudioMNTP is identical to that of AudioNTP, except for the inclusion of an additional target positional embedding, which is derived by incrementing the content positional embedding by one.

4. Implementation

Continuous-valued Audio Tokenizer. We leverage the tokenization pipeline of AudioLDM (Liu et al., 2023a), including a pre-trained VAE and a pre-trained Hifi-GAN (Kong et al., 2020) vocoder. Given an audio, the pipeline extracts the Mel-spectrogram, and the VAE encoder processes it into a sequence of continuous-valued tokens $x = \{x^1, ..., x^n\}$, which is used for training. During inference, we sample $\tilde{x} = \{\tilde{x}^1, ..., \tilde{x}^n\}$ according to equation 3. We use the VAE decoder to decode $\tilde{x}$ back to a Mel-spectrogram. Then, the vocoder is used to synthesize the waveform. Refer to Appendix F.1 for details.

Conditional Audio Language Modeling. We mostly follow the implementation in MAR (Li et al., 2024b), including the training/inference details of the MLP diffusion head, and the architecture design of MLP and the Transformer decoder. We train an audio version of the bidirectional MAR as the topline (**AudioMAR**), a causal LM as the baseline (**AudioNTP**), and our main proposal (**AudioMNTP**), which is also causal. For most of the studies, we train AudioMAR, AudioNTP and AudioMNTP with the 193M Base model. We also train a 462M AudioMNTP Large to study the scaling behavior. For the text prompt, we use CLAP (Wu et al., 2023) and FLAN-T5 (Chung et al., 2024) to extract the text embeddings and concatenate them as the input prompt. Refer to Appendix F.2 for details.

Initialization. By default, AudioMAR, AudioNTP, and AudioMNTP are obtained by fine-tuning the pre-trained image MAR[9] on audio data. That is, they all start by learning the MAR task on images and diverge by using MAR, NTP, and MNTP task on audio data, respectively. As shown in the subsequent ablation study, this image initialization only slightly boosts performance and is not critical.

[8]Appendix D.2 illustrates the idea more formally.

[9]Either the Base or Large model in https://github.com/LTH14/mar.

Masking Schedule. For each training iteration, we sample a masking ratio from a distribution over $[0, 1]$ defined by the schedule, and randomly drop the tokens according to the ratio. Specifically, we apply a mixture of normal distribution and truncated normal distribution, where the former emphasizes the high masking ratio for MLM and the latter preserves the long-tailed distribution on the low masking ratio for next-token prediction. See Appendix B for details.

5. Experiments

Training. We train our model on AudioCaps (AC) (Kim et al., 2019) and WavCaps (WC) (Mei et al., 2024). See Appendix A for details. We use AdamW (Loshchilov & Hutter, 2017) optimizer with a fixed learning rate 1.0×10^{-4}. We train the Base model with 40 NVIDIA V100 GPUs, and the Large model requires 104. Our effective batch size is 2048 10-second clips. We train the Base and the Large model for 1000 epochs, about 2 days and 5 days, respectively.

Evaluation. We evaluate our model on the AC evaluation set, following the protocol in AudioLDM and using its evaluation toolkit[10]. For each generated audio and its ground-truth counterpart in AC, we compute several metrics: Fréchet Audio Distance (FAD) (Kilgour et al., 2019), Fréchet Distance (FD), Kullback–Leibler divergence (KL), Inception Score (IS) (Liu et al., 2023a), and Contrastive Language-Audio Pretraining (CLAP) score (Huang et al., 2023b). See Appendix G.1 for details on these metrics. For subjective evaluation, we rate audio samples on text relevance (REL) and overall quality (OVL) using a 1-5 scale (Ghosal et al., 2023; Liu et al., 2023a); additional details are in Appendix G.2. Because speech is challenging in TTA—and even the most advanced systems often produce unintelligible speech—we split our evaluation into speech and non-speech categories.

5.1. Main results

Continuous-valued tokens with diffusion loss are competitive for audio language modeling. Table 1 shows that our 193M baseline AudioNTP Base outperforms the 285M AudioGen Base and 1B UniAudio by a large margin on the FAD and KL metrics. Specifically, we achieve 20%, 40% relative improvements over AudioGen Base on FD and KL scores, respectively. Both AudioGen and UniAudio are based on the discrete tokens and use much more data, parameters and compute compared to our method. The results gauge the effectiveness of the continuous-valued tokens and the diffusion loss for audio language modeling[11].

[10]https://github.com/haoheliu/audioldm_eval

[11]We do not train our model with discrete tokens due to computational constraints imposed by their long sequence length. For

Table 1. **Main results.** FD, FAD, KL, IS, and CLAP metrics on the AudioCaps evaluation set. The AS, AC, and WC stand for AudioSet, AudioCaps, and WavCaps, respectively. **Bold**: the best performance. Underline: the second best performance. *Re-inference with the public checkpoint. The detailed datasets used by each model are listed in the Appendix A.

Token	Method	Datasets	#Params	FD ↓	FAD ↓	KL ↓	IS ↑	CLAP ↑
Bi-directional								
Discrete	MAGNET-Small (Ziv et al., 2024)	AS + AC + 8 others	300M	23.02	3.22	1.42	9.72	0.287
	MAGNET-Large (Ziv et al., 2024)	AS + AC + 8 others	1.5B	26.19	2.36	1.64	9.10	0.253
Continuous	Tango (Ghosal et al., 2023)	AC	866M	24.52	1.59	1.37	7.70	0.313
	Tango-Full-FT (Ghosal et al., 2023)	AS + AC + WC + 5 others	866M	18.93	2.19	1.12	8.80	0.340
	Tango-AF&AC-FT (Kong et al., 2024b)	AC + 1 other	866M	21.84	2.35	1.32	9.59	0.343
	Tango 2 (Majumder et al., 2024)	AS + AC + WC + 6 others	866M	20.66	2.69	1.12	9.09	**0.375**
	Make-an-Audio 2 (Huang et al., 2023a)	AS + AC + WC + 11 others	937M	16.23	2.03	1.29	**9.95**	0.345
	AudioLDM2-AC (Liu et al., 2024c)	AC	346M	-	1.67	1.10	-	-
	AudioLDM2-AC-Large (Liu et al., 2024c)	AC	712M	-	1.42	**0.98**	-	-
	AudioLDM2-Full (Liu et al., 2024c)	AS + AC + WC + 2 others	346M	-	1.78	1.60	-	-
	- re-inference*		346M	32.14	2.17	1.62	6.92	0.273
	AudioLDM2-Full-Large (Liu et al., 2024c)	AS + AC + WC + 2 others	712M	-	1.86	1.64	-	-
	- re-inference*		712M	33.18	2.12	1.54	8.29	0.281
Uni-directional								
Discrete	AudioGen Base (Kreuk et al., 2023)	AS + AC + 8 others	285M	-	2.84	2.14	-	-
	AudioGen Large (Kreuk et al., 2023)	AS + AC + 8 others	1B	-	1.82	1.69	-	-
	UniAudio (Yang et al., 2023a)	AS + AC + WC + 10 others	1B	-	3.12	2.60	-	-
	Ours							
Continuous	**AudioNTP Base**	AC + WC	193M	18.52	2.28	1.29	9.42	0.308
	AudioMNTP Base	AC + WC	193M	14.81	1.68	1.16	9.67	0.336
	AudioMNTP Large	AC + WC	462M	**14.30**	**1.22**	1.17	9.81	0.341

Table 2. Human Evaluation Results. We use AudioMNTP Large to compare to other models.

