FastDiff 2: Revisiting and Incorporating GANs and Diffusion Models in High-Fidelity Speech Synthesis

Rongjie Huang[1], **Yi Ren**[1], **Ziyue Jiang**[1], **Chenye Cui**[1], **Jinglin Liu**[1], **Zhou Zhao**[1*]
Zhejiang University[1]

Abstract

Generative adversarial networks (GANs) and denoising diffusion probabilistic models (DDPMs) have recently achieved impressive performances in image and audio synthesis. After revisiting their success in conditional speech synthesis, we find that 1) GANs sacrifice sample diversity for quality and speed, 2) diffusion models exhibit outperformed sample quality and diversity at a high computational cost, where achieving high-quality, fast, and diverse speech synthesis challenges all neural synthesizers. In this work, we propose to converge advantages from GANs and diffusion models by incorporating both classes, introducing dual-empowered modeling perspectives: 1) FastDiff 2 (DiffGAN), a diffusion model whose denoising process is parametrized by conditional GANs, and the non-Gaussian denoising distribution makes it much more stable to implement the reverse process with large steps sizes; and 2) FastDiff 2 (GANDiff), a generative adversarial network whose forward process is constructed by multiple denoising diffusion iterations, which exhibits better sample diversity than traditional GANs. Experimental results show that both variants enjoy an efficient 4-step sampling process and demonstrate superior sample quality and diversity.[1]

1 Introduction

Speech synthesis has seen extraordinary progress with the recent development of deep generative models in machine learning (Lv et al., 2023b; Ye et al., 2023b; Zhang et al., 2021, 2022c; Li et al., 2023). Previous models (Oord et al., 2016; Kalchbrenner et al., 2018) generate waveforms autoregressively from mel-spectrograms yet suffer from slow inference speed. Non-autoregressive methods (Huang et al., 2022c, 2023a; Ye et al., 2023a; Jiang et al., 2021) have been designed to address this issue, they generate samples with extremely fast speed and achieve comparable voice quality with autoregressive models.

Among them, Generative adversarial networks (GANs) (Creswell et al., 2018; Mao et al., 2019; Jiang et al., 2022) and denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020; Song et al., 2020) are two popular classes of deep generative models that have demonstrated surprisingly good results and dominated speech synthesis: Jang et al. (2021) utilize local-variable convolution to capture different waveform intervals with adversarial learning. Kong et al. (2020a) propose multi-receptive field fusion (MRF) to model the periodic patterns matters. (Kong et al., 2020b) introduce a time-aware wavenet for conditional diffusion modeling. Huang et al. (2022b) and Lam et al. (2022) utilize a noise predictor to learn a tight inference schedule for skipping denoising steps.

Despite their success in the high-fidelity generation, few studies have compared these two classes of deep generative models in conditional speech synthesis. In this work, we conduct a comprehensive study to revisit GANs and diffusion models, and empirically demonstrate that: 1) GANs tend to generate high-quality speeches but do not cover the whole distribution, which sacrifice sample diversity for quality and speed; and 2) diffusion models exhibit outperformed sample quality and diversity, buy they typically require a large number of iterative refinements. To this end, simultaneously achieving high-quality and diverse speech synthesis at a low computational cost has become an open problem for all neural synthesizers.

In this work, we converge advantages from both classes by incorporating GANs and diffusion models, introducing dual-empowered modeling perspectives for high-fidelity speech synthesis: 1) FastDiff 2 (DiffGAN): a **diffusion model** whose denoising process is parametrized by conditional GANs, and the non-Gaussian denoising distribution makes

[*] Corresponding author

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

Findings of the Association for Computational Linguistics: ACL 2023, pages 6994–7009
July 9-14, 2023 ©2023 Association for Computational Linguistics

it much more stable to implement the reverse process with large step sizes; and 2) FastDiff 2 (GANDiff): a **generative adversarial network** whose forward process is constructed by multiple denoising diffusion iterations, which exhibits better sample diversity than traditional GANs. Experimental results show that both variants enjoy an effective 4-iter sampling process and demonstrate the outperformed sample quality and diversity. Moreover, we show that both variants generalize well to the mel-spectrogram inversion of unseen speakers.

The main contributions of this work are summarized as follows:

- We revisit two popular deep generative models (diffusion models and GANs) in conditional speech synthesis, introducing dual-empowered modeling perspectives to converge advantages from both classes.
- FastDiff 2 (DiffGAN) removes the common assumption of Gaussian distribution and utilizes conditional GANs to parametrize the multimodal denoising distribution, implementing the reverse process with large step sizes more stably.
- FastDiff 2 (GANDiff) breaks the one-shot forward of conditional GANs into several denoising diffusion steps in which each step is relatively simple to model, and thus it exhibits better sample diversity than traditional GANs.
- Experimental results show that both enjoy an effective 4-iter sampling process, providing a principled way for high-fidelity and diverse speech synthesis at a low computational cost.

2 Background on Speech Synthesis

With the development of deep generative models (Ye et al., 2023b; Lv et al., 2023a, 2022; Zhang et al., 2022a,b), speech synthesis technology has made rapid progress up to date. Most models (Wang et al., 2017; Ren et al., 2019; Huang et al.; Cui et al., 2021; Huang et al., 2023b; Ye et al., 2022) first convert input text or phoneme sequence into mel-spectrogram, and then transform it to waveform using a separately trained vocoder (Kumar et al., 2019; Kong et al., 2020a; Huang et al., 2022a). In this work, we focus on designing the second-stage model that efficiently synthesizes high-fidelity waveforms from mel-spectrograms.

Neural vocoders require diverse receptive field patterns to catch audio dependencies, and thus previous models (Oord et al., 2016; Kalchbrenner et al., 2018) generate waveforms autoregressively from mel-spectrograms yet suffer from slow inference speed. In recent years, non-autoregressive methods (Prenger et al., 2019; Kumar et al., 2019; Kong et al., 2020b) have been designed to address this issue, which generates samples with extremely fast speed while achieving comparable voice quality with autoregressive models. Below we mainly introduce two popular classes of deep generative models (diffusion models and GANs) for conditional speech synthesis:

2.1 Generative Adversarial Networks

Generative adversarial networks (GANs) (Kumar et al., 2019; Huang et al., 2021) are one of the most dominant non-autoregressive models in speech synthesis. Morrison et al. (2021) propose a chunked autoregressive GAN for conditional waveform synthesis, Lee et al. (2022) utilize a large-scale pretraining to improve out-of-distribution quality, Bak et al. (2022) investigate GAN-based neural vocoders and proposes an artifact-free GAN-based neural vocoder.

