Masked Text-to-Audio Flow-Matching and Reward Feedback Optimization

Rongjie Huang[1], **Dongchao Yang**[1], **Wenxiang Guo**[2], **Huadai Liu**[2], **Xize Cheng**[2], **Zehan Wang**[2], **Zhou Zhao**[2], **Xixin Wu**[1], **Helen Meng**[1]
The Chinese University of Hong Kong[1], Zhejiang University[2]

Abstract

Flow-matching generative models have created significant milestones in text-to-audio generation, powered by scalable training with increased data, computational resources, and model size, while their scalable inference remains less explored. In this work, we propose MaskAudioFlow, a continuous flow-matching transformer with masked generative modeling designed for scaling text-to-audio inference-time prediction. Specifically, MaskAudioFlow 1) masks spans of audio frames in training and approximates the continuous velocity vector field with flow-matching objective, and 2) performs inference via masked prediction, where we mask out generation and re-predict them through iterative decoding. To reduce the gap between generation and human preferences, we fine-tune MaskAudioFlow using reward signals from text-audio correspondence and perceptual aesthetics. Experimental results demonstrate that MaskAudioFlow achieves state-of-the-art performance in text-to-audio generation, effectively scaling inference-time computation through iterative masked prediction. Moreover, the preference-tuned model demonstrates superior text-audio alignment faithfulness and enhanced perceptual aesthetics. [1]

1 Introduction

Flow-matching generative models (Goodfellow et al., 2020; Lipman et al., 2022; Ho et al., 2020) have recently exhibited high-quality generation across various domains, emerging as a powerful generative modeling technique for high-dimensional data. With scalable training data and model design, image and video flow generative models demonstrate remarkable improvements for creating long context audios (OpenAI., 2024), high-resolution images (Esser et al., 2024), and videos with different aspect ratios (Polyak et al., 2025).

Audio flow-matching models (Karras et al., 2022) also have demonstrated superior fidelity and controllability due to their ability to scale training by increasing data, computational resources, and model size. AudioBox (Vyas et al., 2023) generates various audio modalities scale training data and parameters with large-scale flow-matching generative modeling. Wang et al. (2023) build a strong in-context learning TTS framework leveraging scaling transformer architecture and web-scale training data. Despite the success achieved in scaling data usage and training computation in audio generative models, the inference-time scaling behaviors are relatively less explored.

Scaling inference-time prediction in audio generative models can be approached via 1) increasing the number of denoising steps (Karras et al., 2022; Salimans and Ho, 2022), while the performance gains typically flatten after a few dozen steps; or 2) iterative decoding (Sun et al., 2024; Ziv et al., 2024) which inherently relies on discrete tokens prediction for masked prediction, while the discretization often leads to a information loss.

In this work, we propose MaskAudioFlow, a masked flow-matching transformer approximating the continuous velocity vector field for text-to-audio generation. Motivated to scale inference-time prediction in audio generative models, MaskAudioFlow masks spans of audio frames in training and generates samples via iterative decoding in inference, where we mask out some generation and re-predict them. To reduce the gap between generation and human preferences, we fine-tune MaskAudioFlow with reward-weighting flow-matching objective from text-audio correspondence and perceptual aesthetics, without the need for reward's gradients or filtered datasets (in Tango 2 (Majumder et al., 2024)).

Both subjective and objective evaluations demonstrate that MaskAudioFlow achieves state-of-the-art results in text-to-audio generation with natu-

[1]Audio samples are available at https://MaskAudio.github.io/

Findings of the Association for Computational Linguistics: ACL 2026, pages 37953–37966

ral and faithful synthesis, and MaskAudioFlow effectively scale inference-time prediction through iterative masked prediction. Furthermore, the preference-tuned models exhibit superior text-audio alignment faithfulness and enhanced perceptual aesthetics.The key contributions are as follows:

- We present MaskAudioFlow, a masked flow-matching transformer approximating the continuous velocity vector field for text-to-audio generation.
- We scale inference-time prediction with masked generative modeling, where MaskAudioFlow masks out some generations and re-predict them in an iterative refinement manner.
- We reduce the gap between generation and human preference (text-audio correspondence, perceptual aesthetics) through reward feedback fine-tuning.
- Experimental results present state-of-the-art results in text-to-audio generation, with perference text-audio alignment faithfulness and enhanced perceptual aesthetics.

2 Related Works

2.1 Text-to-Audio Generation

Text-to-Audio generation is a multimodal task that has witnessed notable advancements in recent years. Diffsound (Yang et al., 2022) leverages a pre-trained VQ-VAE (Van Den Oord et al., 2017) on mel-spectrograms to encode audio into discrete codes, subsequently utilized by a diffusion model for text-to-audio synthesis. Another series of models (Huang et al., 2023b; Liu et al.) rely on the Latent Diffusion Model (LDM), which significantly enhances sample quality. Auffusion (Xue et al., 2024) leverages the text-to-image system's inherent generative strength and precise cross-modal alignment to generate audios that accurately match textual descriptions. Mo et al. (2024) propose a multimodal text-to-audio model aligned with video, which employs an audio-visual control net to adeptly merge temporal visual representations with text embeddings. In this work, we study the inference-time behaviors of audio generative models with continuous masked prediction, to iteratively predict samples via masked refinement.