	Non-Speech		Speech	
Method	REL ↑	OVL ↑	REL ↑	OVL ↑
Reference	4.06 ± 1.09	3.87 ± 1.11	4.47 ± 0.83	4.61 ± 0.95
bi-directional				
AudioLDM 2	3.10 ± 1.29	3.32 ± 1.17	3.17 ± 1.14	3.01 ± 0.91
Tango 2	3.95 ± 0.97	3.80 ± 0.99	4.20 ± 0.79	3.15 ± 0.94
uni-directional				
AudioGen	3.05 ± 1.09	3.02 ± 1.17	3.33 ± 0.97	2.56 ± 1.13
AudioMNTP	3.79 ± 1.02	3.46 ± 1.10	3.91 ± 0.94	3.69 ± 0.88

MNTP significantly outperforms next-token prediction. Table 1 shows that AudioMNTP Base outperforms our baseline AudioNTP Base by a large margin, with 26%, 10%, and 9% relative improvements on FAD, KL and CLAP scores respectively, demonstrating the effectiveness of MNTP.

Scaling up MNTP reaches SOTA performances on FD and FAD. We scale up the model to produce the 462M AudioMNTP Large model. The scaling boosts the performance across most of the metrics, including the 27% and 1.5% relative improvements on FAD and CLAP scores compared to AudioMNTP Base. Our model is intrinsically less expressive compared to SOTA diffusion models, (1) our model is uni-directional and (2) our model is significantly smaller, i.e. 866M Tango 2. However, AudioMNTP Large demonstrates the best FD and FAD scores across all the models, and reach the similar KL, IS and CLAP scores compared to the leading diffusion models[12]. We believe performance can be improved by further scaling up the model and data, and we leave this for future work due to computational constraints.

each 10-second clip, our method uses only 256 tokens, whereas AudioGen uses 5,000 tokens.

5.2. Subjective evaluation

We compare with the best existing decoder-only solution, AudioGen, and two leading diffusion models, AudioLDM 2 and Tango 2. Table 2 shows that AudioMNTP is significantly better than AudioGen in both speech and non-speech categories, and approaching the performance of Tango 2. In both categories, our REL scores lag behind those of Tango 2; however, Tango 2 is a bi-directional model and utilizes the preference optimization dataset to enhance text-audio alignment. Interesting, we find that our method is especially good at generating authentic speech, indicated by the highest OVL score in the speech category, potentially due to the incorporation of MLM, the de facto standard in speech pre-training (Liu et al., 2024a; Baevski et al., 2020)[13].

5.3. Ablation studies

We ablate the components of AudioMNTP in Table 3 and visualize them in the Appendix E. Note that different mask-

[12]We place 5th, 2nd, and 4th based on the KL, IS, and CLAP scores, respectively. Most of the scores are close. Tango 2's especially high CLAP score is attributed to the additional preference dataset Audio-Alpaca (Majumder et al., 2024).

[13]MLM is effective in both discriminative (Baevski et al., 2020) and generative (Liu et al., 2024a) speech pre-training.

Table 3. Ablate the components of MNTP with the Base configuration. INIT. means initialization; POS. EMB. means positional embedding. By default, the masking schedule is the mixture of the normal and truncated normal distribution, as described in Section 4. *indicates the masking schedule used in MAR (Li et al., 2024b), a truncated normal distribution over $[0.7, 1]$. †indicates the fixed masking ratio at 0.7. ‡indicates a masking schedule of uniform distribution over $[0, 1]$.

ID	CAUSAL	MAR INIT.	ZERO MASK	GAUSSIAN MASK	DROP TOKEN	PREDICT NEXT	PREDICT SKIP	TARGET POS. EMB.	PREDICT MASKED	FD ↓	FAD ↓	KL ↓	IS ↑	CLAP ↑
(A)	BASELINE - AUDIONTP													
	✓	✓	✗	✗	✗	✓	✗	✗	✗	18.52	2.28	1.29	9.42	0.308
(B)	✗→✓	✓	✓*	✗	✗	✓	✗	✗	✓	17.15	2.13	1.25	9.45	0.321
(C)	✓	✓	✓†	✗	✗	✓	✗	✗	✗	16.78	1.97	1.28	9.33	0.333
(D)	✓	✓	✗	✓‡	✗	✓	✗	✗	✗	15.15	1.82	1.18	9.22	0.324
(E)	✓	✓	✗	✗	✓	✓	✗	✗	✗	16.62	1.77	1.32	9.25	0.315
(F)	✓	✓	✓	✗	✗	✓	✗	✗	✗	24.45	4.49	1.68	5.87	0.229
(G)	AUDIOMNTP													
	✓	✓	✗	✗	✓	✓	✓	✓	✗	**14.81**	**1.68**	**1.16**	**9.67**	**0.336**
(H)	✓	✓	✗	✗	✓*	✓	✓	✓	✗	15.55	1.89	**1.16**	9.56	0.327
(I)	✓	✓	✗	✗	✓	✓	✓	✗	✗	20.82	3.12	1.55	8.13	0.301
(J)	✓	✗	✗	✗	✓	✓	✓	✓	✗	14.85	1.70	1.17	9.62	0.335
(K)	TOPLINE - AUDIOMAR													
	✗	✓	✓*	✗	✗	✗	✗	✗	✓	14.86	1.35	1.17	9.75	0.346

ing strategies, such as masking or dropping, favor different masking schedules. We explored various schedules and report the best result.

Combining MLM and NTP. Table 3 (B) naively combines MLM with NTP by using the bidirectional AudioMAR as the initialization and fine-tuning with next-token prediction on audio data. That is, Table 3 (A) uses the image MAR as the initialization, but Table 3 (B) uses AudioMAR as the initialization, which additionally learns the MAR task on audio data. Table 3 (B) shows that MLM on audio data improves performance, but only slightly.

Migrate MLM into NTP: masking. The bidirectional MLM is essentially not designed for the causal decoding. We investigate incorporating input masking into next-token prediction. Table 3 (C) follows MAR to apply zero masking, but uses the causal decoder for next-token prediction. Table 3 (D) tries the Gaussian noises as the masking, since the numerical value space is more aligned to that of the clean tokens, which are sampled from the same Gaussian by the diffusion head. Appendix E Figure 8 (C) illustrates the idea. Table 3 (E) tries dropping the tokens as a type of masking, which significantly reduces the compute cost, as shown by Appendix E Figure 8 (E). We predict the next tokens directly on top of the visible tokens instead of the masked tokens, and the masked tokens are completely removed. Given that our masking ratio is typically above 70% (Li et al., 2024b; Huang et al., 2022), dropping only requires less than half of the original computation during training. According to Table 3 (C), (D), and (E), incorporating input masking to next-token prediction improves the performance on most metrics compared to Table 3 (A), including FD, FAD and CLAP, except for IS. We proceed with the dropping scenario due to its computational efficiency. However, our proposal is independent of the exact realization of the masking. Note that other masking strategies, such as zero masking, do not work well with our final masking schedule, as indicated in Table 3 (F). In contrast, dropping tokens remains effective with different masking schedules in subsequent studies.