The generator G aims to transform noise z into $G(z)$ that mimics real data, while the discriminator D learns to distinguish the generated samples $G(z)$ from real ones. GANs jointly train a powerful generator G and discriminator D with a min-max game:

$$\begin{aligned}\min_G \max_D V(G,D) &= \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[\log(D(\mathbf{x}))] \\ &+\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}[\log(1-D(G(\mathbf{z})))],\end{aligned} \tag{1}$$

However, GAN-based models are often difficult to train, collapsing (Creswell et al., 2018) without carefully selected hyperparameters and regularizers, and showing less sample diversity.

2.2 Diffusion Probabilistic Models

Denoising diffusion probabilistic models (DDPMs) (Ho et al., 2020) are likelihood-based generative models that have recently advanced the state-of-the-art results in most image and audio synthesis tasks. Denote data distribution as $q(\mathbf{x}_0)$, the diffusion process is defined by a fixed Markov chain from data $\mathbf{x}_0$ to the latent variable $\mathbf{x}_T$, which gradually adds noise to the data $q(\mathbf{x}_0)$ in T steps

Model	Quality MOS (↑)	 MCD (↓)	 PESQ (↑)	Speed RTF (↓)	Diversity NDB (↓)	 JS (↓)
GT	4.32±0.06	/	/	/	/	/
GAN	4.08±0.07	**1.48**	3.87	**0.001**	34	0.0016
Diffusion	**4.16±0.09**	1.62	**3.92**	4.70	**22**	**0.0010**

Table 1: Comparison of GANs and diffusion models for speech synthesis. We crowd-source 5-scale MOS tests via Amazon Mechanical Turk, which are recorded with 95% confidence intervals (CI). We implement real-time factor (RTF) assessment on a single NVIDIA V100 GPU.

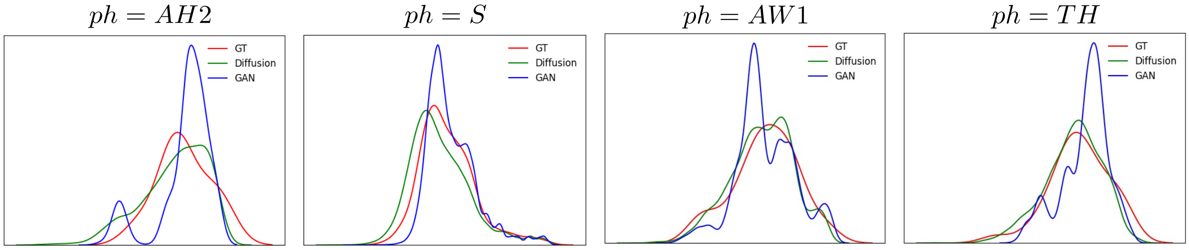


Figure 1: Comparison of sample distribution coverage between diffusion models and GANs. We randomly choose 4 different phonemes ($ph = AH2, S, AW1, TH$) in this case study.

with pre-defined noise schedule β_t:

$$q(\mathbf{x}_1, \cdots, \mathbf{x}_T|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2}$$
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I})$$

The reverse process is to recover samples from Gaussian noises parameterized by shared θ. A guarantee of high sample diversity typically comes at the cost of hundreds of denoising steps:

$$p_\theta(\mathbf{x}_0, \cdots, \mathbf{x}_{T-1}|\mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \tag{3}$$
$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)^2 \boldsymbol{I})$$

It has been demonstrated that diffusion probabilistic models (Dhariwal and Nichol, 2021; Xiao et al., 2021) can learn diverse data distribution in multiple domains, such as images and time series. However, an apparent degradation could be witnessed when reducing reverse iterations, making it challenging to get accelerated.

3 Preliminary Study

In image generation, superior sample diversity (Dhariwal and Nichol, 2021; Ho et al., 2020; Song et al., 2020) is a crucial reason for the diffusion model to produce high-quality samples even on the challenging dataset. Due to the distinctive advantages of diversity and distribution coverage over GANs, diffusion models have been demonstrated to generate realistic and vivid images, achieving the current state-of-the-art measured by FID.

Despite the comprehensive studies of GANs and diffusion models for image generation, few have compared these two classes of deep generative models in speech synthesis, where an audio signal is different (Oord et al., 2016; Kalchbrenner et al., 2018) for its long-term dependencies, high sampling rate, and strong condition. In this section, we provide an empirical study and investigate the characteristic of both classes with close model capacity in speech. Specifically, we evaluate the performance (including sample quality, speed, and diversity) and explore how distribution coverage impacts sample quality by auditory sensation.

3.1 Experimental Setup

We prepare 20 unseen samples from the benchmark LJSpeech dataset (Ito and Johnson, 2017) for evaluation. For a fair comparison, we implement the GAN and diffusion model with a shared backbone (Huang et al., 2022b), which comprises three Diffusion-UBlock and DBlock with the up/downsample rate of [8, 8, 4]. Following the common practice (Kumar et al., 2019; Yamamoto et al., 2020), we remove the time embedding in GAN and introduce an auxiliary multi-resolution STFT loss to stabilize adversarial learning. More information has been attached in Appendix D.1.

3.2 Visualization

We further visualize the marginal distributions $P(\mathbf{x}|ph)$ of diffusion models and GANs in Figure 1. Specifically, we 1) randomly sample 100 latent noises z for each testing audio and obtain

2000 utterances in total. 2) split the generated utterances into phoneme-level samples according to the boundary obtained by forced alignment (McAuliffe et al., 2017) and transform them into linear spectrograms; 3) compute the histograms [2] and smooth them into probability density functions with kernel density estimation for better visualization.

3.3 Analyses

Based on the evaluation results presented in Table 1 and the marginal distributions illustrated in Figure 1, we have the following observations:

Diffusion models demonstrate better sample diversity at the cost of slow inference speed. A more diverse data distribution could be observed in samples generated by diffusion models, demonstrating a better mode convergence. Diffusion models are better at data sharpness, diversity, and matching marginal label distribution of training data. However, sampling from diffusion models often requires thousands of network iterations, which is significantly slower than GAN and makes their application expensive in practice.

GANs trade off diversity for quality and speed. A distinct degradation of mode convergence could be witnessed in GANs, which tend to produce samples but do not cover the whole distribution, indicating a collapsed distribution and less sample diversity. To conclude, GANs sacrifice diversity for quality and speed, while the constrained distribution does not hinder their ability to generate high-fidelity samples. Compared to diffusion models, GANs enjoy high-quality speech synthesis with a minor gap of 0.08 in MOS, while even demonstrating an outperformed performance in MCD evaluation. Regarding inference speed, GANs enjoy an effective one-shot sampling process, significantly reducing the inference time compared with competing diffusion mechanisms.

