ReGen: Hierarchical Multi-Prompt Representation Generation for Efficient Waveform Diffusion Models

Sang-Hoon Lee[1] and Ha-Yeong Choi[2]
[1]Department of Artificial Intelligence, Ajou University, Suwon, Korea [2]KT Corp., Seoul, Korea

ICML International Conference On Machine Learning

Introduction

<Waveform Generation using conditional flow matching>

Waveform Diffusion Models

1. **Conditional Flow Matching (CFM)**
   - CFM has recently gained increasing attention as a new generation paradigm for audio modeling
2. **Diffusion Transformers (DiTs) with Raw Waveform**
   - DiTs has strong generative capabilities, but Waveform modeling with DiTs remains relatively underexplored, partly due to the dominance of powerful GAN-based models
3. **Generation VS Reconstruction**
   - Low-bitrate neural audio codecs or High-compressed WaveVAEs can benefit from generative models, which may improve perceptual quality, speech intelligibility, and speaker similarity

Our Papers

-2023: Diffusion
- DiffWave [ICLR, 2021]
- WaveGrad [ICLR, 2021]
- PriorGrad [ICLR, 2022]
- MBD [NeurIPS, 2023]

2024-: CFM
- PeriodWave [ICLR, 2025]
- RFWave [ICLR, 2025]
- FlowDec [ICLR, 2025]

→ Streaming Generation
- StreamFlow [NeurIPS, 2025]

→ Acceleration (CFM+GAN)
- PeriodWave-Turbo [2024]
- Flow2GAN [ICLR, 2026]

→ CFM Neural Audio Codec with Representation Alignment (REPA)
- FlowTokenizer [ACL Findings, 2026]

↓
Representation Generation (ReGen)
- ReGen [ICML, 2026]

<A timeline of existing diffusion waveform generative models>

End-to-End TTS

Motivations

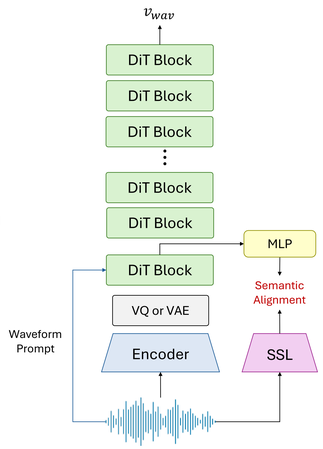
Efficient audio generation relies on compressed latent representations, but extreme compression can remove the semantic and acoustic information required for high-quality waveform generation

1. **Information bottleneck**
   - Low-bitrate neural codec or high-compressed VAE representations can degrade speech intelligibility, speaker identity, and high-frequency details.
2. **Representation Alignment is not enough**
   - Although REPA accelerates DiTs training, intermediate representation regularization may entangle latents and limit generative capacity.

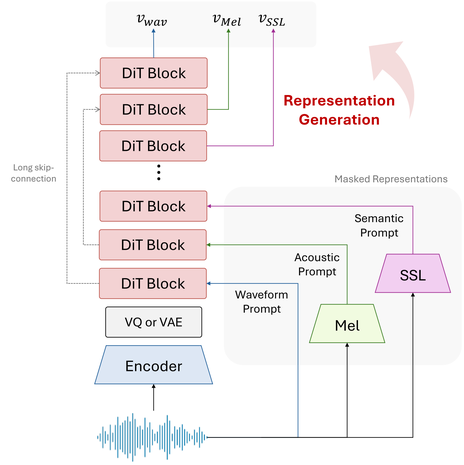
Contributions

1. **Representation Generation (ReGen)**
   - Reframes representation alignment as representation generation within a single diffusion Transformer
2. **Hierarchical Multi-prompting**
   - Uses semantic, acoustic, and waveform prompts to guide raw waveform generation
3. **Generalized Flow Matching (GFM)**
   - Adds a velocity-space repulsive loss to prevent multiple stochastic paths from collapsing into one averaged flow

Methods



(a) REPA (Baseline)



(b) ReGen (Proposed)

<Overall architecture of ReGen>

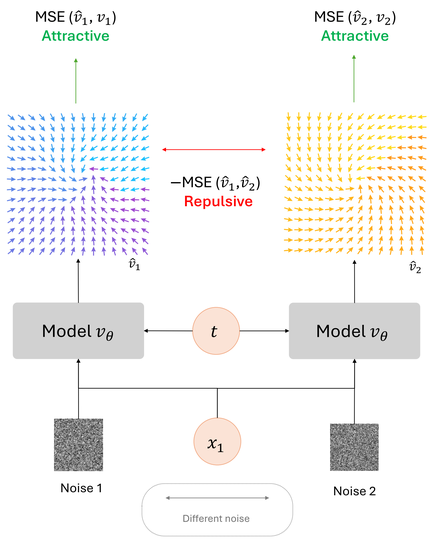
Experiment (Target Waveform: 24,000 Hz → 50Hz Patchify)

