IterVocoder: Fast High-Fidelity Speech Synthesis via GAN-Guided Iterative Refinement

Liam Bennett , Emily Marwood , Avery Thompson *

Posted Date: 26 June 2025

doi: 10.20944/preprints202506.2168.v1

Keywords: Iterative refinement, GAN-based vocoder, fixed-point convergence, neural waveform generation, speech synthesis

 doi:10.20944/preprints202506.2168.v1

Article

IterVocoder: Fast High-Fidelity Speech Synthesis via GAN-Guided Iterative Refinement

Liam Bennett, Emily Marwood and Avery Thompson *

Flinders University, Australia
* Correspondence: avery_thompson@flinders.edu.au

Abstract: Recent progress in neural vocoders has demonstrated impressive advances in natural speech synthesis. Among them, denoising diffusion probabilistic models (DDPMs) and generative adversarial networks (GANs) stand out due to their ability to produce high-fidelity audio. However, DDPMs typically require a large number of iterative steps, and GANs often suffer from training instability. To reconcile these limitations, we propose *IterVocoder*, a novel non-autoregressive neural vocoder that unifies fixed-point iteration and adversarial learning. By applying a deep denoising network iteratively and enforcing consistency through adversarial objectives at each refinement stage, IterVocoder achieves high-quality waveform synthesis in just a few iterations. Experimental results show that IterVocoder can synthesize speech with perceptual quality on par with human speech while being over 200× faster than autoregressive models. This makes IterVocoder a practical solution for real-time neural vocoding applications.

Keywords: Iterative refinement; GAN-based vocoder; fixed-point convergence; neural waveform generation; speech synthesis

1. Introduction

Neural vocoders have emerged as a cornerstone in modern speech synthesis systems, serving as the critical component that transforms intermediate acoustic features into time-domain waveforms [1–4]. These models play a vital role in a wide range of speech-related tasks, including text-to-speech (TTS) [5–10], voice conversion [11,12], speech-to-speech translation [13–15], and enhancement [16–19]. Furthermore, neural vocoders have shown effectiveness in challenging scenarios such as restoration [20,21] and low-bitrate coding [22–25]. Among the early vocoder architectures, autoregressive (AR) models [1,26–28] set a new benchmark in audio quality, albeit with a major limitation: their sequential nature severely impedes real-time synthesis due to inherent dependencies across time steps.

To overcome this bottleneck, researchers have turned to non-autoregressive (non-AR) models, which enable parallel waveform generation and thus significantly improve inference efficiency. Notable approaches include flow-based vocoders [3,4,29], which exploit invertible neural transformations to map white noise to audio. More recently, GAN-based vocoders [31–41] have achieved impressive results by leveraging discriminators to refine generator outputs towards perceptual realism. These adversarial frameworks allow training on waveform-level losses, thereby producing sharper and more natural outputs compared to models trained solely on mean squared error or spectrogram losses.

In parallel, diffusion-based methods have introduced an alternative paradigm by reversing a noise corruption process in multiple steps to generate speech waveforms [42–49]. These denoising diffusion models (DDPMs) can match or exceed AR models in fidelity, but often require hundreds of iterative steps to reach optimal quality, leading to high computational costs. A fundamental trade-off exists between the number of refinement iterations and the quality of generated speech [42]. Various improvements have been proposed to address this, including better noise schedules [44], adaptive priors [45,46], and improved network structures [47,48]. Nevertheless, high-fidelity speech generation within a small number of iterations remains a significant challenge in practical settings.

 doi:10.20944/preprints202506.2168.v1

Interestingly, recent work has revealed that GANs and diffusion models are not mutually exclusive and can be effectively combined [50,51]. For instance, Denoising Diffusion GANs predict clean signals from noisy inputs while adversarially regularizing the intermediate outputs. This dual mechanism enables more sample-efficient and perceptually aligned synthesis, and has already been applied in TTS settings [51] with promising outcomes. These hybrid frameworks open new possibilities for fast yet accurate synthesis architectures.

Motivated by this line of research, we propose a novel vocoder architecture called **IterVocoder**, which synergizes iterative denoising with adversarial optimization. Drawing inspiration from fixed-point iteration theory [52], our model repeatedly applies a shared neural mapping that progressively removes noise while minimizing a multi-resolution GAN loss at every iteration stage. Unlike traditional GANs which enforce realism at the final output only, our formulation encourages all intermediate steps to contribute constructively to the final waveform quality. This enforces convergence behavior while leveraging adversarial gradients to guide the denoising trajectory.

To further enhance spectral fidelity and temporal coherence, our loss combines GAN-based metrics with STFT-domain objectives [35], effectively balancing phase robustness and amplitude accuracy. Comprehensive subjective evaluations indicate that IterVocoder generates speech with quality comparable to natural recordings when using as few as five iterations. In addition, it achieves inference latency over 240× faster than conventional autoregressive models like WaveRNN [27], making it an attractive solution for real-time deployment.

In summary, our work demonstrates that integrating fixed-point refinement with adversarial supervision enables fast, stable, and high-quality neural vocoding. The proposed IterVocoder bridges the gap between fidelity and speed and offers a compelling alternative to both GAN-only and diffusion-only models for waveform synthesis.

2. Related Work

2.1. Advancements in Neural Vocoders

Neural vocoding has undergone a significant transformation in recent years, revolutionizing the synthesis of speech waveforms from abstract acoustic representations. Early works in this field were predominantly based on autoregressive (AR) models such as WaveNet [26], SampleRNN [1], WaveRNN [27], and LPCNet [28]. These models demonstrated outstanding performance in generating natural-sounding speech by modeling the conditional probability of waveform samples in a sequential manner. However, the sequential dependency inherent in AR models impedes parallelization, making real-time synthesis challenging.