4 Methods

After revisiting GAN and diffusion models for speech synthesis, we witness that 1) GANs sacrifice sample diversity for better quality and speed, producing high-quality samples but not covering the whole distribution. 2) Diffusion models exhibit outperformed sample quality and diversity, requiring iterative refinement at a high computational cost. In this section, we aim to converge advantages from both classes, introducing dual-empowered modeling perspectives for high-fidelity, fast, and diverse speech synthesis.

4.1 Overview

This section presents our proposed models dually empowered by GANs and diffusion: 1) FastDiff 2 (DiffGAN): a diffusion model whose denoising process is parametrized by conditional GANs, and thus the non-Gaussian denoising distribution makes it much more stable to implement the reverse process with large step sizes; and 2) FastDiff 2 (GANDiff): a generative adversarial network whose forward process is constructed by multiple denoising diffusion distributions, thus exhibiting better sample diversity than traditional GANs.

4.2 Diffusion Mechanism Leveraging GAN

Diffusion models commonly assume that the denoising distribution can be approximated by Gaussian distributions. However, the Gaussian assumption holds only in the infinitesimal limit of small denoising steps, which requires numerous steps in the reverse process. As such, reducing the number of iterative steps always causes a distinct degradation in perceptual quality.

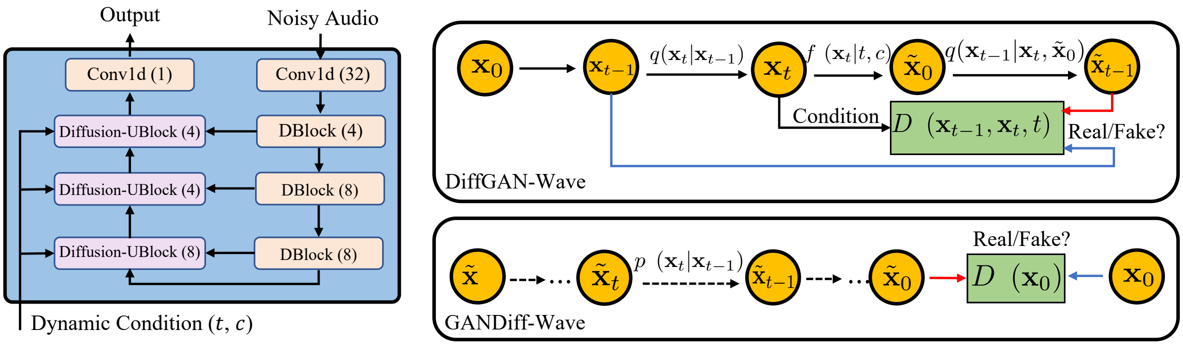
In this work, we propose **FastDiff 2 (DiffGAN)** leveraging conditional GANs to model the denoising distribution $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, and thus the non-Gaussian multimodal distribution makes it much more stable to implement the reverse process with large steps sizes. Specifically, our forward diffusion process is set up with the main assumption that the number of diffusion iterations is small ($T = 4$). The training is formulated by matching the conditional GAN generator $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ and $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ using an adversarial loss that minimizes a divergence D_{adv} per denoising step. The discriminator $D_\phi\left(\mathbf{x}_{t-1}, \mathbf{x}_t, t\right)$ is designed to be diffusion-step-dependent, which supervises the generator to produce high-fidelity speech sample. The min-max objective can be expressed as:

$$\min_\theta \sum_{t\geq 1} \mathbb{E}_{q(t)}\left[D_{\mathrm{adv}}\left(q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right) \| p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)\right)\right], \quad (4)$$

$$\mathcal{L}_G = \sum_{t\geq 1} \mathbb{E}_{q(\mathbf{x}_t)} \mathbb{E}_{p_\theta\left(\mathbf{x}_{t-1}|\mathbf{x}_t\right)}\left[\left(D_\phi\left(\mathbf{x}_{t-1}, \mathbf{x}_t, t\right)-1\right)^2\right], \quad (5)$$

$$\mathcal{L}_D = \sum_{t\geq 1} \mathbb{E}_{q(\mathbf{x}_t) q\left(\mathbf{x}_{t-1}|\mathbf{x}_t\right)}\left[\left(D_\phi\left(\mathbf{x}_{t-1}, \mathbf{x}_t, t\right)-1\right)^2\right] + \mathbb{E}_{p_\theta\left(\mathbf{x}_{t-1}|\mathbf{x}_t\right)}\left[D_\phi\left(\mathbf{x}_{t-1}, \mathbf{x}_t, t\right)^2\right], \quad (6)$$

[2] We obtain similar results among different frequency bands and choose the 70-th bin for illustration.



(a) Shared architecture

(b) Training process

Figure 2: The overall architecture for dual-empowered speech models. In subfigure (a), it takes noisy audio $\mathbf{x}_t$ as input and conditions on diffusion time index t and Mel-spectrogram c.

Where D_{adv} depends on the adversarial training setup, and the fake samples from $p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$ are contrasted against the real one from $q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)$.

Reparameterization on diffusion model. Different from the conventional diffusion models that require hundreds of steps with small β_t to estimate the gradient for data density, recent works (Salimans and Ho, 2022; Liu et al., 2022) have witnessed that approximating some surrogate variables, e.g., the noiseless target data gives better quality. We reparameterize the denoising model by directly predicting the clean data $\mathbf{x}_0$. Free from estimating the gradient for data density, it only needs to predict unperturbed $\mathbf{x}_0$ and then add perturbation with the posterior distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ (formulated in Appendix B), and the reverse transition distribution can be expressed as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, c) = q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \tilde{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t|t, c)\right) \quad (7)$$

4.3 GAN Leveraging Diffusion Mechanism

GAN-based models are often difficult to train, collapsing (Mao et al., 2019) without carefully selected hyperparameters and regularizers, and showing less sample diversity. Besides, these models show distinct degradation in training stability, which cannot generate deterministic values due to the complex data distribution.