1. **ReGenTokenizer (FSQ): Low-bitrate neural audio codec (25Hz, 400 bps)**
2. **ReGenVAE (VAE): High-compressed WaveVAE (12.5 Hz, 32 dims)**
3. **ReGenVoice: LDM Text-to-Speech using the latents of ReGenVAE (12.5Hz → 6.25Hz LDM)**

Representation Generation (ReGen)

Generate representations, rather than merely aligning them: jointly predict vector fields in SSL, Mel, and waveform spaces under a hierarchical masked-prompt objective

$$\mathcal{L}_{\text{CFM}} = \sum_{r \in \{\text{ssl,mel,wav}\}} \mathbb{E}_{t,x_0,x_1}\left[\left\|\hat{v}_\theta^r(x_t^r, t, \bar{x}_1^r) - u_t^r\right\|^2 \odot M\right]$$



<Conceptual visualization of GFM>

Generalized Flow Matching (GFM)

Inspired by generalized energy distance (GED) which introduces a repulsive term between generated samples, we introduce generalized flow matching (GFM), which reformulates repulsive regularization in the velocity space, aligning the objective with the optimization domain of CFM.

$$\mathcal{L}_{\text{att}} = \frac{1}{2}\left(\|v_\theta(x_t^{(1)}, t, c) - u_t^{(1)}\|^2 + \|v_\theta(x_t^{(2)}, t, c) - u_t^{(2)}\|^2\right) \quad (4)$$

$$\mathcal{L}_{\text{rep}} = \|v_\theta(x_t^{(1)}, t, c) - \text{sg}(v_\theta(x_t^{(2)}, t, c))\|^2 \quad (5)$$

$$\mathcal{L}_{\text{GFM}} = \mathcal{L}_{\text{att}} - \lambda_{\text{neg}}\mathcal{L}_{\text{rep}}$$

$$\mathcal{L} = \mathcal{L}_{\text{GFM}}^{\text{wav}} + \lambda_{regen}\left(\mathcal{L}_{\text{GFM}}^{\text{Mel}} + \mathcal{L}_{\text{GFM}}^{\text{SSL}}\right)$$

Adversarial Post-Training

To improve audio fidelity, the pre-trained ReGen is further fine-tuned with adversarial feedbacks (MPD, MS-STFTD, MS-SB-CQTD) combined with multi-scale STFT losses.

Experiments

Reconstruction Results on LibriSpeech Benchmark

Method	Hz	TPS	N_q	Bitrate	Stream.	WER (↓)	STOI (↑)	PESQ -WB (↑)	PESQ -NB (↑)	SPK-SIM (↑)	UTMOS (↑)
GT	16k	-	-	-	-	1.96	1.00	4.64	4.55	1.00	4.09
SpeechTokenizer (Zhang et al., 2024)	16k	100	2	1000	✓	3.92	0.77	1.25	1.59	0.36	2.28
BigCodec (Xin et al., 2024)	16k	80	1	1040	✗	2.76	0.93	2.68	3.27	0.84	4.11
X-codec2 (Ye et al., 2025)	16k	50	1	800	✗	2.47	0.92	2.43	3.04	0.82	4.13
StableCodec(Parker et al., 2025)	16k	50	2	700	✗	5.12	0.91	2.24	2.91	0.62	4.23
EnCodec (Défossez et al., 2023)	24k	600	8	6000	✓	2.15	0.94	2.77	3.18	0.89	3.09
EnCodec (Défossez et al., 2023)	24k	150	2	1500	✓	4.90	0.85	1.56	1.94	0.60	1.58
WavTokenizer (Ji et al., 2025)	24k	75	1	900	✗	3.98	0.90	2.13	2.63	0.65	3.79
WavTokenizer(Ji et al., 2025)	24k	40	1	480	✗	11.20	0.85	1.62	2.06	0.48	3.57
Mimi (Défossez et al., 2024)	24k	100	8	1100	✓	2.92	0.90	2.27	2.80	0.73	3.63
ReGenTokenizer-LibriTTS	24k	25	1	400	✓	2.83	0.84	1.39	1.65	0.67	3.99
ReGenTokenizer-Emilia	24k	25	1	400	✓	2.70	0.86	1.52	1.85	0.74	3.77
VAE											
ReGenVAE-LibriTTS	24k	12.5	-	-	✓	2.05	0.94	2.90	3.38	0.84	4.21
ReGenVAE-Emilia	24k	12.5	-	-	✓	2.11	0.95	2.99	3.45	0.89	4.16