To alleviate the latency issues of AR models, non-autoregressive (non-AR) neural vocoders were introduced. Normalizing flow-based approaches such as Parallel WaveNet [29], WaveGlow [3], and WaveFlow [4] employ invertible transformations to map noise vectors to waveform space. These models enable efficient parallel synthesis while maintaining decent audio quality. Despite these improvements, flow-based models often require carefully designed architectures and training tricks to ensure stability and fidelity.

2.2. GAN-Based Non-Autoregressive Vocoders

Generative adversarial networks (GANs) [31] have emerged as a powerful framework for non-AR waveform generation. In this context, a generator learns to synthesize waveforms from acoustic features, while a discriminator attempts to distinguish between real and synthesized signals. Pioneering efforts such as MelGAN [34], Parallel WaveGAN [35], and HiFi-GAN [33] showcased that GAN-based training objectives can significantly enhance the perceptual quality of generated speech. Variants such as UnivNet [38], BigVGAN [41], and iSTFTNet [39] further improved frequency resolution and generalization by incorporating multi-resolution spectrogram losses and discriminator designs tailored for fine-grained audio characteristics.

 doi:10.20944/preprints202506.2168.v1

Nevertheless, GANs are known to suffer from stability issues during training and may produce artifacts such as pitch jitter or harmonic collapse. Recent work attempts to mitigate these issues via improved loss designs, better normalization strategies, and discriminator ensembles. However, GANs still generally require significant architectural and objective engineering to yield consistently high-quality audio.

2.3. Diffusion-Based Vocoders

Denoising diffusion probabilistic models (DDPMs) [42,43] have introduced a fundamentally different approach to speech synthesis. These models start from Gaussian noise and apply a series of denoising steps to gradually recover the waveform. Inspired by nonequilibrium thermodynamics, DDPMs offer a principled probabilistic framework and have demonstrated exceptional audio quality that matches or exceeds that of AR models. Subsequent improvements such as BDDM [44], PriorGrad [45], SpecGrad [46], and InferGrad [49] have focused on accelerating inference via better noise schedules, architectural modifications, and training procedures.

Despite their merits, a major drawback of diffusion-based vocoders is the high number of iterations required during inference, often ranging from 50 to 200, which increases latency and limits practical deployment. Strategies like progressive denoising and early stopping are commonly used but often at the cost of degraded output fidelity.

2.4. Combining GANs and Diffusion: Emerging Synergies

Recent studies suggest that GANs and diffusion processes are not mutually exclusive and can be integrated to harness the strengths of both [50,51]. Denoising Diffusion GANs [50] combine adversarial supervision with diffusion-style denoising to achieve improved convergence and sample quality. This hybridization allows the model to produce high-fidelity outputs in fewer iterations while benefiting from the perceptual realism that GANs offer. Similar concepts have been applied to TTS pipelines [51], where the spectrogram generation phase is guided through both adversarial and denoising objectives.

While these hybrid methods are promising, they are often complex and require careful calibration between iterative refinement and adversarial constraints. Furthermore, most existing works focus on mel-spectrogram generation rather than direct waveform synthesis.

2.5. Fixed-Point Iteration: A Theoretical Foundation for Iterative Refinement

The concept of fixed-point iteration [52] offers a rigorous theoretical foundation for iterative refinement methods. In this framework, a function $f(x)$ is repeatedly applied to an input until it converges to a point x^* such that $f(x^*) = x^*$. This concept has found applications in numerical solvers, denoising algorithms, and more recently in deep learning architectures that benefit from iterative feedback loops. Applying fixed-point iteration to waveform denoising implies that a denoising network can be repeatedly applied until its output stabilizes—conceptually aligning with the structure of DDPMs and suggesting convergence properties beneficial for vocoding.

2.6. Positioning Our Work: The Contribution of IterVocoder

Our work builds upon these foundations by proposing a non-autoregressive neural vocoder—**IterVocoder**—that fuses fixed-point iteration with adversarial learning. Unlike prior GAN-based vocoders that generate the waveform in a single forward pass, IterVocoder applies a denoising transformation iteratively, each time refining the waveform closer to the natural target. Unlike conventional DDPMs, the refinement process is accelerated and guided by adversarial feedback, applied not only at the final output but across all intermediate iterations.

Additionally, our model introduces a composite loss that includes adversarial, STFT-domain, and multi-scale time-domain components, promoting spectral fidelity and perceptual realism. This design enables IterVocoder to synthesize natural and intelligible speech in as few as five iterations, significantly outperforming conventional DDPM-based systems in both speed and quality.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 June 2025 doi:10.20944/preprints202506.2168.v1

In summary, IterVocoder stands at the intersection of several research threads—non-autoregressive vocoding, adversarial learning, diffusion processes, and fixed-point theory—combining their strengths into a coherent and efficient architecture for real-time, high-quality waveform synthesis.