Migrate MLM into NTP: masked prediction. We study the effectiveness of skip-token prediction, our form of masked prediction, which we hypothesize to benefit next-token prediction. Table 3 (G), AudioMNTP, justifies our hypothesis and outperforms all the rows on all metrics. We ablate our mixture of distributions masking schedule to verify the robustness of MNTP. We replace the schedule by the one used in MAR[14]. Table 3 (H) shows that our MNTP task, even with the masking schedule of MAR, remains competitive compared to Table 3 (A) and Table 3 (B), only slightly underperforming Table 3 (G). This demonstrates MNTP's robustness to the masking schedule. Next, we ablate the target positional embedding used for skip-token prediction. Table 3 (I) is significantly worse than Table 3 (G) on all metrics, suggesting that different future positions conflict each other. Table 3 (I) is even worse than Table 3 (E), which simply predicts the immediate next token. The results demonstrate the importance of the target positional embedding.

Initialization. We replace the image MAR initialization by random initialization for AudioMNTP. Table 3 (J) underperforms Table 3 (G) on all metrics, but slightly, suggesting that the image MAR initialization is helpful but not neces-

[14]We find that MAR's truncated normal distribution on $[0.7, 1.0]$ works poorly, but simply shifting it to the left to $[0.55, 0.85]$ works reasonably well.

Table 4. **Ablating different MLP diffusion head sizes** with the AudioMNTP Base configuration. The dimension of each MLP layer is 1024. L denotes the layers of the MLP. The relative size of the heads in the whole model is 5.4%, 10.7%, and 17.4% for $L = 1, 3, 6$, respectively. S denotes the inference speed, measured by seconds per 10-second audio with batch size 40 on an NVIDIA V100 GPU.

MLP L	MLP Size	FD ↓	FAD ↓	KL ↓	IS ↑	CLAP ↑	S
1	9.84M	16.97	1.68	1.20	8.87	0.321	**3.22**
3	20.34M	14.81	1.68	1.16	9.67	0.336	3.93
6	36.09M	**14.74**	**1.21**	**1.15**	**9.78**	**0.337**	4.84

Table 5. Comparing the text embedding with the Base configuration. EMB. denotes embedding.

Text Emb.	FD ↓	FAD ↓	KL ↓	IS ↑	CLAP ↑
CLAP + FLAN-T5	**14.82**	**1.68**	**1.16**	**9.67**	**0.336**
CLAP	15.94	1.75	1.29	8.36	0.297
FLAN-T5	16.43	1.75	1.37	8.22	0.285

sary for MNTP to work.

The size of the MLP diffusion head. Table 4 shows that the size of the diffusion head matters. Increasing the head from 1 layer to 3 layers significantly improve the performances on most metrics. Further increasing to 6 layers significantly decrease the FAD score. However, with the larger MLP size, the inference time increases. Our results highlight that the size of the diffusion head is an important hyperparameter for balancing performance and efficiency during trade-offs.

The ensemble text embedding. Table 5 shows that concatenating CLAP and FLAN-T5 yields the best performance, as both embeddings provide complementary information to the model, and neither individual embedding can match the performance of the joint embeddings.

5.4. MAR vs. MNTP vs. Next-Token Prediction

MNTP is approaching MAR. First, Table 6A indicates that AudioMAR yields the best results on most metrics except for FD. This is expected, as it is a bidirectional model with the full context. However, compared to AudioNTP, AudioMNTP significantly bridges the gap and achieves performance comparable to that of AudioMAR on IS and CLAP score. AudioMNTP even surpasses AudioMAR on FD and KL, suggesting that MNTP is a competitive task for learning causal language modeling on continuous data.

The effectiveness of MNTP in causal decoding. MAR is competitive, but its use of future context does not support streaming in the LLM framework well. We verify this hypothesis by comparing AudioMAR, AudioNTP, and AudioMNTP under causal decoding. AudioNTP and AudioMNTP are intrinsically developed for causal decoding, while AudioMAR can perform causal decoding by unmasking from left to right. Causal decoding degrades the performance of AudioMAR when comparing Table 6 (A) and (C). Interesting, by increasing the decoding steps as shown by Table 6 (D), AudioMAR works much better, and even surpasses the naive AudioNTP in Table 6 (E). However, Table 6 (F) indicates that AudioMNTP is the best in most metrics in the causal decoding scenario[15].

Table 6. **Comparing MAR and MNTP on different decoding scenarios.** The random-order decoding is the default decoding method used in MAR. The **steps** means the number of decoding steps. 64 is the default decoding step of MAR. 256 is the total sequence length of the 10-second audio. AudioMAR supports parallel decoding so it can reduce the decoding steps, denoted by †. The **bold** denotes the best performance with the causal decoding. The underline denotes the globally best performance.

ID	Model	Steps	FD ↓	FAD ↓	KL ↓	IS ↑	CLAP ↑
			random decoding				
(A)	AudioMAR	64†	14.86	1.35	1.17	9.75	0.35
(B)		256	15.02	1.57	1.13	9.47	0.347
			causal decoding				
(C)	AudioMAR	64†	21.43	2.17	1.17	7.31	0.287
(D)		256	16.32	1.84	**1.14**	9.09	0.317
(E)	AudioNTP	256	18.52	2.28	1.29	9.42	0.308
(F)	AudioMNTP	256	**14.82**	**1.68**	1.16	**9.67**	**0.336**

6. Conclusion

We propose **AudioNTP** and **AudioMNTP**, a high-performing framework for audio language modeling. Our results demonstrate that **continuous-valued tokens** are competitive for audio language modeling, yielding 20% improvements over the discrete token-based audio language models. Our **masked next-token prediction (MNTP)** further validates the benefits of applying masked language modeling concept to causal language modeling on continuous data, achieving another 20% improvements over next-token prediction while preserving the causality during inference. By combining continuous-valued tokens with MNTP, our approach achieves SOTA audio generation quality, rivaling the leading latent diffusion models and reviving the language modeling approach for sound generation. Moreover, our models are significantly smaller than most existing solutions, showcasing the efficiency and effectiveness of the proposed learning techniques. These findings point toward a promising direction for directly generating high-fidelity audio with the streamable and scalable large language models.

[15] We verified the idea of simply applying MLM to a Transformer decoder in Table 3 (C) and (D), which are worse than AudioMNTP.

Impact Statement

Scaling with Large Language Models (LLMs). Our work demonstrates that causal language modeling can produce high-fidelity audio even with small models and limited data. A natural next step is to scale up to billions of parameters like LLMs and incorporate larger datasets, such as AudioSet[16]. With this scaling, our approach has the potential to set a new SOTA in streamable sound generation.

Merge with other Modalities in LLMs. Since our method only requires changing the prediction head of the Transformer decoder, it can be easily integrated into existing multi-modality LLM training to jointly learn with other modalities such as text and image (Sun et al., 2024). Compared to post-synthesizing with the external vocoders (Huang et al., 2024; Liu et al., 2023b; Wang et al., 2024b), the end-to-end approach avoids error propagation and allows audio data to interact directly and losslessly with other modalities in a single model. Audio generation also benefits from the LLM scaling with other modalities.

Efficient Inference. Since we primarily adhere to the standard Transformer decoder, our model naturally benefits from resources developed for this standardized architecture, including the KV-cache (Pope et al., 2023) and PagedAttention (Kwon et al., 2023). Furthermore, our model learns to *predict any future timestamp given any subset of past information*. Theoretically, one can devise a parallel decoding algorithm by manipulating the target positional embedding. With these techniques, combined with our inherently streaming capability, our method can potentially establish a new SOTA in efficient, high-fidelity sound generation.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Aghajanyan, A., Huang, P.-Y. B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., and Zettlemoyer, L. Cm3: A causal masked multimodal model of the internet. *ArXiv*, abs/2201.07520, 2022. URL https://api.semanticscholar.org/CorpusID:246035820.