In this work, we propose **FastDiff 2 (GANDiff)** leveraging diffusion mechanism to construct the forward process by multiple denoising iterations, and thus we expect it exhibits better training stability and sample diversity compared to traditional one-shot GANs. To be more specific, we 1) initialize the generator G with a pre-trained diffusion teacher; 2) conduct 4-iter denoising to generate $\tilde{\mathbf{x}}_0$ with gradient, which is regarded as the forward process of the generator; and finally 3) G plays an adversarial game with the discriminator D, and the min-max objective can be expressed as:

$$\mathcal{L}_G = \mathbb{E}_{q_{data}}\left[\left(D_\phi\left(\tilde{\mathbf{x}_0}\right) - 1\right)^2\right] \quad (8)$$

$$\mathcal{L}_D = \mathbb{E}_{q_{data}}\left[\left(D_\phi\left(\tilde{\mathbf{x}_0}\right)\right)^2 + \left(D_\phi\left(\mathbf{x}_0 - 1\right)\right)^2\right] \quad (9)$$

We empirically find that the initialization of diffusion teacher provides a better understanding of noise schedules, and it reduces the difficulties of adversarial learning by orders of magnitude. FastDiff 2 (GANDiff) breaks the forward process of one-shot conditional GAN into several denoising diffusion iterations, in which each step is relatively simple to model. Thus, it exhibits better sample diversity than traditional one-shot GANs.

4.4 Architecture

As illustrated in Figure 2(a), we take a stack of time-aware location-variable convolution (Huang et al., 2022b) as a shared backbone to model long-term time dependencies with adaptive conditions efficiently. Convolution is conditioned on dynamic variations (diffusion steps and spectrogram fluctuations) in speech, which equips the model with diverse receptive field patterns and promotes robustness.

We build the basic architecture of discriminator upon WaveNet (Oord et al., 2016). It consists of ten layers of non-causal dilated 1-D convolutions with weight normalization. The discriminator is trained to correctly classify the generated sample as fake while classifying the ground truth as real. More details have been attached in Appendix C.

4.5 Loss Objective

Adversarial GAN Objective. For the generator and discriminator, the training objectives follow (Mao et al., 2017), which replaces the binary cross-entropy terms of the original GAN objectives (Goodfellow et al., 2014) with least squares loss functions for non-vanishing gradient flows.

Frequency-domain Reconstruction Objective. To stabilize adversarial learning, we include frequency-domain sample reconstruction loss objective by applying the multi-resolution STFT (Short Time Fourier Transform) operation $STFT(\cdot)$ (given in Appendix F):

$$\mathcal{L}_\theta = \mathcal{L}_{STFT}(\tilde{\boldsymbol{x}}_0, \boldsymbol{x}_0) \tag{10}$$

4.6 Training Algorithm

The training procedures of the proposed FastDiff 2 (GANDiff) and FastDiff 2 (DiffGAN) have been illustrated as follows. The sampling algorithms have been attached in Appendix D.2.

Algorithm 1 Training FastDiff 2 (DiffGAN)

```
Require: FastDiff 2 (DiffGAN) generator $\theta$, discriminator $\phi$, and mel condition $c$.
repeat
    Sample $\mathbf{x}_0 \sim q_{\text{data}}$, $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and $t \sim \text{Unif}(\{1, \cdots, T\})$
    Sample $\mathbf{x}_t$, $\mathbf{x}_{t-1}$ according to E.q (2)
    $\tilde{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t | t, c)$
    Sample $\tilde{\mathbf{x}}_{t-1} \sim q(\mathbf{x}_{t-1} | \mathbf{x}_t, \tilde{\mathbf{x}}_0)$ according to E.q (7)
    Take gradient descent steps on $\nabla_\theta(\mathcal{L}_\theta + \mathcal{L}_G)$ according to E.q (10) and (5)
    Take gradient descent steps on $\nabla_\phi \mathcal{L}_D$ according to E.q (6)
until FastDiff 2 (DiffGAN) converged
```

Algorithm 2 Training FastDiff 2 (GANDiff)

```
Require: Diffusion teacher $\alpha$ with schedule $\beta$ ($T = 4$) derived by noise predictor, FastDiff 2 (GANDiff) generator $\theta$, discriminator $\phi$, and mel condition $c$.
Initialize $\theta$ parameters using teacher $\alpha$
repeat
    for $t = T, \cdots, 1$ do
        Sample $\tilde{\mathbf{x}}_{t-1} \sim p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t, c)$
    end for
    Take gradient descent steps on $\nabla_\theta(\mathcal{L}_\theta + \mathcal{L}_G)$ according to E.q (10) and (8)
    Take gradient descent steps on $\nabla_\phi \mathcal{L}_D$ according to E.q (9)
until FastDiff 2 (GANDiff) converged
```

5 Related Works

5.1 Diffusion Probabilistic Model

The diffusion probabilistic model is a family of generative models with the capacity to learn complex data distribution, which has recently attracted a lot of research attention in several important domains. Diffusion models generate high-fidelity samples yet inherently suffer from slow sampling speed, and thus multiple methods have conducted extensive investigations to accelerate the sampling process: Chen et al. (2020) utilize a grid search algorithm for a shorter inference schedule. Liu et al. (2021) introduces a shallow diffusion mechanism that starts denoising at a particular distribution instead of Gaussian white noise. Huang et al. (2022b); Lam et al. (2022) utilize a noise predictor to learn a tight inference schedule for skipping denoising steps. Their designs make diffusion models more applicable to real-world deployment, while the diffusion/denoising mismatch leads to quality degradation during jumping sampling steps. In this work, we avoid this mismatch by incorporating GANs into diffusion models, which makes it much more stable to implement the reverse process with large step sizes.

5.2 Generative Adversarial Network

Generative adversarial networks (GANs) (Jang et al., 2021; Kong et al., 2020a) are one of the most dominant deep generative models for speech generation. UnivNet (Jang et al., 2021) has demonstrated its success in capturing different waveform intervals with local-variable convolution. HIFI-GAN (Kong et al., 2020a) proposes multi-receptive field fusion (MRF) to model the periodic patterns matters. However, GAN-based models are often difficult to train, collapsing (Creswell et al., 2018) without carefully selected hyperparameters and regularizers, and showing less sample diversity. Differently, we incorporate diffusion models into GANs and break the generation process into several conditional denoising steps, in which each step is relatively simple to model. Thus, we expect our model to exhibit better sample diversity.