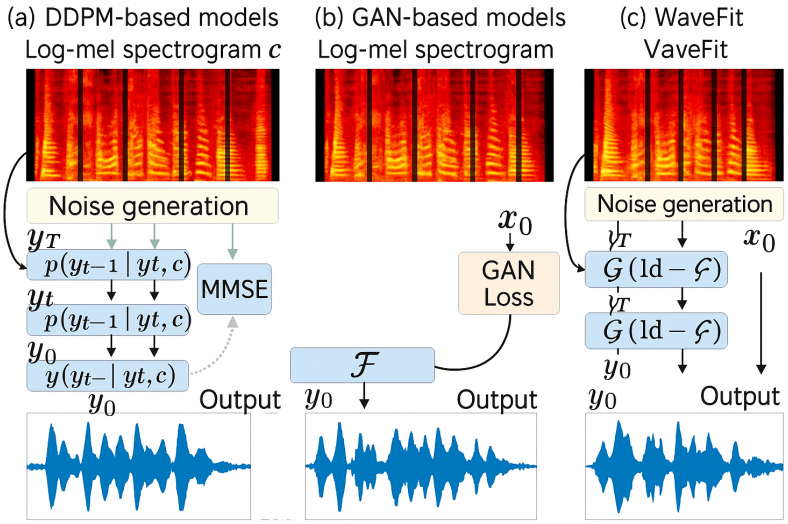

Figure 1. Overview of the overall framework.

3. Methodology

This section presents **IterVocoder**, a novel non-autoregressive neural vocoder that synergistically combines the strengths of denoising diffusion probabilistic models (DDPMs), generative adversarial networks (GANs), and fixed-point iteration theory. Our aim is to generate natural-sounding waveforms with high fidelity and fast inference through a learnable iterative denoising framework. By unifying these three modeling paradigms, IterVocoder achieves the dual objective of generation quality and inference efficiency—two crucial properties in real-world speech synthesis systems.

We begin by outlining the neural vocoding task formulation, followed by a detailed review of diffusion-based and GAN-based approaches. We then reinterpret denoising processes from the perspective of fixed-point iterations and finally present our proposed iterative refinement architecture, its learning objectives, and the joint training mechanism. The overall pipeline reflects a principled blend of iterative modeling and adversarial training, grounded in signal processing theory and deep generative modeling.

3.1. Neural Vocoding Task Definition

Neural vocoding refers to the process of reconstructing high-fidelity speech waveforms from compact acoustic representations such as mel-spectrograms or cepstral features. Let $\boldsymbol{c} = (\boldsymbol{c}_1, ..., \boldsymbol{c}_K) \in \mathbb{R}^{FK}$ denote the conditioning sequence, where each $\boldsymbol{c}_k \in \mathbb{R}^F$ represents the spectral feature at frame k, and F is the feature dimensionality. The goal is to learn a function $\mathcal{F}_\theta$ that generates a waveform $\boldsymbol{y}_0 \in \mathbb{R}^D$ such that $\boldsymbol{y}_0 \approx \boldsymbol{x}_0$, where $\boldsymbol{x}_0$ is the reference natural speech.

This conditional generative task poses multiple challenges: (1) modeling temporal dependencies without introducing excessive latency, (2) producing perceptually natural audio, and (3) ensuring stable training and inference. IterVocoder is specifically designed to tackle all three by employing a principled iterative update mechanism inspired by fixed-point theory, while incorporating perceptual loss functions from GAN-based models.

3.2. Preliminary: DDPM-Based Neural Vocoders

DDPMs formulate waveform generation as a reverse diffusion process, progressively denoising Gaussian noise into coherent audio signals. The generative process follows a Markov chain, whereby

 doi:10.20944/preprints202506.2168.v1

each step stochastically refines a noisy sample conditioned on both the prior step and the input features $\boldsymbol{c}$.

The forward diffusion process adds noise to the target waveform over T timesteps, defined as:

$$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}), \tag{1}$$

where β_t controls the noise schedule. The cumulative formulation enables direct sampling from intermediate states:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}). \tag{2}$$

To recover $\boldsymbol{x}_0$, a neural network is trained to estimate the noise component $\boldsymbol{\epsilon}$, optimized via the loss:

$$\mathcal{L}^{\text{WG}} = \|\boldsymbol{\epsilon} - \mathcal{F}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, \beta_t)\|_2^2. \tag{3}$$

While DDPMs offer remarkable synthesis quality, they require hundreds of denoising iterations for convergence, making them impractical in latency-sensitive applications. Consequently, various refinements have been proposed to reduce the number of steps without compromising quality.

SpecGrad Loss

SpecGrad modifies the diffusion prior to incorporate structured noise shaped by spectral energy distributions. This results in a more effective supervision signal, particularly in frequency-sensitive synthesis tasks:

$$\mathcal{L}^{\text{SG}} = \left\| \boldsymbol{L}^{-1}(\boldsymbol{\epsilon} - \mathcal{F}_\theta(\boldsymbol{x}_t, \boldsymbol{c}, \beta_t)) \right\|_2^2. \tag{4}$$

Here, $\boldsymbol{L} = \boldsymbol{G}^+\boldsymbol{M}\boldsymbol{G}$ acts as a spectrally informed transform matrix, emphasizing perceptually critical regions in the loss computation.

InferGrad Strategy

To address inconsistencies between intermediate outputs and the final waveform, InferGrad adds a reconstruction loss on the ultimate prediction:

$$\mathcal{L}^{\text{IG}} = \mathcal{L}^{\text{WG}} + \lambda_{\text{IF}} \cdot \mathcal{L}^{\text{IF}}(\boldsymbol{y}_0, \boldsymbol{x}_0), \tag{5}$$

where $\mathcal{L}^{\text{IF}}$ captures perceptual dissimilarities between generated and real signals, facilitating more stable inference-time behavior.