2.2 Flow Matching Generative Models

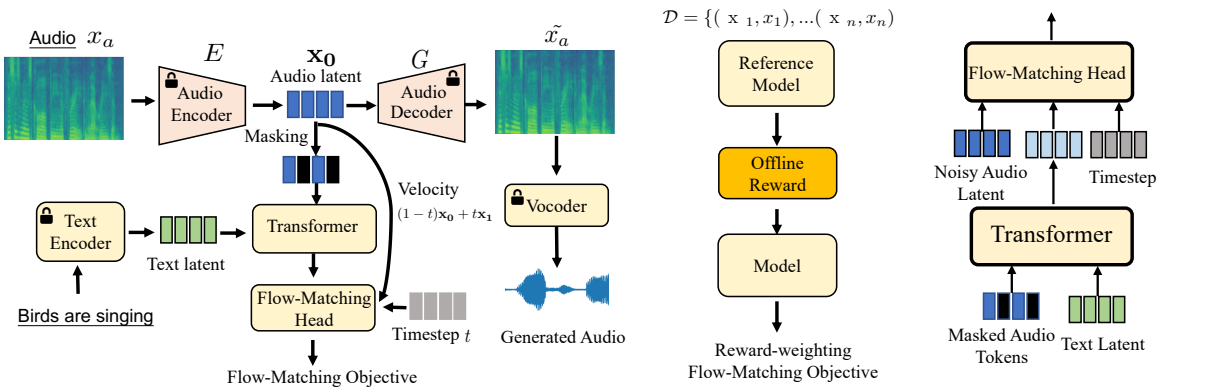
Flow matching Lipman et al. (2022) is a synthesis framework using ODEs that unifies and extends probability flow with a simple vector-field regression loss. Further leveraging ideas from optimal transport, rectified flows (Lee et al., 2023c; Liu et al., 2022) aim to minimize the trajectory curvature and further connect data and noise on a straight line. Flow-matching models (Vyas et al., 2023; Wang et al., 2024) have simpler formulations and fewer constraints but better quality compared to diffusion models (Ho et al., 2020; Xiao et al., 2021), demonstrating improvement in audio generation. In this work, we approximate the velocity vector field with continuous flow-matching objective, enhancing the model's capabilities in high-fidelity audio generation.

2.3 Learning from Human Feedback

There is often a gap between generative models' training objectives and human preference, and thus human feedback has been utilized to align model performance with user intent to improve the performance in downstream tasks. DiffusionDPO (Wallace et al., 2024) adapts the Direct Preference Optimization (DPO) (Rafailov et al., 2023) and aligns diffusion models to human preferences by directly optimizing on human comparison data. A series of work DRAFT (Clark et al., 2023) shows that diffusion models can be finetuned directly for downstream differentiable reward models using end-to-end backpropagation, and AlignProp (Prabhudesai et al., 2023) leverages a randomized number of steps of the denoising process and demonstrates the reduced over-optimization. In text-to-audio generation, Tango 2 (Majumder et al., 2024) fine-tunes the model using diffusion-DPO loss on the constructed preference dataset and show that it leads to improved audio quality. Differently, we finetune MaskAudioFlow with reward-weighting flow-matching objective (Lee et al., 2023a), guiding the model to prioritize high-reward samples in the data manifold, which doesn't rely on gradients of rewards or filtered datasets.

3 Method

In this section, we first overview MaskAudioFlow and discuss the motivation for scaling inference-time prediction via masked prediction. Next, we discuss the masking strategies including choosing the span of tokens and masking ratio, following



(a) Pipeline Overview **(b) Reward fine-tuning** **(c) MaskAudioFlow**

Figure 1: MaskAudioFlow overview. In subfigure (a), the modules printed with a *lock* are frozen for training the flow transformer. In subfigure (c), MaskAudioFlow leverages the cross-attention transformer to learn text-audio correspondence and applies a transformation head for calculating the continuous flow-matching objective.

which we present the audio/text representation for text-to-audio modeling. Lastly, a novel transformer architecture with flow-matching head is introduced to inject conditions and learn text-audio correspondence.

3.1 Overview

As illustrated in Figure 1, MaskAudioFlow consists of the following main components: 1) VAE to encode spectrogram into a latent and convert it back to spectrogram; 2) text encoder to derive high-level textual representation, 3) flow-matching transformer that masks span of continuous audio latent in training, and inference with ODE sampler via iterative decoding, and 4) separately-trained neural vocoder to convert mel-spectrograms to raw waveforms. In the following sections, we describe these components in detail.

3.2 Motivation

As stated before, audio flow-matching models (Vyas et al., 2023; Wang et al., 2024) have demonstrated superior performance due to their ability to scale training by increasing data, computational resources, and model size, while their **inference-time behaviors** are less explored.

MaskAudioFlow is motivated to scale inference-time prediction in audio generative models. In this work, we scale inference with masked generative modeling, where MaskAudioFlow masks out generation randomly and re-predict them in an iterative refinement manner.

3.3 Masking Strategy in Training

To enable iterative refinement in inference where masked audio "tokens" are predicted conditioning

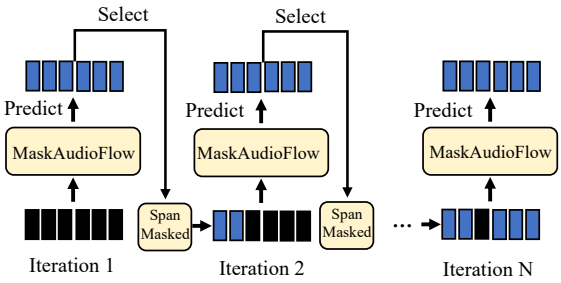


Figure 2: Iterative decoding. We mask out generation randomly and re-predict them in the predict-select iterative manner.

on the unmasked regions, MaskAudioFlow needs to be trained using teacher forcing (i.e., masked prediction (Chang et al., 2022; Li et al., 2023)). In training, we mask the continuous audio latents and compute the flow-matching training objective only in masked positions. Here we describe the following design spaces, and refer the reader to ablation section for a detailed discussion of our findings.

- **Span of tokens**. Adjacent audio tokens are often extracted from a temporal context sharing information with significant redundancy, due to the receptive field of the audio encoder. It (Ziv et al., 2024) suggests that masking a span of tokens instead of scattered tokens is more efficient for model optimization. As such, we validate by masking and predicting spans of tokens with a length of $l = 4$ as the atomic building block in time.

- **Masking ratio**. With bidirectional attention, we can predict multiple frames of unknown tokens given any number of known tokens

simultaneously at training time. Designing a pretext task with proper hardness (Huang et al., 2022) is important for effective masked generative modeling. In this work, we use a masking ratio of $p = 0.6$ to randomly mask 60% of the tokens in training.

3.4 Iterative Decoding in Inference

Here we describe the inference-time iterative decoding in Algorithm 1, where we mask out generation and re-predict them in an iterative manner. For each iteration, the algorithm runs as follows:

- **Predict**. To sample with the flow-matching transformer, the model predicts all the tokens in parallel by solving the probability flow ODE backward in time. By default, we use torchdiffeq (Chen et al., 2018) package to implement the ODE solvers with 25 steps.

- **Select**. It (Li et al., 2025; Fan et al., 2024) is reported that randomly select the tokens to be predicted presents improvement over raster order models, or over relying on a confidence score to indicate the model's belief of a prediction.