6 Experiments

6.1 Experimental Setup

6.1.1 Dataset

For a fair and reproducible comparison against other competing methods, we use the benchmark

Model	Quality			Speed	Diversity	
	MOS (↑)	STOI (↑)	PESQ (↑)	RTF (↓)	NDB (↓)	JS (↓)
GT	4.32±0.06	/	/	/		
WaveNet (MOL)	3.95±0.08	/	/	85.23	0.34	**0.002**
WaveGlow	3.86±0.08	0.961	3.20	0.029	0.73	0.015
HIFI-GAN	4.06±0.10	0.970	3.63	0.002	0.70	0.012
UnivNet	4.05±0.09	0.969	3.54	0.002	0.71	0.010
Diffwave (6 steps)	4.06±0.09	0.966	3.72	0.093	0.81	0.012
WaveGrad (50 steps)	4.00±0.00	0.954	3.33	0.390	0.68	0.012
FastDiff (4 steps)	4.09±0.10	0.971	3.78	0.017	0.66	0.014
FastDiff 2 (DiffGAN) (4 steps)	**4.16±0.10**	0.972	3.73	0.017	0.47	0.004
FastDiff 2 (GANDiff) (4 steps)	4.12±0.08	**0.979**	**3.90**	0.017	**0.27**	**0.002**

Table 2: Comparison with other neural vocoders in terms of quality, diversity and synthesis speed. For sampling, we used 50 steps in WaveGrad, 6 steps in DiffWave and 4 steps in FastDiff, respectively, following (ivanvovk, 2020), (philsyn, 2021), and (Huang, 2022).

LJSpeech dataset (Ito and Johnson, 2017) which consists of 13,100 audio clips of 22050 Hz from a female speaker for about 24 hours. To evaluate the model generalization ability over unseen speakers in multi-speaker scenarios, we prepare the VCTK dataset (Yamagishi et al., 2019), which is downsampled to 22050 Hz to match the sampling rate with the LJSpeech dataset. VCTK consists of approximately 44,200 audio clips uttered by 109 native English speakers with various accents. Following the common practice, we conduct preprocessing and extract the spectrogram with the FFT size of 1024, hop size of 256, and window size of 1024 samples.

6.1.2 Model Configurations

FastDiff 2 (DiffGAN) and FastDiff 2 (GANDiff) share the same backbone comprising three Diffusion-UBlocks and DBlocks with the up/downsample rate of $[8, 8, 4]$, respectively. The discriminator consists of ten layers of non-causal dilated 1-D convolutions, whose strides are linearly increasing from one to eight except for the first and last layers. Channels and kernel sizes are set to 64 and 5, respectively. Both variants share the same number of denoising steps ($T = 4$) in both training and inference. The multi-resolution STFT loss is computed by the sum of three different STFT losses described in Appendix F.

6.1.3 Training and Evaluation

Both models are trained with constant learning rate $lr = 2 \times 10^{-4}$ on 4 NVIDIA V100 GPUs. We use random short audio clips of 25600 samples from each utterance with a batch size of 16 for each GPU. We crowd-source 5-scale MOS tests via Amazon Mechanical Turk to evaluate the audio quality. The MOS scores are recorded with 95% confidence intervals (CI). Raters listen to the test samples randomly and are allowed to evaluate each audio sample once. We adopt additional objective evaluation metrics including STOI (Taal et al., 2010), PESQ (Rix et al., 2001) to test sample quality, and NDB, JS (Richardson and Weiss, 2018) for sample diversity. To evaluate the inference speed, we implement the real-time factor (RTF) assessment on a single NVIDIA V100 GPU. More information about objective and subjective evaluation is attached in Appendix E.

6.2 Comparsion With Other Models

We compared our proposed models in audio quality and sampling speed with competing models, including 1) WaveNet (Oord et al., 2016), the autoregressive generative model for raw audio. 2) WaveGlow (Prenger et al., 2019), the parallel flow-based model. 3) HIFI-GAN V1 (Kong et al., 2020a) and UnivNet (Jang et al., 2021), the most popular GAN-based models. 4) Diffwave (Kong et al., 2020b), WaveGrad (Chen et al., 2020), and FastDiff (Huang et al., 2022b), three diffusion probabilistic models that generate high-fidelity speech samples. For easy comparison, the results are compiled and presented in Table 2, and we have the following observations:

For our GAN-empowered diffusion model, FastDiff 2 (DiffGAN) has achieved the highest MOS compared with the baseline models, with a gap of 0.16 compared to the ground truth audio. Regarding inference speed, it enjoys an effective 4-iter sampling process and enables a speed of 58x faster than real-time on a single NVIDIA V100 GPU without engineered kernels. FastDiff 2 (DiffGAN) provides a principled way to accelerate DDPMs in both training and inference, avoiding quality degradation caused by a training-inference mismatch in baseline diffusion models (FastDiff, WaveGrad, Diffwave). It is worth mentioning that FastDiff 2 (DiffGAN) maintains the outperformed sample diversity inherited in DDPMs.

For diffusion-empowered GANs, FastDiff 2 (GANDiff) also demonstrates high-quality speech synthesis with the MOS of 4.12. For objective evaluation, it further presents the new state-of-the-art results in PESQ and STOI, superior to all baseline models. Moreover, we can see that it achieves a higher JSD and NDB compared to baseline GAN models. It breaks the generation process into several conditional denoising diffusion steps, in which each step is relatively simple to model. Thus, we expect our model to exhibit better mode coverage and sample diversity than traditional GANs (HIFI-GAN, UnivNet).

To conclude, by incorporating GAN and diffusion models, the dual-empowered speech models converge advantages from both classes and achieve high-quality and diverse speech synthesis at a low computational cost.

6.3 Ablation Study

We conduct ablation studies to demonstrate the effectiveness of several designs, including the diffusion reparameterization and frequency-domain objective in dual-empowered speech models. The results of both subjective and objective evaluation have been presented in Table 3, and we have the following observations: 1) Replacing the diffusion reparameterization design and parameterizing the denoising model by predicting the Gaussian noise ϵ has witnessed a distinct degradation in perceptual quality. Specifically, FastDiff 2 (DiffGAN) directly predicts clean data to avoid significant degradation when reducing reverse iterations. 2) Removing the sample reconstruction loss objective results in blurry predictions with distinct artifact (Kumar et al., 2019) in both variants, demonstrating the effectiveness of the multi-resolution STFT regularization in stabilizing adversarial learning, which is helpful to improve the quality of generated waveforms with a MOS gain.

Model	MOS (↑)	STOI(↑)	PESQ (↑)
GT	4.32±0.06	/	/
FastDiff 2 (DiffGAN)	4.16±0.10	**0.972**	3.73
w/o DR	2.40±0.08	0.922	3.19
w/o RO	2.40±0.08	0.922	3.19
FastDiff 2 (GANDiff)	4.12±0.08	**0.979**	**3.90**
w/o RO	2.71±0.07	0.954	3.15

Table 3: Ablation study results. Comparison of the effect of each component on quality. DR: diffusion reparameterization, RO: reconstruction objective.