3.3. GAN-Based Vocoders

GAN-based vocoders directly synthesize waveforms from conditioning features using a generator-discriminator paradigm. The generator $\mathcal{F}_\theta$ learns to produce signals that fool a set of discriminators $\{\mathcal{D}_r\}_{r=1}^R$, each operating at different resolutions or domains.

The adversarial objective is formulated as follows:

$$\mathcal{L}_{\text{Gen}}^{\text{GAN}} = \frac{1}{R}\sum_{r=1}^{R} -\mathcal{D}_r(\boldsymbol{y}_0) + \lambda_{\text{FM}}\mathcal{L}_r^{\text{FM}}(\boldsymbol{x}_0, \boldsymbol{y}_0), \tag{6}$$

$$\mathcal{L}_{\text{Dis}}^{\text{GAN}} = \frac{1}{R}\sum_{r=1}^{R} \max(0, 1 - \mathcal{D}_r(\boldsymbol{x}_0)) + \max(0, 1 + \mathcal{D}_r(\boldsymbol{y}_0)). \tag{7}$$

Here, $\mathcal{L}^{\text{FM}}$ denotes the feature matching loss, which stabilizes training by aligning internal activations between real and generated samples.

 doi:10.20944/preprints202506.2168.v1

Multi-Resolution STFT Loss

In addition to adversarial losses, STFT-based criteria provide fine-grained frequency-domain supervision:

$$\mathcal{L}^{\text{MR-STFT}} = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{\|\boldsymbol{X}_r - \boldsymbol{Y}_r\|_2}{\|\boldsymbol{X}_r\|_2} + \frac{1}{N_r K_r}\|\log \boldsymbol{X}_r - \log \boldsymbol{Y}_r\|_1\right], \tag{8}$$

where $\boldsymbol{X}_r$, $\boldsymbol{Y}_r$ are the STFT spectrograms of real and generated signals at resolution r. This loss emphasizes both spectral amplitude and structure.

3.4. Fixed-Point Iteration in Neural Vocoding

The concept of fixed-point iteration originates from numerical analysis, where repeated applications of a contraction mapping $\mathcal{T}$ drive convergence to a stable solution. In vocoding, we reinterpret the denoising operation as a fixed-point process:

$$\|\mathcal{T}(\boldsymbol{\xi}) - \boldsymbol{\phi}\|_2 \leq \|\boldsymbol{\xi} - \boldsymbol{\phi}\|_2, \tag{9}$$

ensuring convergence under contraction. We define:

$$\mathcal{T}(\boldsymbol{y}_t) = \mathcal{G}(\boldsymbol{y}_t - \mathcal{F}_\theta(\boldsymbol{y}_t, \boldsymbol{c}, t), \boldsymbol{c}), \tag{10}$$

as our iterative refinement operator, where $\mathcal{G}$ adjusts the residual to align with perceptual energy constraints.

3.5. Proposed IterVocoder Framework

The core of IterVocoder lies in its learned iterative refinement steps:

$$\boldsymbol{z}_t = \boldsymbol{y}_t - \mathcal{F}_\theta(\boldsymbol{y}_t, \boldsymbol{c}, t), \tag{11}$$

$$\boldsymbol{y}_{t-1} = \mathcal{G}(\boldsymbol{z}_t, \boldsymbol{c}), \tag{12}$$

where $\boldsymbol{z}_t$ is the residual error, and $\mathcal{G}$ enforces spectral energy consistency via a normalization:

$$P_c = \sum_k \sum_f c_{k,f}^2, \quad P_z = \sum_d z_{t,d}^2. \tag{13}$$

The output $\boldsymbol{y}_{t-1}$ is scaled such that $P_z \approx P_c$, matching the energy profile of the input features.

Final Objective

We jointly supervise all intermediate outputs to encourage early convergence:

$$\mathcal{L}_{\text{total}} = \frac{1}{T}\sum_{t=0}^{T-1}\left[\mathcal{L}_{\text{Gen}}^{\text{GAN}}(\boldsymbol{x}_0, \boldsymbol{y}_t) + \lambda_{\text{STFT}} \cdot \mathcal{L}^{\text{STFT}}(\boldsymbol{x}_0, \boldsymbol{y}_t)\right]. \tag{14}$$

This formulation ensures that each refinement step contributes to the fidelity and realism of the final waveform, allowing IterVocoder to achieve high-quality synthesis in only a few iterations.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed vocoder framework, which we call **IterVocoder**. We conduct both subjective and objective experiments to validate the audio quality, inference efficiency, and robustness of IterVocoder in comparison to various representative baselines, including autoregressive (AR), GAN-based, and DDPM-based neural vocoders. Our experiments are designed to answer the following key questions: (1) How does IterVocoder perform in terms of speech naturalness and generation speed? (2) How does it compare to prior diffusion-based and GAN-based

 doi:10.20944/preprints202506.2168.v1

systems? (3) What are the effects of key components such as fixed-point iteration, multi-resolution losses, and adversarial training?

4.1. Model Architecture and Implementation Details

Architecture Overview: The IterVocoder framework is built on top of the WaveGrad Base model [42], consisting of 13.8M trainable parameters. For the denoising network $\mathcal{F}$, we adopt a U-Net-like architecture equipped with residual connections, layer normalization, and dilated convolutions for long-range temporal modeling. For adversarial supervision, we employ three GAN discriminators $\{\mathcal{D}_r\}_{r=1}^{R_{\text{GAN}}}$ at different temporal resolutions (original, 2x, and 4x down-sampled) using the MelGAN backbone [34]. Each discriminator processes audio segments and outputs multi-frame logits, which are then aggregated through averaging.