- **Masking Schedule**. In each inference iteration, we compute the number of tokens to be masked according to a masking schedule function. It progressively reduces the masking ratio from 1.0 to 0 with a cosine schedule (Chang et al., 2022; Li et al., 2023). By default, we use 8 steps in this schedule.

3.5 Audio Tokens

Most masked generation models typically rely on tokenization to transform samples into a set of discrete tokens with a finite vocabulary and then estimate a categorical distribution over the vocabulary, while such discretization (Sun et al., 2024; Chang et al., 2023) often leads to a loss of information from the sample. In this work, we use continuous audio latents for masked generative modeling by applying a transformation head to approximate the continuous distribution and calculate the flow-matching loss objective. It (Li et al., 2025; Fan et al., 2024) eliminates the need for vector quantization, allowing to model audio with continuous tokenizers which yield much better reconstruction quality.

3.6 Architecture

We present MaskAudioFlow's architecture in Figure 1(c) and a detailed illustration is attached in Appendix A.

Cross-attention transformer. We leverage the contrastive language-audio pretraining (CLAP) (Elizalde et al., 2023) representation for text, and we leverage the cross-attention module (Rombach et al., 2022; Podell et al., 2023) with rotary positional embedding (RoPE) (Su et al., 2024; Heo et al., 2024) for injecting temporal positional embedding into the model, which enables the model to grasp the temporal relationships between successive frames.

Flow-matching transformation head. To avoid quantization in masked generative modeling, we use a lightweight flow-matching head to model the per-token distribution. As shown in Figure 1, the flow-matching head is a smaller transformer with fewer blocks. For integrating denoising timestep t into the model, we leverage the adaptive layer normalization (AdaLN) (Peebles and Xie, 2023; Ma et al., 2024b).

Algorithm 1 Iterative Decoding

Dataset: Text descriptions $\mathcal{D} = \{\text{txt}_1, \ldots, \text{txt}_n\}$
Required: MaskAudioFlow $\mathbf{v}_\theta$ with cross-attention transformer w, flow-matching head w_f; number of decoding iterations s; span length L; ODE sampler ODE with steps T. REPEATINTERLEAVED denotes the repeat interleaved operation.
$N_{\text{spans}} \leftarrow T/L$
$m^0_{\text{spans}} \leftarrow \text{ONES}(T), \quad x \leftarrow \text{ZEROS}(T)$ ▷ Initialize span mask and output tokens
for $i = 0$ to s **do**
  $m^i \leftarrow \text{REPEATINTERLEAVED}(m^i_{\text{spans}}, L)$
  $z \leftarrow w(\text{txt}, y, m^i)$
  $p \leftarrow \cos\left(\frac{\pi+i}{2s}\right)$
  $N_{\text{mask}} \leftarrow \max\left(\lfloor p \cdot N_{\text{spans}} \rfloor, 1\right)$
  $m^{i+1}_{\text{spans}} \leftarrow \text{RANDOMSELECT}(N_{\text{mask}})$ ▷ Sample new mask spans
  $\tilde{x} \leftarrow \text{ODE}_{w_f}(z, T)$
  $\tilde{m}_{\text{spans}} \leftarrow m^i_{\text{spans}} \oplus m^{i+1}_{\text{spans}}$
  $\tilde{m} \leftarrow \text{REPEATINTERLEAVED}(\tilde{m}_{\text{spans}}, L)$
  $x[\tilde{m}] \leftarrow \tilde{x}[\tilde{m}]$ ▷ Update selected positions
end for
return x

4 Reward Feedback Fine-tuning

In this section, we first overview the motivation for reward fine-tuning and discuss the reward-weighting objective, following which we introduce CLAP (Elizalde et al., 2023) and audio aesthetic (Tjandra et al., 2025) as reward models to respectively align with human preference towards text-audio correspondence and perceptual aesthetics.

4.1 Motivation

There is often a gap between generative models' training objectives and human preference (Prabhudesai et al., 2023; Clark et al., 2023), and thus human feedback can been utilized to align model performance with human's intent. In this section, we leverage reward-weighting finetuning to adjust the probability of selecting samples proportionally to human preference rewards, encouraging the model to focus on the high-reward generation and mitigate the preference gap.

4.2 Reward-weighting Flow-matching

Tango 2 (Majumder et al., 2024) fine-tunes the model using diffusion-DPO (Wallace et al., 2024) loss on the constructed preference dataset with respective preferred (winner) and undesirable (loser) audios, while a major limitation of this approach is its heavy reliance on a filtered dataset containing both positive and negative samples. Additionally, DPOK series of work (Wu et al., 2024; Clark et al., 2023) backpropagate through a differentiable reward function in the sampling process, while the feedback from inference-time gradients requires a careful design to avoid policy collapse and maintain diversity.

To reduce the requirement of positive-negative pairs, we fine-tune MaskAudioFlow without relying on gradients of rewards or filtered datasets. Specifically, we 1) generate audio from training dataset using the trained MaskAudioFlow, and predict the human preference score with reward models, and 2) fine-tune MaskAudioFlow with reward-weighting flow-matching objective (Lee et al., 2023a) to prioritize high-reward samples in the data manifold, where samples that yield higher rewards are assigned greater importance, and the policy is updated by re-weighting the flow-matching objective loss:

$$\mathcal{L}_{CFM} = \min_{\theta} \mathbb{E}_{t, p_t(\mathbf{x}|\mathbf{x}_1)} \left[w(\mathbf{x}_1) \| \mathbf{v}_\theta^{\text{ft}}(\mathbf{x}, t) - \mathbf{u}_t(\mathbf{x} \mid \mathbf{x}_1) \|^2 \right], \tag{1}$$

where $w(\mathbf{x}_1)$ is a weighting function, and we have a weighting function $w(\mathbf{x}_1) \propto r(\mathbf{x}_1)$. Assuming an exponential weighting function $w(\mathbf{x}_1) = \exp(\tau * r(\mathbf{x}_1))$ case, τ can control the collapse speed of learned policy, and $\tau \to \infty$ also induces a policy collapse problem.

Typically, the diversity of the model-generated dataset is limited and can result in overfitting. To mitigate this, we further include a regularization loss $\mathcal{L}_{Reg}$ by penalizing divergence from the pre-trained model $\mathbf{v}_\theta^{\text{ref}}$, thus maintaining diversity in the learned model with final objective $\mathcal{L}_{Reg} = \left\| \mathbf{v}_\theta^{\text{ft}}(\mathbf{x}, t) - \mathbf{v}_\theta^{\text{ref}}(\mathbf{x}, t) \right\|^2, \mathcal{L} = \mathcal{L}_{Reg} + \mathcal{L}_{CFM}$.