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

[16]We currently rely only on AudioCaps and WavCaps, totaling about 1,000 hours of audio.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Baevski, A., Babu, A., Hsu, W.-N., and Auli, M. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pp. 1416–1429. PMLR, 2023.

Bavarian, M., Jun, H., Tezak, N., Schulman, J., McLeavey, C., Tworek, J., and Chen, M. Efficient training of language models to fill in the middle. *arXiv preprint arXiv:2207.14255*, 2022.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31: 2523–2533, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Chen, H., Xie, W., Vedaldi, A., and Zisserman, A. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020a.

Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., and Sutskever, I. Generative pretraining from pixels. In *International conference on machine learning*, pp. 1691–1703. PMLR, 2020b.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Chung, Y.-A. and Glass, J. Improved speech representations with multi-target autoregressive predictive coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2353–2358, 2020.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023.

Deshmukh, S., Elizalde, B., and Wang, H. Audio retrieval with wavtext5k and clap training. In *INTERSPEECH 2023*, pp. 2948–2952, 2023. doi: 10.21437/Interspeech.2023-1136.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Drossos, K., Lipping, S., and 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, 2020.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pp. 5547–5569. PMLR, 2022.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Fonseca, E., Pons Puig, J., Favory, X., Font Corbera, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., and Serra, X. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR), 2017.

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Fried, D., Aghajanyan, A., Lin, J., Wang, S., Wallace, E., Shi, F., Zhong, R., Yih, S., Zettlemoyer, L., and Lewis, M. Incoder: A generative model for code infilling and synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Ghosal, D., Majumder, N., Mehrish, A., and Poria, S. Text-to-audio generation using instruction-tuned LLM and latent diffusion model. *CoRR*, abs/2304.13731, 2023. doi: 10.48550/ARXIV.2304.13731. URL https://doi.org/10.48550/arXiv.2304.13731.

Gong, Y., Chung, Y.-A., and Glass, J. Ast: Audio spectrogram transformer. In *Interspeech 2021*, pp. 571–575, 2021. doi: 10.21437/Interspeech.2021-698.

Guo, Z., Leng, Y., Wu, Y., Zhao, S., and Tan, X. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hershey, S., Ellis, D. P., Fonseca, E., Jansen, A., Liu, C., Moore, R. C., and Plakal, M. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 366–370. IEEE, 2021.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 30016–30030, 2022.

Hou, L., Pang, R. Y., Zhou, T., Wu, Y., Song, X., Song, X., and Zhou, D. Token dropping for efficient bert pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3774–3784, 2022.

Huang, J., Ren, Y., Huang, R., Yang, D., Ye, Z., Zhang, C., Liu, J., Yin, X., Ma, Z., and Zhao, Z. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *CoRR*, abs/2305.18474, 2023a. doi: 10.48550/ARXIV.2305.18474. URL https://doi.org/10.48550/arXiv.2305.18474.

Huang, P.-Y., Xu, H., Li, J., Baevski, A., Auli, M., Galuba, W., Metze, F., and Feichtenhofer, C. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022.

Huang, R., Chen, F., Ren, Y., Liu, J., Cui, C., and Zhao, Z. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3945–3954, 2021.

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., and Zhao, Z. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pp. 13916–13932. PMLR, 2023b.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 23802–23804, 2024.

Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7669–7673. IEEE, 2020.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota, 2019.

Kilgour, K., Zuluaga, M., Roblek, D., and Sharifi, M. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech 2019*, pp. 2350–2354, 2019. doi: 10.21437/Interspeech.2019-2219.

Kim, C. D., Kim, B., Lee, H., and Kim, G. 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)*, pp. 119–132, 2019.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., and Khudanpur, S. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5220–5224. IEEE, 2017.

Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

Kong, Z., Goel, A., Badlani, R., Ping, W., Valle, R., and Catanzaro, B. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*, 2024a.

Kong, Z., Lee, S.-g., Ghosal, D., Majumder, N., Mehrish, A., Valle, R., Poria, S., and Catanzaro, B. Improving text-to-audio models with synthetic captions. In *Proc. SynData4GenAI 2024*, pp. 1–5, 2024b.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y., and Adi, Y. Audiogen: Textually guided audio generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=CYK7RfcOzQ4.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp.

7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Li, T., Chang, H., Mishra, S., Zhang, H., Katabi, D., and Krishnan, D. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2142–2152, 2023.

Li, T., Katabi, D., and He, K. Return of unconditional generation: A self-supervised representation generation method. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

Li, T., Tian, Y., Li, H., Deng, M., and He, K. Autoregressive image generation without vector quantization. *arXiv preprint arXiv:2406.11838*, 2024b.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Liu, A. H., Le, M., Vyas, A., Shi, B., Tjandra, A., and Hsu, W.-N. Generative pre-training for speech with flow matching. In *The Twelfth International Conference on Learning Representations*, 2024a.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W., and Plumbley, M. D. Audioldm: text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 21450–21474, 2023a.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024b.

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., and Plumbley, M. D. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:2871–2883, 2024c. doi: 10.1109/TASLP.2024.3399607. URL https://doi.org/10.1109/TASLP.2024.3399607.

Liu, X., Zhu, Z., Liu, H., Yuan, Y., Cui, M., Huang, Q., Liang, J., Cao, Y., Kong, Q., Plumbley, M. D., et al. Wavjourney: Compositional audio creation with large language models. *arXiv preprint arXiv:2307.14335*, 2023b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Majumder, N., Hung, C., Ghosal, D., Hsu, W., Mihalcea, R., and Poria, S. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *CoRR*, abs/2404.09956, 2024. doi: 10.48550/ARXIV.2404.09956. URL https://doi.org/10.48550/arXiv.2404.09956.

Martín-Morató, I. and Mesaros, A. What is the ground truth? reliability of multi-annotator data for audio tagging. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 76–80. IEEE, 2021.

Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M. D., Zou, Y., and 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*, 2024.

Meng, L., Zhou, L., Liu, S., Chen, S., Han, B., Hu, S., Liu, Y., Li, J., Zhao, S., Wu, X., et al. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551*, 2024.

Mesaros, A., Heittola, T., and Virtanen, T. Tut database for acoustic scene classification and sound event detection. In *2016 24th European Signal Processing Conference (EUSIPCO)*, pp. 1128–1132. IEEE, 2016.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *Proceedings of the International Conference on Machine Learning*, 2021.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Peng, P., Huang, P.-Y., Li, S.-W., Mohamed, A., and Harwath, D. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12442–12462, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.673. URL https://aclanthology.org/2024.acl-long.673.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Pope, R., Douglas, S., Chowdhery, A., Devlin, J., Bradbury, J., Heek, J., Xiao, K., Agrawal, S., and Dean, J. Efficiently scaling transformer inference. *Proceedings of Machine Learning and Systems*, 5:606–624, 2023.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. Mls: A large-scale multilingual dataset for speech research. In *Interspeech 2020*, pp. 2757–2761, 2020. doi: 10.21437/Interspeech.2020-2826.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Salamon, J., Jacoby, C., and Bello, J. P. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Sennrich, R., Haddow, B., and Birch, A. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pp. 371–376, 2016.

Shi, Y., Bu, H., Xu, X., Zhang, S., and Li, M. Aishell-3: A multi-speaker mandarin tts corpus. In *Interspeech 2021*, pp. 2756–2760, 2021. doi: 10.21437/Interspeech.2021-755.