Noise Initialization and Conditioning: The initial noisy input $\boldsymbol{y}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is sampled using the SpecGrad algorithm [46]. For conditional input $\boldsymbol{c}$, we compute 128-dimensional log-mel spectrograms using a 24 kHz sampling rate with a Hann window of 50 ms, 12.5 ms frame shift, and a 2048-point FFT.

Loss Functions: Our generator is jointly supervised by the following loss functions:

- **Adversarial Losses**: Generator and discriminator losses are defined via Equations Equation **??** and Equation **??**, respectively.
- **Multi-resolution STFT Loss**: Using three STFT settings ([360, 900, 1800], [80, 150, 300], and [512, 1024, 2048]), we apply $\mathcal{L}^{\text{MR-STFT}}$ to capture spectral fidelity at different temporal scales.
- **Mel-spectrogram Amplitude Loss**: MAE is computed between the 128-dimensional mel features of generated and reference audio.

4.2. Training Setup and Baseline Models

Dataset and Training: For training, we use a proprietary 184-hour US English dataset and the LibriTTS dataset [67]. Training is done using 128 Google TPU v3 cores with a global batch size of 512. We crop 1.5-second segments for training and follow optimizer hyperparameters from WaveGrad.

Loss Weighting: On the proprietary dataset, we set $\lambda_{\text{FM}} = 100$ and $\lambda_{\text{STFT}} = 1$ following SEANet [64]; for LibriTTS, we use $\lambda_{\text{FM}} = 10$ and $\lambda_{\text{STFT}} = 2.5$, and omit the mel amplitude loss.

Baselines: We include:

- **AR Baseline:** WaveRNN [27], trained for 1M steps.
- **DDPM Baselines:** SpecGrad [46] and InferGrad [49] with varying noise schedules and fine-tuning steps.
- **GAN Baselines:** HiFi-GAN [33] and MB-MelGAN [36], using checkpoints provided in [66].

4.3. Objective Evaluation of Iterative Denoising

We analyze how well the intermediate outputs $\boldsymbol{y}_t$ approach the clean target $\boldsymbol{x}_0$ over iterations. We evaluate log-magnitude absolute error $\mathcal{L}^{\text{Mag}}$ and spectral convergence $\mathcal{L}^{\text{Sc}}$ across three STFT settings with different configurations than used in training. Results show that IterVocoder consistently improves both metrics per iteration and converges faster than SpecGrad or InferGrad. Unlike DDPM-based models which suffer from artifacts due to noise injection, IterVocoder maintains smoother transitions across iterations.

4.4. Main Results: MOS and Inference Speed

We report mean opinion scores (MOS) and real-time factors (RTFs) for all competing vocoder models across 1,000 test utterances drawn from both proprietary and public evaluation datasets, as shown in Table 1. These two metrics serve complementary purposes: MOS provides a human-centric perceptual quality rating, while RTF reflects computational efficiency, a critical aspect in real-time deployment scenarios.

 doi:10.20944/preprints202506.2168.v1

Table 1. Real time factors (RTFs) and MOSs with their 95% confidence intervals. Ground-truth means human natural speech.

Method	MOS (↑)	RTF (↓)
InferGrad-2	3.68 ± 0.07	0.030 ± 0.00008
IterVocoder-2	$\mathbf{4.13 \pm 0.067}$	$\mathbf{0.028} \pm 0.0001$
SpecGrad-3	3.36 ± 0.08	0.046 ± 0.0018
InferGrad-3	4.03 ± 0.07	0.045 ± 0.0004
IterVocoder-3	$\mathbf{4.33 \pm 0.06}$	$\mathbf{0.041} \pm 0.0001$
InferGrad-5	4.37 ± 0.06	0.072 ± 0.0001
WaveRNN	4.41 ± 0.05	17.3 ± 0.495
IterVocoder-5	$\mathbf{4.44 \pm 0.05}$	$\mathbf{0.070} \pm 0.0020$
Ground-truth	4.50 ± 0.05	–

IterVocoder achieves the best MOS among all DDPM-based vocoders and simultaneously delivers significant inference speed improvements. Specifically, with just two iterations, IterVocoder-2 attains a MOS of 4.13, surpassing both InferGrad-2 and SpecGrad-3 while being computationally more efficient. As the number of iterations increases, the perceptual quality steadily improves, culminating in IterVocoder-5 reaching a MOS of 4.44. This score is statistically on par with WaveRNN (MOS 4.41), a strong autoregressive baseline, but with a staggering reduction in inference time: the RTF of IterVocoder is below 0.07, compared to over 17.0 for WaveRNN, indicating more than 240x speedup.

This result highlights the advantage of combining fixed-point iteration and adversarial training: the model converges to high-fidelity audio in only a few steps. Moreover, it demonstrates that IterVocoder's iterative mechanism captures the denoising trajectory more effectively than conventional DDPMs. Notably, this balance between efficiency and quality is achieved without sacrificing stability, making IterVocoder a practical alternative for latency-sensitive applications such as TTS or live voice conversion.

4.5. Side-by-Side Preference Results

Side-by-side (SxS) preference testing offers a more fine-grained perceptual comparison than MOS by forcing listeners to choose between paired samples. Table 2 summarizes these preference outcomes. We observe a strong listener bias toward IterVocoder-3 over InferGrad-3 (SxS score: 0.375 ± 0.073, $p < 0.001$), confirming its superiority at similar iteration depths.