Model	MOS (↑)	STOI(↑)	PESQ (↑)
GT	4.30±0.06	/	/
WaveNet (MOL)	3.80±0.07	/	/
WaveGlow	3.65±0.07	0.870	3.10
HIFI-GAN	3.76±0.09	0.862	3.14
UnivNet	3.79±0.08	0.887	3.21
Diffwave (6)	3.80±0.09	0.873	3.22
WaveGrad (50)	3.73±0.07	0.856	3.15
FastDiff (4)	3.84±0.08	0.894	3.25
FastDiff 2 (DiffGAN) (4)	**3.96±0.07**	0.910	3.28
FastDiff 2 (GANDiff) (4)	3.92±0.08	**0.912**	**3.57**

Table 4: Comparison with other neural vocoders of synthesized utterances for unseen speakers.

6.4 Generalization To Unseen Speakers

We use 40 randomly selected utterances of 5 unseen speakers in the VCTK dataset that are not used in training for out-of-distribution testing. Table 4 shows the experimental results for the mel-spectrogram inversion of the samples from unseen speakers: We notice that both variants produce high-fidelity samples and outperform baseline models. They universally generate audio with strong robustness from entirely new speakers outside the training set.

7 Conclusion

In this work, through revisiting two popular classes (diffusion models and GANs) of deep generative models, we observed that 1) GANs tended to generate samples but did not cover the whole distribution, and 2) diffusion models exhibited outperformed sample quality and diversity while requiring iterative refinement at a high computational cost. To achieve high-quality, fast and diverse speech synthesis, we converged advantages by incorporating GANs and diffusion models, introducing dual-empowered modeling perspectives: 1) FastDiff 2 (DiffGAN), a diffusion model whose denoising process was parametrized by conditional GANs, and the non-Gaussian denoising distribution made it much more stable to implement the reverse process with large step sizes; and 2) FastDiff 2 (GANDiff): a generative adversarial network whose forward process was constructed by multiple denoising diffusion iterations, and it exhibited better mode coverage and sample diversity. Experimental results showed that both variants enjoyed an efficient 4-step sampling and demonstrated superior sample quality and diversity. We envisage that our work serve as a basis for future speech synthesis studies.

8 Limitations and Potential Risks

The adversarial learning still requests a proper selection of hyperparameters, otherwise the training procedure could be unstable. Besides, training speech diffusion probabilistic models typically require more computational resources, and degradation could be witnessed with decreased training data.

Our proposed model lowers the requirements for high-quality speech synthesis, which may cause unemployment for people with related occupations, such as broadcasters and radio hosts. In addition, there is the potential for harm from non-consensual voice cloning or the generation of fake media, and the voices of the speakers in the recordings might be overused than they expect.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000, National Natural Science Foundation of China under Grant No.62222211, Grant No.61836002 and Grant No.62072397.

References

Taejun Bak, Junmo Lee, Hanbin Bae, Jinhyeok Yang, Jae-Sung Bae, and Young-Sun Joo. 2022. Avocodo: Generative adversarial network for artifact-free vocoder. *arXiv preprint arXiv:2206.13404*.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. Wavegrad: Estimating gradients for waveform generation. In *Proc. of ICLR*.

Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. 2018. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*.

Chenye Cui, Yi Ren, Jinglin Liu, Feiyang Chen, Rongjie Huang, Ming Lei, and Zhou Zhao. 2021. Emovie: A mandarin emotion speech dataset with a simple emotional text-to-speech model. *arXiv preprint arXiv:2106.09317*.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *Proc. of NeurIPS*, volume 34.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

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

Huang. 2022. Fastdiff.

Rongjie Huang, Feiyang Chen, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. 2021. Multi-singer: Fast multi-singer singing voice vocoder with a large-scale corpus. In *Proc. of ACM MM*.

Rongjie Huang, Chenye Cui, Feiyang Chen, Yi Ren, Jinglin Liu, Zhou Zhao, Baoxing Huai, and Zhefeng Wang. 2022a. Singgan: Generative adversarial network for high-fidelity singing voice generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2525–2535.

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

Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. 2022b. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2023b. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*.

Rongjie Huang, Yi Ren, Jinglin Liu, Chenye Cui, and Zhou Zhao. Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech. In *Advances in Neural Information Processing Systems*.

Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui, and Yi Ren. 2022c. Prodiff: Progressive fast diffusion model for high-quality text-to-speech. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2595–2605.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/. Accessed: 2022-01-01.

ivanvovk. 2020. Wavegrad.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. In *Proc. of InterSpeech*.

Ziyue Jiang, Yi Ren, Ming Lei, and Zhou Zhao. 2021. Fedspeech: Federated text-to-speech with continual learning. *arXiv preprint arXiv:2110.07216*.

Ziyue Jiang, Zhe Su, Zhou Zhao, Qian Yang, Yi Ren, and Jinglin Liu. 2022. Dict-tts: Learning to pronounce with prior dictionary knowledge for text-to-speech. *Advances in Neural Information Processing Systems*, 35:11960–11974.

Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *International Conference on Machine Learning*, pages 2410–2419. PMLR.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020a. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. of NeurIPS*, 33:17022–17033.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020b. Diffwave: A versatile diffusion model for audio synthesis. In *Proc. of ICLR*.

Robert Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*.

Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre de Brébisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Max WY Lam, Jun Wang, Dan Su, and Dong Yu. 2022. Bddm: Bilateral denoising diffusion models for fast and high-quality speech synthesis. In *Proc. of ICLR*.

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*.

Linjun Li, Tao Jin, Wang Lin, Hao Jiang, Wenwen Pan, Jian Wang, Shuwen Xiao, Yan Xia, Weihao Jiang, and Zhou Zhao. 2023. Multi-granularity relational attention network for audio-visual question answering. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1.

Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, Peng Liu, and Zhou Zhao. 2021. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. *arXiv preprint arXiv:2105.02446*, 2.

Songxiang Liu, Dan Su, and Dong Yu. 2022. Diffgan-tts: High-fidelity and efficient text-to-speech with denoising diffusion gans. *arXiv preprint arXiv:2201.11972*.

Zheqi Lv, Zhengyu Chen, Shengyu Zhang, Kun Kuang, Wenqiao Zhang, Mengze Li, Beng Chin Ooi, and Fei Wu. 2023a. Ideal: Toward high-efficiency device-cloud collaborative and dynamic recommendation system. *arXiv preprint arXiv:2302.07335*.