Sun, Y., Bao, H., Wang, W., Peng, Z., Dong, L., Huang, S., Wang, J., and Wei, F. Multimodal latent language modeling with next-token diffusion. *arXiv preprint arXiv:2412.08635*, 2024.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tschannen, M., Eastwood, C., and Mentzer, F. Givt: Generative infinite-vocabulary transformers. In *European Conference on Computer Vision*, pp. 292–309. Springer, 2025.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Vyas, A., Shi, B., Le, M., Tjandra, A., Wu, Y.-C., Guo, B., Zhang, J., Zhang, X., Adkins, R., Ngan, W., et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A. (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446.

Wang, X., Li, Y., Fu, C., Xie, L., Li, K., Sun, X., and Ma, L. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*, 2024a.

Wang, Y., Huang, R., Song, S., Huang, Z., and Huang, G. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. *Advances in neural information processing systems*, 34:11960–11973, 2021.

Wang, Y., Wang, X., Zhu, P., Wu, J., Li, H., Xue, H., Zhang, Y., Xie, L., and Bi, M. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. In *Interspeech 2022*, pp. 4242–4246, 2022. doi: 10.21437/Interspeech.2022-48.

Wang, Z., Tai, Y.-W., and Tang, C.-K. Audio-agent: Leveraging llms for audio generation, editing and composition. *arXiv preprint arXiv:2410.03335*, 2024b.

Wei, C., Mangalam, K., Huang, P.-Y., Li, Y., Fan, H., Xu, H., Wang, H., Xie, C., Yuille, A., and Feichtenhofer, C. Diffusion models as masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16284–16294, 2023.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and 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, 2023.

Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., Shi, J., Zhao, S., Bian, J., Wu, X., Zhao, Z., Watanabe, S., and Meng, H. Uniaudio: An audio foundation model toward universal audio generation. *CoRR*, abs/2310.00704, 2023a. doi: 10.48550/ARXIV.2310.00704. URL https://doi.org/10.48550/arXiv.2310.00704.

Yang, D., Yu, J., Wang, H., Wang, W., Weng, C., Zou, Y., and Yu, D. Diffsound: Discrete diffusion model for

text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1720–1733, 2023b.

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Yu, L., Simig, D., Flaherty, C., Aghajanyan, A., Zettlemoyer, L., and Lewis, M. Megabyte: Predicting million-byte sequences with multiscale transformers. *Advances in Neural Information Processing Systems*, 36:78808–78823, 2023.

Žagar, A. and Robnik-Šikonja, M. Slovene SuperGLUE benchmark: Translation and evaluation. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S. (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2058–2065, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.221.

Zaremba, W., Sutskever, I., and Vinyals, O. Recurrent neural network regularization. *ArXiv*, abs/1409.2329, 2014. URL https://api.semanticscholar.org/CorpusID:17719760.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. Libritts: A corpus derived from librispeech for text-to-speech. In *Interspeech 2019*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441.

Zhang, X., Lyu, X., Du, Z., Chen, Q., Zhang, D., Hu, H., Tan, C., Zhao, T., Wang, Y., Zhang, B., et al. Intrinsicvoice: Empowering llms with intrinsic real-time voice interaction abilities. *arXiv preprint arXiv:2410.08035*, 2024.

Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., and Levy, O. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

Ziv, A., Gat, I., Lan, G. L., Remez, T., Kreuk, F., Copet, J., Défossez, A., Synnaeve, G., and Adi, Y. Masked audio generation using a single non-autoregressive transformer. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Ny8NiVfi95.

A. Data

We compare the datasets used by our method to the existing systems. We only use AudioCaps (Kim et al., 2019) and WavCaps (Mei et al., 2024), highlighting the data efficiency of our methods. Audios longer than 10 seconds are randomly cropped into 10 second. That is, the number of the text-audio pairs are the same after the pre-processing. To speedup the training, we pre-extract the text embedding and the continuous-valued audio tokens.

Table 7. **The datasets used by the TTA systems.** The AS, AC, WC stand for AudioSet (Gemmeke et al., 2017), AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024), respectively. WavCaps is composed of FreeSound (Fonseca et al., 2017), BBC Sound Effects[18], SoundBible[19], and the AudioSet Strongly-Labeled Subset (Hershey et al., 2021), augmented with the ChatGPT-generated text prompt. Some systems use the individual datasets in WavCaps without the natural language prompt, i.e. AudioGen.

METHOD	DATASETS
AUDIOGEN BASE (KREUK ET AL., 2023) AUDIOGEN LARGE (KREUK ET AL., 2023) MAGNET-SMALL (AUDIO[20]) (ZIV ET AL., 2024) MAGNET-LARGE (AUDIO) (ZIV ET AL., 2024)	**AS** + **AC** + BBC SOUND EFFECTS + CLOTHO V2 (DROSSOS ET AL., 2020) + VGG-SOUND (CHEN ET AL., 2020A) + FSD50K (FONSECA ET AL., 2021) + FREE TO USE SOUNDS[21] + SONNISS GAME EFFECTS[22] + WESOUNDEFFECTS[23] + ODEON SOUND EFFECTS[24]
TANGO (GHOSAL ET AL., 2023)	**AC**
TANGO-FULL-FT (GHOSAL ET AL., 2023)	**AS** + **AC** + **WC** + MUSICCAPS (AGOSTINELLI ET AL., 2023) + ESC (PICZAK, 2015) + URBANSOUND (SALAMON ET AL., 2014) + GTZAN (TZANETAKIS & COOK, 2002) + MUSICAL INSTRUMENTS[25]
TANGO-AF&AC-FT (KONG ET AL., 2024B)	**AC** + AF-AUDIOSET[26]
TANGO 2 (MAJUMDER ET AL., 2024)	TANGO-FULL-FT DATA + AUDIO-ALPACA (AA)[27]
MAKE-AN-AUDIO 2 (HUANG ET AL., 2023A)	**AS** + **AC** + **WC** + WAVTEXT5K (DESHMUKH ET AL., 2023) + ADOBE AUDITION SOUND EFFECTS[28] + MACS (MARTÍN-MORATÓ & MESAROS, 2021) + CLOTHV2 (DROSSOS ET AL., 2020) + AUDIOSTOCK[29] + EPIDEMIC SOUND[30] + FSD50K (FONSECA ET AL., 2021) + ODEON SOUND EFFECTS[24] URBANSOUND (SALAMON ET AL., 2014) + ESC (PICZAK, 2015) + TUT (MESAROS ET AL., 2016)
AUDIOLDM2-AC (LIU ET AL., 2024C) AUDIOLDM2-AC-LARGE (LIU ET AL., 2024C)	**AC**
AUDIOLDM2-FULL (LIU ET AL., 2024C) AUDIOLDM2-FULL-LARGE (LIU ET AL., 2024C)	**AS** + **AC** + **WC** + VGG-SOUND (CHEN ET AL., 2020A) + MSD (BERTIN-MAHIEUX ET AL., 2011)
UNIAUDIO (YANG ET AL., 2023A)	**AS** + **AC** + **WC** + LIBRILIGHT (KAHN ET AL., 2020) + LIBRITTS (ZEN ET AL., 2019) + MLS (PRATAP ET AL., 2020) + AISHELL3 (SHI ET AL., 2021) + OPENCPOP (WANG ET AL., 2022) + OPENSINGER (HUANG ET AL., 2021) + MSD (BERTIN-MAHIEUX ET AL., 2011) + PROMPTSPEECH (GUO ET AL., 2023) + OPENSLR26,OPENSLR28F (KO ET AL., 2017)
AUDIO-NTP BASE **AUDIO-MNTP BASE** **AUDIO-MNTP LARGE**	**AC** + **WC**

B. Masking Schedules

B.1. Defining masking schedules.