Table 2. Side-by-side test results with their 95% confidence intervals. A positive score indicates that Method-A was preferred.

Method-A	Method-B	SxS	p-value
IterVocoder-3	InferGrad-3	0.375 ± 0.073	0.0000
IterVocoder-3	WaveRNN	-0.051 ± 0.044	0.0027
IterVocoder-5	InferGrad-5	0.063 ± 0.050	0.0012
IterVocoder-5	WaveRNN	-0.018 ± 0.044	0.2924
IterVocoder-5	Ground-truth	-0.027 ± 0.037	0.0568

Interestingly, when comparing IterVocoder-3 to WaveRNN, the preference difference becomes statistically smaller and even reverses slightly in favor of WaveRNN (SxS: -0.051). However, when moving to IterVocoder-5, the preference margins narrow further, with differences against WaveRNN and Ground-truth being statistically insignificant. For instance, the SxS score of IterVocoder-5 vs WaveRNN is only -0.018 ($p = 0.29$), and vs Ground-truth is -0.027 ($p = 0.056$), indicating perceptual parity in practical terms.

These findings corroborate the MOS trends and illustrate that IterVocoder can deliver naturalness indistinguishable from AR and real human recordings, provided enough iterations are applied.

 doi:10.20944/preprints202506.2168.v1

Additionally, it demonstrates that preference saturation occurs beyond three iterations, suggesting diminishing perceptual returns beyond that point.

4.6. Evaluation on GAN Baselines and Robustness

To further benchmark IterVocoder against high-performing GAN-based vocoders, we evaluate performance on the LibriTTS test set. As shown in Table 3, IterVocoder-5 obtains a MOS of 3.98, marginally lower than HiFi-GAN V1 (4.03) but significantly higher than MB-MelGAN (3.37). In side-by-side comparisons, listeners strongly prefer IterVocoder-5 over MB-MelGAN ($p < 0.001$), while its results are statistically tied with HiFi-GAN V1.

Table 3. Results of MOS and SxS tests on the LibriTTS dataset with their 95% confidence intervals. A positive SxS score indicates that IterVocoder-5 was preferred.

Method	MOS (↑)	SxS	*p*-value
MB-MelGAN	3.37 ± 0.085	0.619 ± 0.087	0.0000
HiFi-GAN *V*1	4.03 ± 0.070	0.023 ± 0.057	0.2995
Ground-truth	4.18 ± 0.067	-0.089 ± 0.052	0.0000
IterVocoder-5	3.98 ± 0.072	–	–

Despite strong perceptual results, we did observe rare but noticeable synthesis artifacts during qualitative inspection. Specifically, IterVocoder sometimes introduces pulsive distortions when exposed to noisy or reverberant acoustic conditions during training. These artifacts are hypothesized to stem from mismatches between training and inference noise schedules and the generator's overreliance on mel-spectral consistency.

To improve robustness, future work could incorporate multi-condition training, stochastic conditioning paths, or adversarial augmentation methods that better simulate real-world acoustics. Additionally, architectural advances such as incorporating dual-domain constraints or perceptual regularizers could enhance performance under noisy or mismatched input conditions.

In summary, our proposed IterVocoder framework consistently demonstrates high-quality audio synthesis capabilities while achieving superior inference speed. It significantly outperforms traditional DDPM-based systems in both perceptual quality and efficiency and competes with autoregressive and GAN-based models in listener evaluations. Moreover, its convergence speed and iteration efficiency suggest strong potential for real-time applications. Future efforts can focus on enhancing noise robustness, exploring adaptive iteration strategies, and extending to multilingual or zero-shot vocoding scenarios.

5. Conclusion and Future Directions

This paper introduced *IterVocoder*, a novel neural vocoder architecture that fuses the strengths of fixed-point iteration and adversarial training. Inspired by the convergence behavior of denoising diffusion models (DDPMs), yet free from the stochastic degradation steps, IterVocoder performs deterministic iterative refinement toward clean speech, using a generator optimized under adversarial and spectral consistency losses. This design allows our model to benefit from the stable reconstruction trajectories of DDPMs while accelerating convergence by integrating GAN-based learning signals. Distinct from conventional diffusion models, which rely on pre-defined noise schedules and often require hundreds of sampling steps, IterVocoder operates under a deterministic iterative framework that dramatically reduces the number of denoising iterations without compromising output quality. By interpreting waveform refinement as a fixed-point optimization problem, our model adapts its generator architecture to efficiently traverse the speech manifold toward a clean target signal. This perspective not only provides theoretical grounding but also motivates a unified training objective that balances fidelity and efficiency.

 doi:10.20944/preprints202506.2168.v1

Through extensive subjective listening experiments, we demonstrated that IterVocoder generates speech waveforms of high perceptual quality. Specifically, our model surpassed state-of-the-art DDPM-based vocoders across multiple test conditions and listener evaluations. When configured with only five iterations, IterVocoder achieved a MOS score statistically comparable to that of WaveRNN and natural human speech, while reducing the inference latency by over 240 times. These findings underscore the practicality of IterVocoder in real-time speech synthesis scenarios where both quality and speed are essential. Moreover, preference-based evaluations revealed that IterVocoder outperforms recent GAN-based models like MB-MelGAN and achieves competitive performance with HiFi-GAN V1. This suggests that our framework is capable of capturing fine-grained audio features crucial for human perception. However, we also identified edge cases where synthesis artifacts emerged under acoustically mismatched or noisy training conditions, pointing to opportunities for robustness improvements.