Zheqi Lv, Feng Wang, Shengyu Zhang, Kun Kuang, Hongxia Yang, and Fei Wu. 2022. Personalizing intervened network for long-tailed sequential user behavior modeling. *arXiv preprint arXiv:2208.09130*.

Zheqi Lv, Wenqiao Zhang, Shengyu Zhang, Kun Kuang, Feng Wang, Yongwei Wang, Zhengyu Chen, Tao Shen, Hongxia Yang, Beng Chin Ooi, and Fei Wu. 2023b. Duet: A tuning-free device-cloud collaborative parameters generation framework for efficient device model generalization. In *Proceedings of the ACM Web Conference 2023*.

Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. 2019. Mode seeking generative adversarial networks for diverse image synthesis. In *Proc. of CVPR*.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502.

Max Morrison, Rithesh Kumar, Kundan Kumar, Prem Seetharaman, Aaron Courville, and Yoshua Bengio. 2021. Chunked autoregressive gan for conditional waveform synthesis. *arXiv preprint arXiv:2110.10139*.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

philsyn. 2021. Diffwave-vocoder.

Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *Proc. of ICASSP*.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. Fastspeech: fast, robust and controllable text to speech. In *Proc. of ICONIP*.

Eitan Richardson and Yair Weiss. 2018. On gans and gmms. In *Proc. of ICONIP*.

Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Proc. of ICASSP*.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. In *Proc. of ICLR*.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proc. of ICASSP*.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

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

Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *Proc. of ICASSP*.

Zhenhui Ye, Rongjie Huang, Yi Ren, Ziyue Jiang, Jinglin Liu, Jinzheng He, Xiang Yin, and Zhou Zhao. 2023a. Clapspeech: Learning prosody from text context with contrastive language-audio pre-training. *2305.10763*.

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, JinZheng He, and Zhou Zhao. 2023b. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*.

Zhenhui Ye, Zhou Zhao, Yi Ren, and Fei Wu. 2022. Syntaspeech: syntax-aware generative adversarial text-to-speech. *arXiv preprint arXiv:2204.11792*.

Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Jianghe Xu, Shouhong Ding, and Chao Wu. 2021. A practical data-free approach to one-shot federated learning with heterogeneity. *arXiv preprint arXiv:2112.12371*.

Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. 2022a. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125.

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. 2022b. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pages 26311–26329. PMLR.

Jie Zhang, Lei Zhang, Gang Li, and Chao Wu. 2022c. Adversarial examples for good: Adversarial examples guided imbalanced learning. *arXiv preprint arXiv:2201.12356*.

A Detailed Formulation of DDPM

We define the data distribution as $q(\mathbf{x}_0)$. The diffusion process is defined by a fixed Markov chain from data $\mathbf{x}_0$ to the latent variable $\mathbf{x}_T$:

$$q(\mathbf{x}_1, \cdots, \mathbf{x}_T | \mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t | \mathbf{x}_{t-1}), \tag{11}$$

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

The whole process gradually converts data $\mathbf{x}_0$ to whitened latents $\mathbf{x}_T$ according to the fixed noise schedule $\beta_1, \cdots, \beta_T$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}) \tag{12}$$

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

$$\mathcal{L}_\theta = \left\| \boldsymbol{\epsilon}_\theta \left(\alpha_t \mathbf{x}_0 + \sqrt{1-\alpha_t^2}\boldsymbol{\epsilon} \right) - \boldsymbol{\epsilon} \right\|_2^2 \tag{13}$$

Unlike the diffusion process, the reverse process is to recover samples from Gaussian noises. The reverse process is a Markov chain from x_T to x_0 parameterized by shared θ:

$$p_\theta(\mathbf{x}_0, \cdots, \mathbf{x}_{T-1} | \mathbf{x}_T) = \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t), \tag{14}$$

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

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_\theta(\mathbf{x}_t, t)^2 \boldsymbol{I}) \tag{15}$$

B Diffusion Posterior Distribution

Firstly we compute the corresponding constants respective to diffusion and reverse process:

$$\alpha_t = \prod_{i=1}^{t} \sqrt{1-\beta_i} \quad \sigma_t = \sqrt{1-\alpha_t^2} \tag{16}$$

The Gaussian posterior in the diffusion process is defined through the Markov chain, where each iteration adds Gaussian noise. Consider the forward diffusion process in Eq. 12, which we repeat here:

$$\begin{aligned} &q(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T | x_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}), \\ &q(\boldsymbol{x}_t | \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}) \end{aligned} \tag{17}$$

We emphasize the property observed by (Ho et al., 2020), the diffusion process can be computed in a closed form:

$$q(\boldsymbol{x}_t | \boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \alpha_t \boldsymbol{x}_0, \sigma_t \mathbf{I}) \tag{18}$$

C Model Hyperparameters

C.1 Architectures

As illustrated in Table 5, we list the hyper-parameters of dual-empowered speech models.

Table 5: Architecture hyperparameters of FastDiff 2 (DiffGAN)/FastDiff 2 (GANDiff).

Module	Parameter
DBlock Hidden Channels	32
DBlock Downsample Ratios	[4, 8, 8]
Diffusion UBlock Hidden Channels	32
Diffusion UBlock Upsample Ratios	[8, 8, 4]
Time-aware LVC layers Each Block	4
Time-aware LVC layers Kernel Size	256
Diffusion Kernel Predictor Hidden Channels	64
Diffusion Kernel Predictor Kernel Size	3
Diffusion Embedding Input Channels	128
Diffusion Embedding Output Channels	512
Use Weight Norm	True
Total Number of Parameters	15 M

C.2 Diffusion hyperparameters

We list the diffusion hyper-parameters in Table 6.

Table 6: Diffusion hyperparameters.

Diffusion Hyperparameter
FastDiff 2 (GANDiff):
$\beta = [3.6701e^{-7}, 1.7032e^{-5}, 7.908e^{-4}, 7.6146e^{-1}]$
FastDiff 2 (DiffGAN):
$\beta = \text{Linear}(1 \times 10^{-4}, 0.1, 4)$

D Training and Inference details

D.1 Preliminary Study

Both models are trained with constant learning rate $lr = 2 \times 10^{-4}$ on 4 NVIDIA V100 GPUs. We conduct preprocessing and extract the spectrogram with the FFT size of 1024, hop size of 256, and window size of 1024 samples.