We explored various masking schedules, including the truncated normal distribution used in MAR, the fixed masking ratio, uniform distribution, and our mixture of normal and truncated normal distribution. Different *masking strategies*, including

[18]https://sound-effects.bbcrewind.co.uk

[19]https://soundbible.com/

[20]There are music and audio versions of Magnet. The music version follows the datasets used in (Copet et al., 2024) and the audio version follows those in AudioGen.

[21]https://www.freetousesounds.com/all-in-one-bundle/

[22]https://sonniss.com/gameaudiogdc

[23]https://wesoundeffects.com/we-sound-effects-bundle-2020/

[24]https://www.paramountmotion.com/odeon-sound-effects

[25]https://www.kaggle.com/datasets/soumendraprasad/musical-instruments-sound-dataset

[26]AF-AudioSet is a synthetic dataset released by (Kong et al., 2024b), where Audio Flamingo (Kong et al., 2024a) provides the caption for the AudioSet audio.

[27]The Audio-Alpaca (AA) preference dataset is released by Tango 2, where a text prompt corresponds to a winner audio and a loser audio. It is used after the standard TTA supervised training for the preference alignment.

[28]https://www.adobe.com/products/audition/offers/adobeauditiondlcsfx.html

[29]https://audiostock.net/

[30]https://www.epidemicsound.com/

zero masking, Gaussian masking, and dropping, work best under different *masking schedules*, as suggested by Table 3. By default, a masking schedule represents a distribution over $[0, 1]$, where we sample a masking ratio $r \sim [0, 1]$ for each training iteration. Then, given a length-n sequence of continuous-valued tokens, we mask (drop) $n \times r$ tokens. For the dropping masking strategy, the remaining sequence length is $n \times (1 - r)$. In Figure 4, we visualize a few representative schedules we find useful in this study.

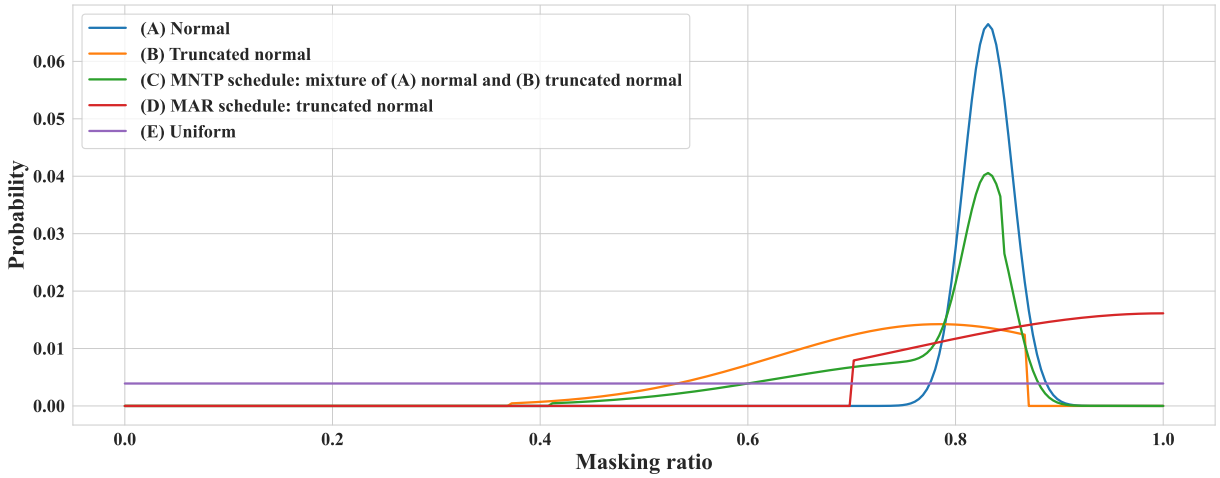


Figure 4. Visualizing the masking schedules. We first sample a masking ratio from the schedule, a probability distribution over $[0, 1]$, and then sample the masking positions in the token sequence based on the ratio. In (C), we average (A) and (B) with equal weights.

B.2. Ablating the MNTP schedule.

We examine the components of the MNTP masking schedule, specifically the normal distribution and the truncated normal distribution. Table 8 shows that the normal distribution (A) plays a more critical role than the truncated normal distribution (B), underscoring the importance of a high masking ratio during MLM. However, the normal distribution alone results in a poor IS score. We hypothesize that this issue arises from a training/testing mismatch. During next-token prediction, all previous tokens are presented—a scenario that is less frequently encountered during training if the focus is solely on a high masking ratio. To address this, we incorporate a long-tailed distribution (B) into the normal distribution by averaging them, resulting in (C), our final masking schedule. As shown in Table 8, (C) outperforms (D), the MAR default masking schedule, across all metrics, demonstrating the effectiveness of our approach.

Table 8. Ablating the masking schedule of MNTP. The schedule IDs match the IDs in Figure 4.

SCHEDULE ID	FD ↓	FAD ↓	KL ↓	IS ↑	CLAP ↑
(A)	15.22	1.70	1.17	8.83	0.326
(B)	15.96	2.17	1.25	8.18	0.293
(C)	**14.81**	**1.68**	**1.16**	**9.67**	**0.336**
(D)	15.55	1.89	1.16	9.56	0.327

C. Classifier-free guidance (CFG) and sampling temperature

Classifier-free guidance. Classifier-free guidance is a method for trading off diversity in favor of fidelity. We reuse the notations in Section 3.1. During training, the input text prompt w is replaced with a fake latent w_f with a probability of 10%. The fake latent has the same length as the text prompt and is learned jointly with the whole model. During testing, we run two inferences for each text prompt: conditional generation and unconditional generation. Specifically, at each decoding position i, we get two conditioning vectors: $z_c^i = C_\theta(w, \beta, x^1, ..., x^{i-1})$ and $z_u^i = C_\theta(w_f, \beta, x^1, ..., x^{i-1})$. At each diffusion sampling step t, we then induce two noise predictions with the diffusion head: $\varepsilon_c = M_\phi(\tilde{x}_t^i,\ z_c^i,\ t)$ and $\varepsilon_u = M_\phi(\tilde{x}_t^i,\ z_u^i,\ t)$. The classifier-free guidance is realized by linear interpolating two predicted noises.

$$\varepsilon = \varepsilon_c + \omega_i \cdot (\varepsilon_c - \varepsilon_u) \tag{4}$$

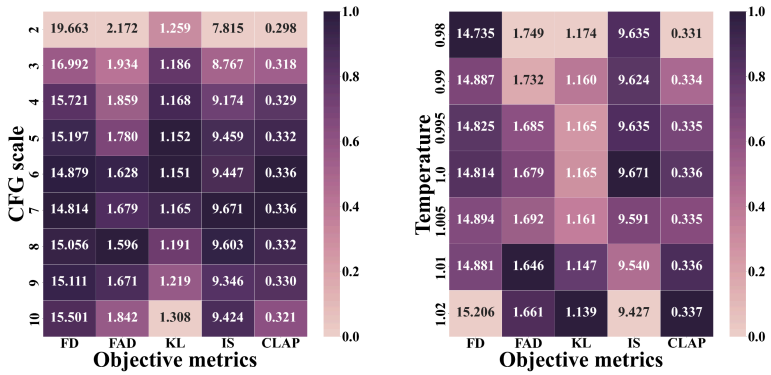
where $\omega_i \in R$ is the guidance scale when decoding the position i. Intuitively, the guidance scale represents the degree of *moving away* from the unconditional generation. With higher ω_i, the generation becomes more aligned with the given text prompt, albeit with reduced diversity. We do not fix a guidance scale ω_i for all positions. We adopt a annealing schedule in (Ziv et al., 2024). That is, given the current decoding position $i \in \{1, ..., n\}$ and an initial CFG scale $\omega_0 \geq 1$, we gradually decrease the guidance scale by:

$$\omega_i = 1 + (\omega_0 - 1) \times \left(1 - \frac{i-1}{n-1}\right) \tag{5}$$

Intuitively, this means that the initial decoded tokens are more tailored to the given condition and gradually allow for greater diversity and uncertainty. We find that this annealing schedule works better than the constant schedule and the linear schedule used in MAR (Li et al., 2024b).