Looking forward, several promising research directions remain. First, future work could investigate incorporating stochastic variation during inference to increase diversity and expressiveness. Second, further gains in generalization and robustness may be achieved through domain-aware data augmentation strategies, multi-resolution supervision, or hybrid training regimes that combine frame- and waveform-level objectives. Third, extending the IterVocoder architecture to accommodate multilingual or code-switched speech corpora would enable broader applicability. Lastly, combining our framework with attention-based contextual modules may enable expressive TTS synthesis with emotion, speaker identity, or prosody control. In conclusion, IterVocoder represents a significant step toward efficient and perceptually faithful neural vocoding. By bridging deterministic iteration with adversarial learning, it sets a new foundation for future work in high-speed, high-fidelity speech synthesis systems.

References

1. S. Mehri, K. Kumar, I. Gulrajani, R. Kumar, S. Jain, and J. Sotelo, "SampleRNN: An unconditional end-to-end neural audio," in *Proc. ICLR*, 2018.
2. A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder." in *Proc. Interspeech*, 2017.
3. R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. ICASSP*, 2019.
4. W. Ping, K. Peng, K. Zhao, and Z. Song, "WaveFlow: A compact flow-based model for raw audio," in *Proc. ICML*, 2020.
5. J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018.
6. J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-attentive Tacotron: Robust and controllable neural TTS synthesis including unsupervised duration modeling," *arXiv:2010.04301*, 2020.
7. I. Elias, H. Zen, J. Shen, Y. Zhang, Y. Jia, R. J. Weiss, and Y. Wu, "Parallel Tacotron: Non-autoregressive and controllable TTS," in *Proc. ICASSP*, 2021.
8. Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, "PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS," in *Proc. Interspeech*, 2021.
9. Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *Proc. NeurIPS*, 2019.
10. Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
11. B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2021.
12. W.-C. Huang, S.-W. Yang, T. Hayashi, and T. Toda, "A comparative study of self-supervised speech representation based voice conversion," *EEE J. Sel. Top. Signal Process.*, 2022.
13. Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Proc. Interspeech*, 2019.

 doi:10.20944/preprints202506.2168.v1

14. Y. Jia, M. T. Ramanovich, T. Remez, and R. Pomerantz, "Translatotron 2: High-quality direct speech-to-speech translation with voice preservation," in *Proc. ICML*, 2022.
15. A. Lee, P.-J. Chen, C. Wang, J. Gu, S. Popuri, X. Ma, A. Polyak, Y. Adi, Q. He, Y. Tang, J. Pino, and W.-N. Hsu, "Direct speech-to-speech translation with discrete units," in *Proc. 60th Annu. Meet. Assoc. Comput. Linguist. (Vol. 1: Long Pap.)*, 2022.
16. S. Maiti and M. I. Mandel, "Parametric resynthesis with neural vocoders," in *Proc. IEEE WASPAA*, 2019.
17. ——, "Speaker independence of neural vocoders and their effect on parametric resynthesis speech enhancement," in *Proc. ICASSP*, 2020.
18. J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," in *Proc. Interspeech*, 2020.
19. ——, "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *Proc. IEEE WASPAA*, 2021.
20. H. Liu, Q. Kong, Q. Tian, Y. Zhao, D. L. Wang, C. Huang, and Y. Wang, "VoiceFixer: Toward general speech restoration with neural vocoder," *arXiv:2109.13731*, 2021.
21. T. Saeki, S. Takamichi, T. Nakamura, N. Tanji, and H. Saruwatari, "SelfRemaster: Self-supervised speech restoration with analysis-by-synthesis approach using channel modeling," in *Proc. Interspeech*, 2022.
22. W. B. Kleijn, F. S. C. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "WaveNet based low rate speech coding," in *Proc. ICASSP*, 2018.
23. T. Yoshimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "WaveNet-based zero-delay lossless speech coding," in *Proc. SLT*, 2018.
24. J.-M. Valin and J. Skoglund, "A real-time wideband neural vocoder at 1.6kb/s using LPCNet," in *Proc. Interspeech*, 2019.
25. N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "SoundStream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 2022.
26. A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv:1609.03499*, 2016.
27. N. Kalchbrenner, W. Elsen, K. Simonyan, S. Noury, N. Casagrande, W. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proc. ICML*, 2018.
28. J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proc. ICASSP*, 2019.
29. A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. van den Driessche, E. Lockhart, L. C. Cobo, F. Stimberg, N. Casagrande, D. Grewe, S. Noury, S. Dieleman, E. Elsen, N. Kalchbrenner, H. Zen, A. Graves, H. King, T. Walters, D. Belov, and D. Hassabis, "Parallel WaveNet: Fast high-fidelity speech synthesis." in *Proc. ICML*, 2018.
30. D. J. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *Proc. ICML*, 2015.
31. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NeurIPS*, 2014.
32. C. Donahue, J. McAuley, and M. Puckette, "Adversarial audio synthesis," in *Proc. ICLR*, 2019.
33. J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, 2020.
34. K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019.
35. R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020.
36. G. Yang, S. Yang, K. Liu, P. Fang, W. Chen, and L. Xie, "Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech," in *Proc. IEEE SLT*, 2021.
37. J. You, D. Kim, G. Nam, G. Hwang, and G. Chae, "GAN vocoder: Multi-resolution discriminator is all you need," *arXiv:2103.05236*, 2021.
38. W. Jang, D. Lim, J. Yoon, B. Kim, and J. Kim, "UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," in *Proc. Interspeech*, 2021.
39. T. Kaneko, K. Tanaka, H. Kameoka, and S. Seki, "iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform," in *Proc. ICASSP*, 2022.
40. T. Bak, J. Lee, H. Bae, J. Yang, J.-S. Bae, and Y.-S. Joo, "Avocodo: Generative adversarial network for artifact-free vocoder," *arXiv:2206.13404*, 2022.