For audio quality, we adopt objective evaluation metrics including MCD (Kubichek, 1993) and PESQ (Rix et al., 2001). We crowd-sourced 5-scale MOS tests via Amazon Mechanical Turk. Raters listened to the test samples randomly, where they were allowed to evaluate each audio sample once. To evaluate the sampling speed, we implement the real-time factor (RTF) assessment on a single NVIDIA V100 GPU. NDB and JSD metrics are employed to explore the diversity of generated mel-spectrograms.

Algorithm 3 Sampling with FastDiff 2 (DiffGAN)

```
Input: FastDiff 2 (DiffGAN) generator $\theta$, and mel condition $c$.
Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T, \cdots, 1$ do
    Sample $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = q(\mathbf{x}_{t-1}|\mathbf{x}_t, \tilde{\mathbf{x}}_0 = f_\theta(\mathbf{x}_t|t, c))$
end for
return $\mathbf{x}_0$
```

Algorithm 4 Sampling with FastDiff 2 (GANDiff)

```
Input: FastDiff 2 (GANDiff) generator $\theta$, and mel condition $c$.
Sample $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
for $t = T, \cdots, 1$ do
    Sample $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$
end for
return $\mathbf{x}_0$
```

D.2 Sampling Algorithm

E Evaluation Matrix

E.1 Objective Evaluation

Perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) and The short-time objective intelligibility (STOI) (Taal et al., 2010) assesses the denoising quality for speech enhancement.

Number of Statistically-Different Bins (NDB) and Jensen-Shannon divergence (JSD). They measure diversity by 1) clustering the training data into several clusters, and 2) measuring how well the generated samples fit into those clusters.

Mel-cepstral distortion (MCD) (Kubichek, 1993) measures the spectral distance between the synthesized and reference mel-spectrum features.

E.2 Subjective Evaluation

All our Mean Opinion Score (MOS) tests are crowd-sourced and conducted by native speakers. The scoring criteria have been included in Table 7 for completeness. The samples are presented and rated one at a time by the testers, each tester is asked to evaluate the subjective naturalness of a sentence on a 1-5 Likert scale. The screenshots of instructions for testers are shown in Figure 3. We paid \$8 to participants hourly and totally spent about \$600 on participant compensation.

F Multi-resolution STFT loss details

By applying the multi-resolution short time fourier transform, we respectively obtain the spectral convergence ($\mathcal{L}_{stft\_sc}$) and log STFT magnitude ($\mathcal{L}_{stft\_mag}$) of $\mathcal{L}_{STFT}$ in frequency domain:

$$\mathcal{L}_{stft\_sc} = \frac{\| \mathrm{STFT}(\mathbf{x}_0) - \mathrm{STFT}(\tilde{\mathbf{x}}_0)\|_F}{\| \mathrm{STFT}(\mathbf{x}_0)\|_F} \tag{19}$$

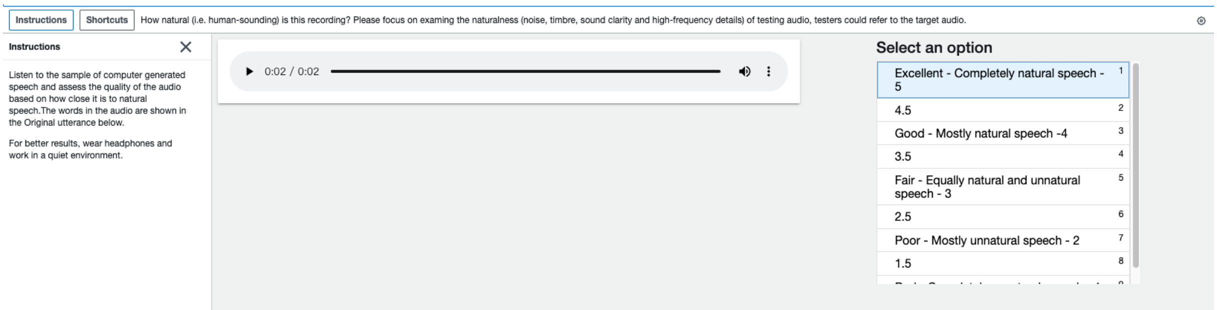
$$\mathcal{L}_{stft\_mag} = \frac{1}{N}\| \log(\mathrm{STFT}(\mathbf{x}_0)) - \log(\mathrm{STFT}(\tilde{\mathbf{x}}_0))\|_1, \tag{20}$$

where $\| \cdot \|_F$ and $\| \cdot \|_1$ denote the Frobenius and L1 norms. N denotes the number of elements in the magnitude; The final multi-resolution STFT loss is the sum of M losses with different analysis parameters(i.e., FFT size, window size, and hop size), and we set $M = 3$:

$$\mathcal{L}_{STFT} = \frac{1}{M}\sum_{m=1}^{M}\left(\mathcal{L}^{(m)}_{stft\_sc} + \mathcal{L}^{(m)}_{stft\_mag}\right) \tag{21}$$

Table 7: Ratings that have been used in the evaluation of speech naturalness of synthetic and ground truth samples.

Rating	Naturalness	Definition
1	Bad	Very annoying and objectionable dist.
2	Poor	Annoying but not objectionable dist.
3	Fair	Perceptible and slightly annoying dist
4	Good	Just perceptible but not annoying dist.
5	Excellent	Imperceptible distortions



How natural (i.e. human-sounding) is this recording? Please focus on examing the naturalness (noise, timbre, sound clarity and high-frequency details) of testing audio, testers could refer to the target audio.

Instructions

Listen to the sample of computer generated speech and assess the quality of the audio based on how close it is to natural speech.The words in the audio are shown in the Original utterance below.

For better results, wear headphones and work in a quiet environment.

0:02 / 0:02

Select an option

Excellent - Completely natural speech - 5 1
4.5 2
Good - Mostly natural speech -4 3
3.5 4
Fair - Equally natural and unnatural speech - 3 5
2.5 6
Poor - Mostly unnatural speech - 2 7
1.5 8

Figure 3: Screenshot of MOS testing.

Table 8: The details of the multi-resolution STFT loss. A hanning window was applied before the FFT process.

FFT size	Frame shift	Window size
1024	600	120
2048	120	250
512	240	50

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
See Section 8.

☑ A2. Did you discuss any potential risks of your work?
See Section 8.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☑ Did you run computational experiments?

See section 6.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
See section 6.1.2

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See section 6.1

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
See section 6.1.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
See section 6.1.1

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

See section 6.1.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
See section 6.1.3 and section E in Appendix.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
See section 6.1.3 and section E in Appendix.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
See section 6.1.3 and section E in Appendix.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.