4.3 Reward Model

CLAP (Elizalde et al., 2023) brings audio and text descriptions into a joint space and demonstrates the zero-shot generalization to multiple downstream domains. For reward function, we have $r(\mathbf{x}_1) = \text{CLAP}(\mathbf{x}_1, \text{txt})$, where CLAP score CLAP is defined as the cosine similarity between the text and audio embeddings with respect to human preferences.

Audio aesthetic (Tjandra et al., 2025) proposes new annotation guidelines that decompose human listening perspectives into four distinct axes. They train no-reference, per-item prediction WavLM (Chen et al., 2022) that assesses audio quality. For the reward function, we calculate the aesthetic score averaged across Production Quality (PQ), Production Complexity (PC), Content Enjoyment (CE), and Content Usefulness (CU).

Algorithm 2 Reward Feedback Fine-tuning

Dataset: Text-audio pairs with preference scores from a reward model:

$$\mathcal{D} = \{(\text{txt}_1, x_1, r_1), \ldots, (\text{txt}_n, x_n, r_n)\}$$

Required: Pre-trained parameters $\mathbf{v}_\theta^{\text{ref}}$; regularization parameter τ

Initialization: Number of ODE sampler steps T; time step range $[T_1, T_2]$

for $(\text{txt}, \mathbf{x}, r) \in \mathcal{D}$ **do**

  $w(\mathbf{x}) \leftarrow \exp(\tau \cdot r)$

  $t \leftarrow \text{UNIFORM}(0, 1)$

  $\mathcal{L}_{\text{CFM}} \leftarrow w(\mathbf{x}) \cdot \| \mathbf{v}_\theta(\mathbf{x}_t, t) - \mathbf{u}_t(\mathbf{x} \mid \mathbf{x}_t) \|^2$

  $\mathcal{L}_{\text{REG}} \leftarrow \left\| \mathbf{v}_\theta(\mathbf{x}, t) - \mathbf{v}_\theta^{\text{ref}}(\mathbf{x}, t) \right\|^2$

  $\mathcal{L} \leftarrow \mathcal{L}_{\text{REG}} + \mathcal{L}_{\text{CFM}}$

end for

5 Training and Evaluation

5.1 Dataset

Following benchmark studies (Yang et al., 2022; Kreuk et al., 2022), we use the training split of Audiocaps dataset (Kim et al., 2019) to train our models. For evaluating text-to-audio models, the Audiocaps test set is adopted as the standard benchmark. We conduct preprocessing on the text and audio data: 1) convert the sampling rate of audios to 16kHz; 2) extract the spectrogram with the FFT size of 1024, hop size of 256 and crop it to a mel-spectrogram of size 80×624; For text-to-music generation, we exclusively employ the LP-Musicaps dataset for training endeavors.

5.2 Model Configurations

We train a continuous VAE to compress the perceptual space with downsampling to a 20-channel latent representation, which balances efficiency and perceptually faithful results. We use 4 V100 GPUs for our main experiments until 1M optimization steps. We utilize BigVGAN (Lee et al., 2023b) trained on Audioset dataset (Gemmeke et al., 2017) as the vocoder to synthesize waveform from the generated mel-spectrogram in all our experiments. The base learning rate is set to 0.005, and we scale it by the number of GPUs and batch size.

In masked decoding, we use iterations of 8 and classifier-free guidance of 3 by default. For flow-matching sampling, we use torchdiffeq (Chen et al., 2018) package to implement 25-step ODE solvers with a step size of 0.04.

5.3 Evaluation Metrics

We evaluate models using objective and subjective metrics over audio quality and text-audio alignment faithfulness. The key automated performance metrics used are KL divergence (↓), and Frechet audio distance (FAD) (↓) to measure audio fidelity. The alignment accuracy (CLAP score) (↑) (Elizalde et al., 2023) is adopted to measure the text-audio alignment faithfulness. The aesthetic scores Tjandra et al. (2025) conduct audio aesthetic evaluation offers a more nuanced assessment of audio quality with Production Quality (PQ), Production Complexity (PC), Content Enjoyment (CE), and Content Usefulness (CU).

For subjective metrics, we use crowd-sourced human evaluation via Amazon Mechanical Turk, where raters are asked to rate MOS (mean opinion score) on a 20-100 Likert scale. We assess the audio quality and text-audio alignment faithfulness by scoring MOS-Q and MOS-F, which are reported with 95% confidence intervals (CI). Detailed information is included in Appendix D.

5.4 Baseline models

We conduct a comparative analysis of the quality of generated audio samples and inference latency across various systems, including GT (i.e., ground-truth audio), AudioLDM 2 (Liu et al.), TANGO 2 (Majumder et al., 2024), Make-An-Audio 2 (Huang et al., 2023a), SoundCTM (Saito et al., 2024), and our MaskAudioFlow model. As a baseline, we also apply the conditional flow matching model (CFM) using the same architecture (without masked generative modeling). We follow their pretrained checkpoints and the number of sampling iterations. The results are compiled and presented in Table 1. For text-to-music generation, we attach the results in Appendix B.

6 Results

6.1 Text-to-Audio Generation

Automatic Objective Evaluation The objective evaluation comparison is presented in Table 1, and we have the following observations: 1) In terms of audio quality, MaskAudioFlow achieves the highest perceptual quality in AudioCaption with FAD of 1.10 and KL of 1.30, demonstrating the SOTA performance of masked autoregressive flow-matching model compared to different baselines. 2) On text-audio similarity, MaskAudioFlow scores a high CLAP with a gap of 0.022 compared to the ground truth audio, suggesting its capability to generate faithful audio that aligns well with descriptions.

Subjective Human Evaluation We also include a human evaluation in Table 1: MaskAudioFlow consistently achieves a high perceptual quality and text-audio alignment with MOS-Q of 78.87 and MOS-F of 77.16. It indicates that raters prefer MaskAudioFlow synthesis against baselines in terms of audio naturalness and faithfulness.