Prompted Reconstruction Results on Seed-en Benchmark

# Stage	Decoder	Method	Hz	TPS	N_q	Stream.	Prompt	WER (↓)	SPK-SIM (↑)	UTMOS (↑)
Single	GAN	EnCodec (Défossez et al., 2023)	24k	150	2	✓	✗	5.36	0.48	1.54
		DAC (RVQ) (Kumar et al., 2023)	24k	75	3	✗	✗	20.08	0.39	1.75
		DAC (VQ) (Kumar et al., 2023)	24k	75	1	✗	✗	12.74	0.45	2.08
		SpeechTokenizer (Zhang et al., 2024)	16k	100	2	✓	✗	7.98	0.46	2.47
		Mimi (Défossez et al., 2024)	24k	100	8	✓	✗	3.99	0.57	3.21
		X-codec2 (Ye et al., 2025)	16k	50	1	✗	✗	2.63	0.62	3.68
		WavTokenizer (Ji et al., 2025)	24k	75	1	✗	✗	6.65	0.48	3.36
		BigCodec (Xin et al., 2024)	16k	80	1	✗	✗	3.25	0.61	3.59
		StableCodec (Parker et al., 2025)	16k	25	1	✗	✗	11.08	0.41	3.87
Two	SPK, GAN	BiCodec (Wang et al., 2025b)	16k	50	1	✗	✓	3.05	0.61	3.68
Three	SPK, Mel-CFM, GAN-Voc.	FireRedTTS (Guo et al., 2024)	48k	25	1	✓	✓	3.35	0.59	3.40
Three	SPK, Mel-CFM, GAN-Voc.	CosyVoice (Du et al., 2024a)	24k	25	1	✓	✓	5.63	0.47	3.65
Three	SPK, Mel-CFM, GAN-Voc.	CosyVoice 2 (Du et al., 2024b)	24k	25	1	✓	✓	4.10	0.68	3.65
Three	LDM, MelVAE, GAN-Voc.	SemantiCodec (Liu et al., 2024)	16k	50	2	✗	✗	5.11	0.49	2.83
Two	Mel-CFM, GAN-Voc.	Vevo (Zhang et al., 2025)	24k	50	1	✗	✓	3.04	0.53	3.50
Two	Mel-CFM, GAN-Voc.	TaDiCodec (Wang et al., 2025c)	24k	12.5	1	✗	✓	2.57	0.69	3.58
Two	Mel-CFM, GAN-Voc.	TaDiCodec (Wang et al., 2025c)	24k	6.25	1	✗	✓	2.73	0.69	3.73
Single	Wave-CFM (+GAN)	**ReGenTokenizer-LibriTTS (ours)**	24k	25	1	✓	✓	3.37	0.61	3.73
Single	Wave-CFM (+GAN)	**ReGenTokenizer-Emilia (ours)**	24k	25	1	✓	✓	3.58	0.71	3.46
VAE										
Single	Wave-CFM (+GAN)	**ReGenVAE-LibriTTS (ours)**	24k	12.5	-	✓	✓	**2.03**	**0.72**	**3.87**
Single	Wave-CFM (+GAN)	**ReGenVAE-Emilia (ours)**	24k	12.5	-	✓	✓	**2.09**	**0.73**	**3.73**

Prompted Waveform Generation

Models	WER	SPK-SIM (to Prompt)	SPK-SIM (to GT)
ReGenVAE-Emilia (w/o Prompt)	2.04	0.65	0.86
ReGenVAE-Emilia (w Prompt)	2.09	0.73	0.87

Zero-shot TTS results on Seed-en Benchmark

Model	Params	Data	WER ↓	SIM ↑
GT	-	-	2.14	0.73
SparkTTS (Wang et al., 2025b)	0.5B	100k	1.98	0.58
TadiCodec-AR (Wang et al., 2025c)	4B	100k	2.28	0.65
F5-TTS (Chen et al., 2024)	0.3B	100k	2.00	0.67
ZipVoice (Zhu et al., 2025)	0.1B	100k	1.70	0.69
CosyVoice2 (Du et al., 2024b)	0.5B	166k	2.57	0.65
CosyVoice3-RL (Du et al., 2025)	0.5B	-	1.68	0.69
VibeVoice (Peng et al., 2025)	1.5B	-	3.04	0.68
VibeVoice-RT (Peng et al., 2025)	0.5B	-	2.05	0.63
VoxCPM-Emilia (Zhou et al., 2025)	0.6B	100k	2.34	0.68
ReGenVoice-S	0.3B	0.5k	2.04	0.65
ReGenVoice-S	0.3B	40k	2.21	0.68
ReGenVoice	0.5B	0.5k	**1.46**	0.64
ReGenVoice	0.5B	40k	1.62	**0.70**

Demo

Paper