Sampling temperature. Conventionally, the higher sampling temperature corresponds to more diverse samples. To facilitate this behavior, we can multiply the noise scale δ in equation 3 by a temperature factor τ (Dhariwal & Nichol, 2021).

Results. Figure 5 demonstrates the effects of the CFG guidance scale ω_0 and the sampling temperature τ. Figure 5 (a) shows that ω_0 is critical for the competitive performances, and the best performance is achieved around 7. As a result, we set $\omega_0 = 7$ as default thorough the article. Next, we explore different sampling temperatures with $\omega_0 = 7$ in Figure 5 (b). Figure 5 (b) shows that sampling temperatures have varying impacts on different metrics. As a result, given a specific metric to optimize, it is helpful to tune the temperature. However, no single τ achieves the best performance across all metrics. Therefore, we set $\tau = 1$ as the default.



(a) Classifier-Free Guidance **(b) Temperature**

Figure 5. Ablating the (a) classifier-free guidance scale (CFG) and the (b) temperature during the MLP diffusion sampling. The darker color represents the better performance. We use the AudioMNTP Base configuration for the ablation. In (a), the temperature is fixed to 1.0; In (b) the CFG is fixed to 7.0, as suggested by (a).

D. MNTP v.s. MLM v.s. Next-Token Prediction

We compare MNTP to the closely related MLM and next-token prediction for their similarities and differences.

D.1. Comparing MNTP to MLM

MNTP draws inspiration from MLM to improve next-token prediction. They share the same spirit in *learning the contextualized dependencies*, where the model generates the representation from a sparse context which is predictive to the unseen data. Figure 6 illustrate the idea. In MLM, the model predicts the unseen x_5 conditioning on a masked sparse context

$\{x^0, x^1, x^4, x^7\}$. In MNTP, the model predicts the unseen x^7 also conditioning on a dropped sparse context $\{x^0, x^1, x^4\}$. The differences include the following:

1. **Directionality**: MLM relies on a bidirectional model, while MNTP is designed for the causal model. As a result, the sparse context of MLM includes both the past $\{x^0, x^1, x^4\}$ and future $\{x^7\}$ context, while that of MNTP includes only the past context $\{x^0, x^1, x^4\}$.
2. **Mask token & Target positional embedding**: We remove the mask tokens which do not appear in the NTP decoding stage and cost extra computation, and predict the unseen tokens directly at the position of the seen tokens with the additional target positional information.
3. **Prediction target**: We consider the right-masked span in Figure 6 as an example. In MLM, all tokens in the masked span are predicted using both left and right context, e.g., $p(x^5 \mid x^4, \ldots)$ and $p(x^6 \mid x^7, \ldots)$. In contrast, MNTP predicts only the rightmost token using left context alone, e.g., $p(x^7 \mid x^4, \ldots)$. To quantify task difficulty, we only list the closest observed token to the unseen token as the condition. As a result, MNTP is uni-directional and relies on more distant tokens on average compared to MLM, making it intuitively more challenging.

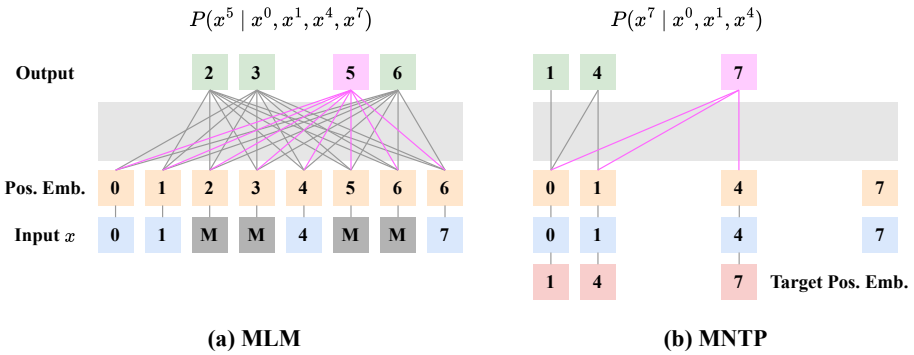


Figure 6. Comparing the differences between MLM and MNTP.

D.2. Comparing MNTP to Next-Token Prediction.

We can view MNTP as a Generalized form of Causal Language Modeling (GCLM), where the model *predicts any future timestamp given any subset of past information*. We illustrate the idea with an simple example in Figure 7. Given the length-5 sequence and the current token x^3, all the possible prediction patterns in GCLM are enumerated in Figure 7 (b), including possible subsets of the past conditions and the possible future prediction targets. All the corresponding masking patterns are listed in Figure 7 (a), and they are all possible masking patterns in our masking schedule since we sample the masking ratio from the entire $[0, 1]$. To be formal, we follow the notation in Section 3.1 and Section 3.2. Given an unmasked sequence $x = \{x^1, ..., x^n\}$, without loss of generality, we pick a current input token x^i where $1 < i < n$. We then have the past token set $P = \{x^1, ..., x^{i-1}\}$ and future token set $F = \{x^{i+1}, ..., x^n\}$. For any prediction pattern in GCLM, there is a past token subset $\bar{P} \subseteq P$ and a target future token $x^f \in F$, we can derive at least one masking pattern $v = \{v^1, ..., v^n\}$ which satisfies this prediction pattern by:

$$v^j = \begin{cases} 1 & \text{if } j \in \{i, f\} \text{ or } (j < i \text{ and } x^j \in \bar{P}), \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Note that this is not the only masking pattern satisfying this prediction pattern, our goal is to show that the prediction pattern would be sampled and learned. This masking pattern can be sampled when setting the masking ratio $r = 1 - \sum_{j=1}^{n} \frac{v^j}{n}$. Since our masking schedule is a distribution over the entire $[0, 1]$ (See Appendix B.1), the pattern would be sampled and learned. As a result, our MNTP at position i models $p_i(x^f \mid \bar{P})$ over all $x^f \in F$ and $\bar{P} \subseteq P$. By setting $\bar{P} = P$ and $x^f = x^{i+1}$, we restores the original next-token prediction with the full context.