Masked CFM compared to CFM. We scale inference with masked generative modeling, where MaskAudioFlow masks out generation and re-predict them in iterative decoding. Compared to the conditional flow matching models (CFM), MaskAudioFlow achieves a higher perceptual quality with the improvement of 0.12 FAD and KL of 0.04, demonstrating the effectiveness of masked CFM. In Figure 3, masked decoding from 1 to

Model	FAD (↓)	KL (↓)	CLAP (↑)	PQ (↑)	PC (↑)	CE (↑)	CU (↑)	MOS-Q(↑)	MOS-F(↑)
GT	/	/	0.674	5.56	3.24	3.44	4.94	87.90	85.48
AudioLDM 2	1.90	1.48	0.622	5.29	2.97	3.25	4.56	73.38	71.22
Make-An-Audio 2	1.80	1.32	0.645	5.32	3.04	3.19	4.66	75.56	73.14
Tango 2	2.84	1.37	**0.680**	5.29	**3.13**	**3.40**	4.87	73.46	76.08
CFM	1.22	1.34	0.640	5.41	3.05	3.26	4.77	76.32	74.75
SoundCTM	1.95	1.36	0.656	5.37	3.01	3.26	4.82	73.87	71.16
MaskAudioFlow	**1.10**	**1.30**	0.657	**5.58**	3.10	3.38	**4.91**	**78.87**	**77.16**

Table 1: Text-to-audio model results. PQ, PC, CE, and CU respectively denote the production quality, production complexity, content enjoyment, content usefulness in aesthetic evaluation.

Model	Reward	FAD (↓)	KL (↓)	CLAP (↑)	PQ (↑)	PC(↑)	CE (↑)	CU(↑)
GT	/	/	/	0.670	5.56	3.24	3.44	4.94
MaskAudioFlow	/	**1.10**	1.30	0.657	5.58	**3.10**	3.38	4.91
MaskAudioFlow	CLAP	2.65	1.29	**0.671**	5.61	3.03	3.38	4.95
MaskAudioFlow	Aesthetic	1.93	**1.27**	0.664	**5.63**	3.05	**3.43**	**4.97**

Table 2: Fine-tuned text-to-audio models with reward feedback optimization.

8 iterations enhance generation quality with a reduction in FAD and KL, which also verifies the effectiveness of masked CFM for scaling inference computation.

6.2 Qualitative Findings

To investigate the behaviors of scaling inference-time prediction via masked generative modeling, we plot the FAD, KL, and CLAP w.r.t iterative decoding steps and illustrate them in Figure 3(a), showing that 1) more iterative refinements consistently enhance generation quality through masked prediction, leading to a improvement in automatic metrics; 2) further increasing the number of iterative decoding steps will yield a distinctly slower improvements.

We visualize the audio and spectrograms at different masked prediction steps (i.e., $1, 3, 5, 7$) in Figure 4(a) showing that: 1) generation begins with masked tokens, which are decoded with a masked schedule until the entire sequence is generated, and 2) varying the number of masked decoding steps offers a flexible mechanism to balance generation quality and latency.

6.3 Perference Text-to-Audio Models

As there is often a gap between generative models' training objectives and preference, the human feedback has been utilized to align model performance with human's intent.

Text-audio alignment faithfulness. Because of the noisy annotation presented in pre-training datasets, there could be a misalignment between the semantics of the generated audio and the associated text prompts. In preference fine-tuning, we conduct reward-weighting using contrastive audio-language pretraining encoders (CLAP) Elizalde et al. (2023) as reward models. The CLAP model learns audio concepts from natural language supervision to bring audio and text descriptions into a joint multimodal space. As can be seen in Table 2, MaskAudioFlow (CLAP) presents a higher text-audio alignment with a similarity score of 0.671, suggesting the effectiveness of reward feedback fine-tuning in learning text-audio correspondence towards while maintaining audio quality.

Aesthetic score. Audio aesthetics (Tjandra et al., 2025) are inherently subjective and deeply intertwined with human perception. In preference fine-tuning, we leverage the aesthetic score prediction model (Tjandra et al., 2025) as rewards to conduct preference reward-weighting. It is a no-reference, per-item prediction model that offers a more nuanced assessment of audio quality. As shown in Table 2, Production Quality (PQ), Content Enjoyment (CE), Content Usefulness (CU) are promoted by scores of $0.05, 0.05, 0.06$ across the testing set compared to the pretrained model. The audio aesthetics predict utterance-level perceptual quality and cover a wide range of human perceptual dimensions, further empowering models to create rich and diverse human-preference audio content.

6.4 Ablation Studies

To verify the effectiveness of several designs in MaskAudioFlow, including masking strategy p, decoding iteration N, and reward weighting factor τ. We conduct ablation studies and discuss the key findings as follows.

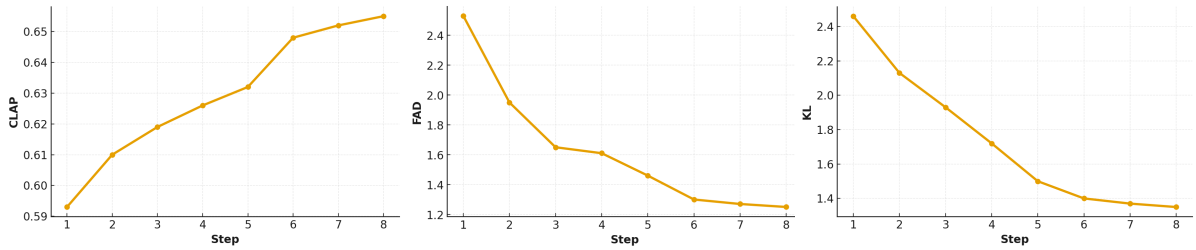
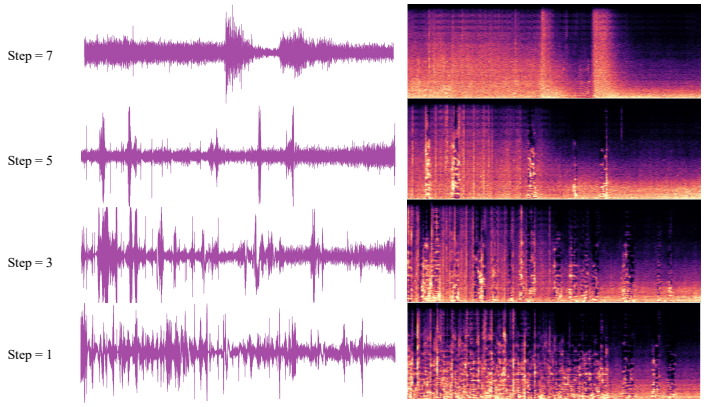


Figure 3: Decoding step analysis for pre-trained masked iterative models, and we use 8 iterations in main experiments.