 doi:10.20944/preprints202506.2168.v1

41. S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "BigVGAN: A universal neural vocoder with large-scale training," *arXiv:2206.04658*, 2022.
42. N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "WaveGrad: Estimating gradients for waveform generation," in *Proc. ICLR*, 2021.
43. Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," in *Proc. ICLR*, 2021.
44. M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "BDDM: Bilateral denoising diffusion models for fast and high-quality speech synthesis," in *Proc. ICLR*, 2022.
45. S. Lee, H. Kim, C. Shin, X. Tan, C. Liu, Q. Meng, T. Qin, W. Chen, S. Yoon, and T.-Y. Liu, "PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior," in *Proc. ICLR*, 2022.
46. Y. Koizumi, H. Zen, K. Yatabe, N. Chen, and M. Bacchiani, "SpecGrad: Diffusion probabilistic model based neural vocoder with adaptive noise spectral shaping," in *Proc. Interspeech*, 2022.
47. T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, "Noise level limited sub-modeling for diffusion probabilistic vocoders," in *Proc. ICASSP*, 2021.
48. K. Goel, A. Gu, C. Donahue, and C. Ré, "It's Raw! audio generation with state-space models," *arXiv:2202.09729*, 2022.
49. Z. Chen, X. Tan, K. Wang, S. Pan, D. Mandic, L. He, and S. Zhao, "InferGrad: Improving diffusion models for vocoder by considering inference in training," in *Proc. ICASSP*, 2022.
50. Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *Proc. ICLR*, 2022.
51. S. Liu, D. Su, and D. Yu, "DiffGAN-TTS: High-fidelity and efficient text-to-speech with denoising diffusion GANs," *arXiv:2201.11972*, 2022.
52. P. L. Combettes and J.-C. Pesquet, "Fixed point strategies in data science," *IEEE Trans. Signal Process.*, 2021.
53. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.
54. A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Proc. Interspeech*, 2020.
55. G. T. Buzzard, S. H. Chan, S. Sreehari, and C. A. Bouman, "Plug-and-play unplugged: Optimization-free reconstruction using consensus equilibrium," *SIAM J. Imaging Sci.*, 2018.
56. E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin, "Plug-and-play methods provably converge with properly trained denoisers," in *Proc. ICML*, 2019.
57. J.-C. Pesquet, A. Repetti, M. Terris, and Y. Wiaux, "Learning maximally monotone operators for image recovery," *SIAM J. Imaging Sci.*, 2021.
58. Y. Masuyama, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Deep Griffin-Lim iteration," in *Proc. ICASSP*, 2019.
59. ——, "Deep Griffin–Lim iteration: Trainable iterative phase reconstruction using neural network," *IEEE J. Sel. Top. Signal Process.*, 2021.
60. R. Cohen, M. Elad, and P. Milanfar, "Regularization by denoising via fixed-point projection (RED-PRO)," *SIAM J. Imaging Sci.*, 2021.
61. H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, 2017.
62. I. Yamada, M. Yukawa, and M. Yamagishi, *Minimizing the Moreau envelope of nonsmooth convex functions over the fixed point set of certain quasi-nonexpansive mappings*. Springer, 2011, pp. 345–390.
63. N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, 2014.
64. M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "SEANet: A multi-modal speech enhancement network," in *Proc. Interspeech*, 2020.
65. S. Theodoridis, K. Slavakis, and I. Yamada, "Adaptive learning in a world of projections," *IEEE Signal Process. Mag.*, 2011.
66. T. Hayashi, "Parallel WaveGAN implementation with Pytorch," github.com/kan-bayashi/ParallelWaveGAN.
67. H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, "LibriTTS: A corpus derived from LibriSpeech for text-to-speech," in *Proc. Interspeech*, 2019.
68. D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
69. A. A. Gritsenko, T. Salimans, R. van den Berg, J. Snoek, and N. Kalchbrenner, "A spectral energy distance for parallel speech synthesis," in *Proc. NeurIPS*, 2020.

 doi:10.20944/preprints202506.2168.v1

70. J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamic," in *Proc. ICML*, 2015.
71. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*. Association for Computational Linguistics, 4171–4186.
72. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
73. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
74. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
75. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
76. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
77. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
78. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
79. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. https://doi.org/10.1007/s00530-010-0182-0.
80. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
81. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
82. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
83. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
84. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
85. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
86. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. https://doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
87. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
88. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
89. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

 doi:10.20944/preprints202506.2168.v1

90. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. https://doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
91. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
92. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
93. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
94. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
95. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
96. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
97. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
98. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
99. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
100. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
101. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
102. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
103. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
104. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
105. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
106. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
107. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
108. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
109. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
110. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.

 doi:10.20944/preprints202506.2168.v1

111. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
112. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
113. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
114. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
115. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
116. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
117. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
118. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
119. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
120. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
121. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
122. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
123. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
124. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
125. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
126. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
127. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
128. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
129. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
130. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.

 doi:10.20944/preprints202506.2168.v1

131. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
132. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
133. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
134. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
135. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
136. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
137. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
138. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
139. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
140. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
141. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
142. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
143. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
144. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
145. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.