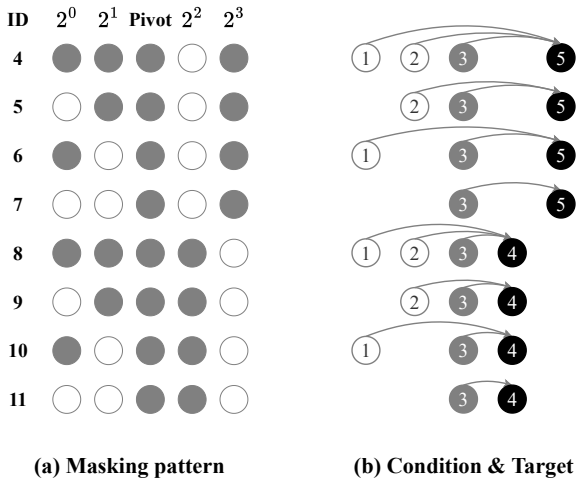


Figure 7. Visualizing the generalized causal language model. (a) shows the masking patterns. We enumerate the masking patterns with binary numbers where a masked position is denoted by 1 (white) and a unmasked position is denoted by 0 (gray). (b) shows the corresponding dropped sequence and the prediction pattern given the current gray token x^3. The conditions are in white and the prediction target is in black.

E. Ablating the components of MNTP

We visualize the process of MNTP ablation in Figure 8. Refer to Section 5.3 for the discussion.

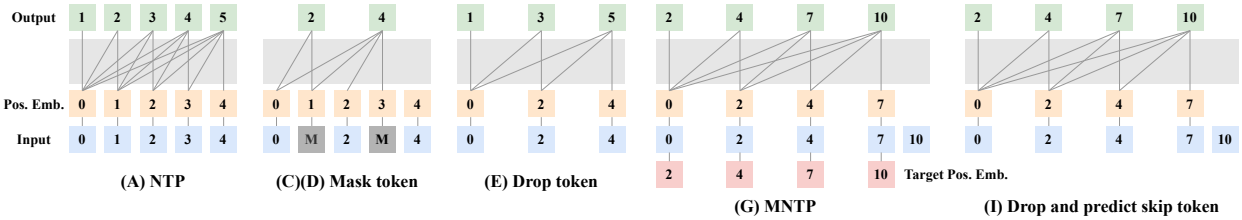


Figure 8. Visualizing the differences between the ablation variants. The subfigure IDs match the IDs in Table 3.

F. Implementation

F.1. Continuous-valued audio tokenizer

Tokenize. We leverage the pre-processing pipeline of AudioLDM (Liu et al., 2023a), including a pre-trained variational autoencoder (VAE) (Kingma & Welling, 2014) $V = \{V_E, V_D\}$ and a pre-trained Hifi-GAN (Kong et al., 2020) vocoder G. V_E and V_D are the encoder and the decoder, respectively. We first extract the 64-band Mel-spectrogram with the hop size of 10 ms from audio: $m \in R^{T \times F}$, where $T = s \cdot 1000/10$ and $F = 64$. Given that all our training samples are 16 kHz 10-second audios, $s = 160000$. The VAE encoder V_E then encodes m into a 2-D map of continuous latents $v \in R^{\frac{T}{r} \times \frac{F}{r} \times c}$ where $r = 4$ is the compression level and $c = 8$ is the latent dimension. To reduce the sequence length and the training cost, we further patchify v into $\bar{x} \in R^{\frac{T}{r \cdot p} \times \frac{F}{r \cdot p} \times (c \cdot p^2)}$ by stacking the latents in a $p \times p$ neighborhood. We set $p = 4$. Finally, we flatten $\bar{x}$ into a 1-D sequence $x = \{x^1, ..., x^n\}$ in a row-major order, where $n = \frac{T}{r \cdot p} \times \frac{F}{r \cdot p}$ and $x^i \in R^{c \cdot p^2}$. That is, different frequency components are placed together in the flatten sequence. With the tokenization, we represent a 10-second clip by 256 continuous-valued tokens with the latent size 128.

De-tokenize. Given the sampled, flattened tokens $\tilde{x} = \{\tilde{x}^1, ..., \tilde{x}^n\}$, we undergo the reverse process of flatten and patchify to get the 2-D latent map $\tilde{v} \in R^{\frac{T}{r} \times \frac{F}{r} \times c}$. The VAE decoder V_D then decodes $\tilde{v}$ into the sampled Mel-spectrogram $\tilde{m}$. Finally, the vocoder G synthesizes waveforms $\tilde{a}$ from $\tilde{m}$.

F.2. Conditional audio language modeling

Diffusion head. We follow (Li et al., 2024b) for its diffusion process, including the cosine noise schedule, training/sampling (1000/100) diffusion steps, and the prediction target (noise). We reuse the MLP architecture introduced in (Li et al., 2024b). The MLP of the Base model has 3 layer, and that of the Large model has 6 layer. Both versions use 1024 dimension for all the MLP layers.

Transformer. We use the Transformer (Vaswani, 2017) implementation in ViT (Wang et al., 2021), same as MAR (Li et al., 2024b). The Base model adopts 24 layers with the hidden dimension of 768; the Large model adopts 32 layers with the hidden dimension of 1024. The model sizes are about 170M and 420M, respectively. Since we use the same architecture, we can leverage the image pre-trained weights in MAR. The image pre-trained weights help the spectrogram-based audio models, as suggested by AST (Gong et al., 2021). We reuse only the weights of the backbone Transformer, excluding those of the diffusion head, since the latter is entangled to the modality-specific token dimension.

Text prompt. We use CLAP (Wu et al., 2023) and FLAN-T5 (Chung et al., 2024) to extract the text embeddings and concatenate them as the input prompt. Our ablation study shows that using both is slightly better than using either individually. In practice, FLAN-T5 produces variable-length embeddings, which we truncate to the first 77 embeddings. Combined with a single CLAP embedding and the linear projections, our text conditioning is equal to 78 tokens, which we use as the initial input prompt in the causal LM for audio generation. The positions of the text prompt do not contribute to the loss.

G. Evaluation

G.1. Objective evaluation

We evaluate our model on the AC evaluation set. Each audio is labeled by 5 captions, and we follow AudioLDM to randomly select one text description as the input prompt. We leverage the AudioLDM evaluation toolkit[31]. Given the generated audio and the ground-truth audio in AC, we calculate several metrics: Fréchet Audio Distance (FAD) (Kilgour et al., 2019), Fréchet Distance (FD), Kullback–Leibler divergence (KL), Inception Score (IS) (Liu et al., 2023a), and Contrastive Language-Audio Pretraining (CLAP) score (Huang et al., 2023b). Intuitively, lower FD and FAD indicate higher similarity to the paired ground-truth. The lower KL indicates that the set of the generated audios share the similar distribution as the ground-truth audios. The higher IS indicates that given an audio classifier, each audio is well-recognized by the classifier[32], and the classifier classifies all the generated audios into diverse classes[33]. Intuitively, these two conditions jointly imply the generated audios are diverse and authentic. Finally, the higher CLAP score indicates the better alignment between the generated audio and the input text prompt.

G.2. Subjective evaluation

We follow a similar approach to Tango (Ghosal et al., 2023), we assessed 20 generated audio samples based on text relevance (REL) and overall quality (OVL) but used a 1-to-5 rating scale instead of a 1-to-100 scale. Each sample was rated by at least 10 participants. We separate the evaluation into speech and non-speech categories to understand the differences in model behaviors. The speech prompts are sampled by selecting the text prompts containing man, woman, and person and people. The non-speech prompts are sampled from the remaining. Note that in the AudioCaps evaluation set, most prompts consist of both speech and non-speech descriptions. We compare our method to the existing decoder-only solution AudioGen and two SOTA diffusion models AudioLDM 2 and Tango 2.

[31] https://github.com/haoheliu/audioldm_eval

[32] Indicated by the low entropy of the output distribution.

[33] Indicated by the high entropy of the average output distribution over the generated audios.