Model	FAD (↓)	FD (↓)	KL (↓)
Masking Ratio			
$p = 0.7$	2.35	38.1	**1.61**
$p = 0.6$	2.51	43.8	1.81
$p = 0.5$	**2.34**	**37.6**	1.72
Masking Span			
w/o Span	2.47	37.6	1.76
Span 2	2.51	35.2	1.58
Span 3	2.98	37.0	1.65
Span 4	**2.43**	**34.1**	**1.38**

Model	FAD (↓)	FD (↓)	KL (↓)	CLAP (↑)
Reward fine-tuning τ				
$\tau = 0.1$	3.1	**33.0**	**1.40**	**0.641**
$\tau = 1$	3.4	35.8	1.51	0.620
$\tau = 10$	**2.4**	35.5	1.46	0.629

Figure 4: (a) Visualization of iterative decoding, and we use "People speaking with loud bangs followed by a slow motion rumble" as a prompt. Left: waveforms. Right: mel-spectrograms. (b) Ablation results among masking ratio, span, and reward finetuning τ.

Masking Strategy. MaskAudioFlow is trained using teacher forcing (i.e., masked prediction), where masking the span of tokens provides the model with a receptive field in learning audio context. In Table 4(b), using a span-length of 4 yields the best performance with FD of 2.43 and KL of 1.38. It suggests that masking a span of tokens instead of scattered tokens span $= 1$ is more efficient for model optimization as spectrograms are extracted from a temporal context with significant redundancy.

Masking Ratio. For the masking ratio, a higher ratio indicates blocking more information during training, and the task with masked flow-matching will become more difficult while random masking improves steadily up to 70%. In summary, MaskAudioFlow with $p = 0.5$ demonstrates the best audio quality among the choices.

Reward Weighting Factor. For reward feedback fine-tuning, τ controls the convergence speed of learned policy to alleviate overfitting. To explore and ablate the reward weighting factor, we use the CLAP encoder as the reward model in feedback fine-tuning. According to experiments with different factors τ, we have $\tau = 0.1$ achieving the best FD, KL, CLAP score results.

7 Conclusion

In this work, we propose MaskAudioFlow, a masked flow-matching transformer approximating the continuous velocity vector field for text-to-audio generation. Motivated to scale inference-time prediction in audio generative models, MaskAudioFlow masked spans of audio frames in training and generated samples via iterative mask-prediction in inference. To bridge the gap between model and human preferences, MaskAudioFlow was fine-tuned using reweighted reward feedback for better aligning on text-audio correspondence and perceptual aesthetics. Both objective and subjective evaluation demonstrated that MaskAudioFlow achieved state-of-the-art results in benchmark text-to-audio generation, which effectively scaled up inference computing with natural and faithful synthesis. The preference-tuned models exhibited superior text-audio alignment faithfulness and enhanced perceptual aesthetics.

Acknowledgements

This study was supported in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government and National Natural Science Foundation of China (62306260).

Limitation

MaskAudioFlow is a flow-based generative model that produces high-quality audio, which typically requires multiple ODE integration steps and masked iterative refinement. Besides, online RL training typically requires more GPU memory in training. One of our future directions is to develop a lightweight and fast iterative flow-based masked transformer for accelerating sampling.

Ethical Considerations

We recognize that general-purpose audio generation can be misused (e.g., to create harmful or deceptive content). As future work, we will study safeguards such as content filtering, watermarking, and usage policies to reduce the risk of inappropriate outputs.

This work is intended solely for academic research. No commercial deployment is planned at this time, so the near-term ethical risk is limited; nevertheless, we encourage responsible use that respects privacy, consent, and copyright.

References

Michael S Albergo and Eric Vanden-Eijnden. 2022. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, and 1 others. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. 2022. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210. IEEE.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. 2018. Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Kevin Clark, Paul Vicol, Kevin Swersky, and David J Fleet. 2023. Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation. *Preprint*, arXiv:2306.05284.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, and 1 others. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*.

Lijie Fan, Tianhong Li, Siyang Qin, Yuanzhen Li, Chen Sun, Michael Rubinstein, Deqing Sun, Kaiming He, and Yonglong Tian. 2024. Fluid: Scaling autoregressive text-to-image generative models with continuous tokens. *arXiv preprint arXiv:2410.13863*.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. 2024. Rotary position embedding for vision transformer. *arXiv preprint arXiv:2403.13298*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proc. of NeurIPS*.

Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. 2023a. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *Preprint*, arXiv:2305.18474.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023b. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *Preprint*, arXiv:2301.12661.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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)*, pages 119–132.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. 2023a. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2023b. BigVGAN: A Universal Neural Vocoder with Large-Scale Training.

Sangyun Lee, Beomsu Kim, and Jong Chul Ye. 2023c. Minimizing trajectory curvature of ode-based generative models. In *International Conference on Machine Learning*, pages 18957–18973. PMLR.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. 2023. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2142–2152.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. 2025. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models.

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. 2024a. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv preprint arXiv:2401.08740*.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. 2024b. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 564–572.

Shentong Mo, Jing Shi, and Yapeng Tian. 2024. Text-to-audio generation synchronized with videos. *arXiv preprint arXiv:2403.07938*.

OpenAI. 2024. https://openai.com/sora. *In OpenAI blog*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, and 69 others. 2025. Movie gen: A cast of media foundation models. *Preprint*, arXiv:2410.13720.

Mihir Prabhudesai, Anirudh Goyal, Deepak Pathak, and Katerina Fragkiadaki. 2023. Aligning text-to-image diffusion models with reward backpropagation.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Koichi Saito, Dongjun Kim, Takashi Shibuya, Chieh-Hsin Lai, Zhi Zhong, Yuhta Takida, and Yuki Mitsufuji. 2024. Soundctm: Uniting score-based and consistency models for text-to-sound generation. *arXiv preprint arXiv:2405.18503*.

Tim Salimans and Jonathan Ho. 2022. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. 2024. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, and 1 others. 2025. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*.

Aaron Van Den Oord, Oriol Vinyals, and 1 others. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, and 1 others. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. 2024. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, and 1 others. 2023. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. 2024. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv preprint arXiv:2406.00320*.

Xiaoshi Wu, Yiming Hao, Manyuan Zhang, Keqiang Sun, Zhaoyang Huang, Guanglu Song, Yu Liu, and Hongsheng Li. 2024. Deep reward supervisions for tuning text-to-image diffusion models. In *European Conference on Computer Vision*, pages 108–124. Springer.

Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. 2021. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*.

Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. 2024. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. 2022. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2024. Masked audio generation using a single non-autoregressive transformer. *arXiv preprint arXiv:2401.04577*.

A Model Configurations

We list the model hyper-parameters of MaskAudioFlow in Table 3.

A.1 Vocoder

We train a BigVGAN (Lee et al., 2022) vocoder from scratch for the spectrogram to waveform generation. The synthesizer includes the generator and multi-resolution discriminator (MRD). The generator is built from a set of look-up tables (LUT) that embed the discrete representation and a series of blocks composed of transposed convolution and a residual block with dilated layers. The transposed convolutions upsample the encoded representation to match the input sample rate.

Table 3: Hyperparameters of MaskAudioFlow. We use T and F to denote the time and frequency moe layers respectively.

Hyperparameter		
MaskAudioFlow	Transformer Layer	16
	Diffusion Head Layer	4
	Transformer Embed Dim	768
	Transformer Attention Headers	12
	Number of Parameters	160 M
BigVGAN Vocoder	Upsample Rates	[5, 4, 2, 2, 2, 2]
	Hop Size	320
	Upsample Kernel Sizes	[9, 8, 4, 4, 4, 4]
	Number of Parameters	121.6M

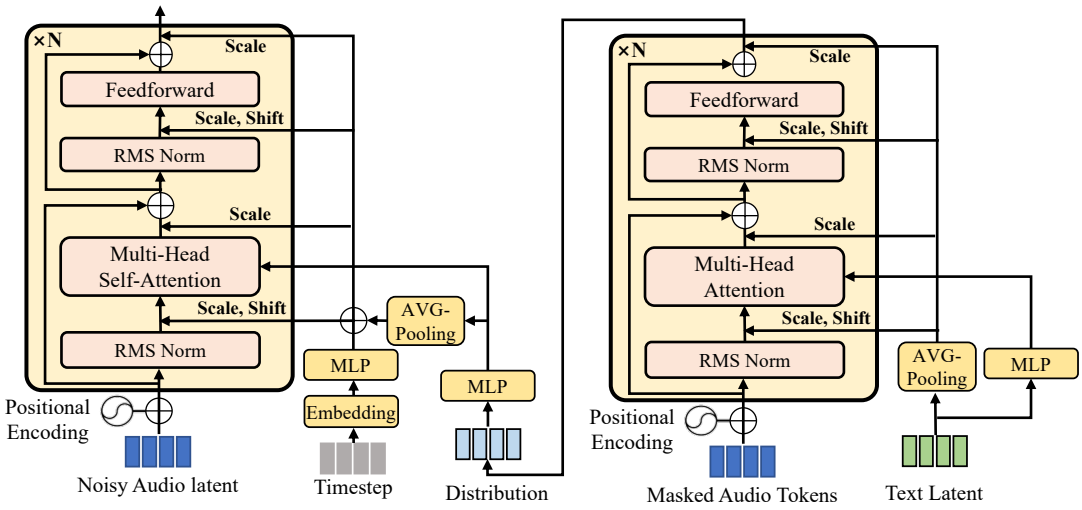


Figure 5: MaskAudioFlow transformer detailed architecture.

A.2 VAE

The continuous spectrogram autoencoder is composed of 1) the audio encoder E takes mel-spectrogram x_a as input and outputs compressed latent $z = E(x_a)$, 2) the audio decoder D reconstructs the mel-spectrogram signals from the compressed representation z; and 3) a discriminator learns to distinguish the generated samples $G(z)$ from real ones in different multi-receptive fields of mel-spectrograms. The training objective is to minimize the weighted sum of reconstruction loss, GAN loss, and KL loss.

B Text-to-Music generation

In this section, we perform a comparative analysis of audio samples generated by FlashAudio against several established music generation systems: 1) GT, the ground-truth audio; 2) MusicGen (Copet et al., 2023); 3) MusicLDM (Chen et al., 2024); 4) AudioLDM 2 (Liu et al.). The results are presented in the Table 4, and we have the following observations:

Model	FAD (↓)	KL (↓)	CLAP (↑)
GT	/	/	0.46
AudioLDM 2	3.81	1.22	**0.43**
MusicGen	4.50	1.41	0.42
MusicLDM	5.20	1.47	0.40
MaskAudioFlow	**3.35**	**1.20**	**0.43**

Table 4: Text-to-music generation comparison.

In terms of audio quality, MaskAudioFlow

achieves the highest perceptual quality with FAD of 3.35 and KL of 1.20, which demonstrates the improved performance of the masked autoregressive flow-matching model compared to previous conditional flow matching or diffusion baselines. This highlights MaskAudioFlow's effectiveness in producing high-quality music samples and its generalization to audio-related domains.

C Preliminaries: Flow-based Generative Models

Here, we give preliminaries on flow generative models. Denote data distribution as p_1, and tractable prior distribution as p_0. Most generative models work by mapping samples $\mathbf{x}_0 \sim p_0(\mathbf{x}_0)$ to data $\mathbf{x}_1$.

A flexible class of generative models (Karras et al., 2022; Song et al., 2020) (i.e., score-based diffusion models) based on turning noise $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$ into data $\mathbf{x}_0$ have been introduced. These models use the time-dependent process $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \varepsilon$, where α_t is a decreasing function of t and σ_t is an increasing function, and set both α_t and σ_t indirectly through different formulations of a stochastic differential equation (SDE).

Common to score-based diffusion models that the process $\mathbf{x}_t$ can be sampled dynamically using SDE, we consider the probability flow ordinary differential equation (ODE) with a velocity field:

$$d\mathbf{x}_t = \mathbf{v}_\theta(\mathbf{x}_t, t)dt, \tag{2}$$

where the velocity $\mathbf{v}$ is parameterized by a neural network θ, and $t \in [0, 1]$. By solving the probability flow ODE backwards in time from $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$, we can generate samples and approximates the ground-truth data distribution $p(x)$. We refer to Eq. 2 as a flow-based generative model.

However, this process is computationally expensive, especially for large network architectures parameterizing $\mathbf{v}_\theta(\mathbf{x}_t, t)$. It is proven that estimating a vector field $\mathbf{u}_t$ that generates a probability path between p_0 and p_1 is equivalent. To construct $\mathbf{u}_t$, we define a forward process, corresponding to a probability path $p_t(\mathbf{x} \mid \mathbf{x}_1)$ with a data sample $\mathbf{x}_t$ between p_0 and $p_1 = \mathcal{N}(0, \mathbf{I})$, with boundary condition $p_{t=0}(\mathbf{x} \mid \mathbf{x}_1) = p_0$ and $p_{t=1}(\mathbf{x} \mid \mathbf{x}_1) = \mathcal{N}(\mathbf{x} \mid \mathbf{x}_1, \sigma^2 \mathbf{I})$ for sufficiently small σ. While regressing $\mathbf{u}_t$ with the *Flow Matching* objective $\mathcal{L}_{FM}$ to learn the velocity field $\mathbf{v}(\mathbf{x}, t)$:

$$\mathcal{L}_{FM} = \min_\theta \mathbb{E}_{t, p_t(\mathbf{x})} \|\mathbf{v}_\theta(\mathbf{x}, t) - \mathbf{u}_t(\mathbf{x})\|^2, \tag{3}$$

For *Conditional Flow Matching*, we have:

$$\mathcal{L}_{CFM} = \min_\theta \mathbb{E}_{t, p_t(\mathbf{x}|\mathbf{x}_1)} \|\mathbf{v}_\theta(\mathbf{x}, t) - \mathbf{u}_t(\mathbf{x} \mid \mathbf{x}_1)\|^2, \tag{4}$$

Rectified Flows (RFs) (Albergo and Vanden-Eijnden, 2022; Liu et al., 2022) define the forward process as straight paths between the data distribution and a standard normal distribution, where we have $\mathbf{x}^{\mathbf{OT}} = (1 - (1 - \sigma_{\min})t)\mathbf{x}_0 + t\mathbf{x}_1$, as the flow from $\mathbf{x}_0$ to $\mathbf{x}_1$.

The objective loss function can be written as:

$$\mathcal{L}_{\text{OT-CFM}} = \min_\theta \mathbb{E}_{t,\, p_t(\mathbf{x}|\mathbf{x}_1)} \big\| \mathbf{v}_\theta(\mathbf{x}^{\text{OT}},\, t) - \mathbf{u}_t^{\text{OT}}\big(\mathbf{x}^{\text{OT}} \mid \mathbf{x}_1\big) \big\|^2. \tag{5}$$

These properties minimize the trajectory curvature and connect data and noise on a straight line. It has been demonstrated that flow probabilistic models (Ma et al., 2024a; Liu et al., 2022; Esser et al., 2024) can learn diverse data distribution in multiple domains, such as images and time series. In this work, we compare the flow formulation to existing DDPM in video-guided audio generative modeling and demonstrate its benefits.

D Evaluation

To probe audio quality, we conduct the MOS-Q (mean opinion score) tests and explicitly instruct the raters to *"focus on examining the audio quality and naturalness."*. The testers present and rate the samples, and each tester is asked to evaluate the subjective naturalness on a 20-100 Likert scale.

To probe video-audio alignment, human raters are shown an audio and a video and asked *"Does the audio align with text faithfully?"*. They must respond with "completely", "mostly", or "somewhat" on a 20-100 Likert scale to score MOS-F.

Our subjective evaluation tests are crowdsourced and conducted via Amazon Mechanical Turk. These ratings are obtained independently for model samples and reference audio. The screenshots of instructions for testers have been shown in Figure. We paid $10 to participants hourly and totally spent about $400 on participant compensation. A small subset of audio samples used in the test is available at https://MaskAudio.github.io/.

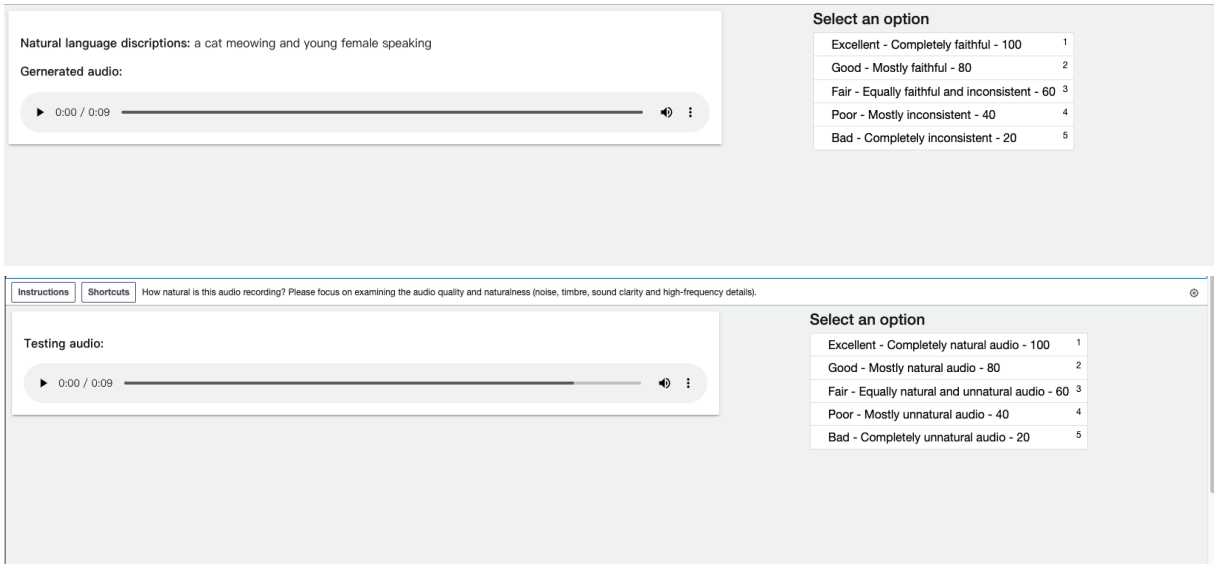


Figure 6: Screenshots of subjective evaluations: (a) MOS-F testing, (b) MOS-Q testing.