StreamFlow: Streaming Audio Generation from Neural Codec Tokens via Streaming Flow Matching

Ha-Yeong Choi and Sang-Hoon Lee

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

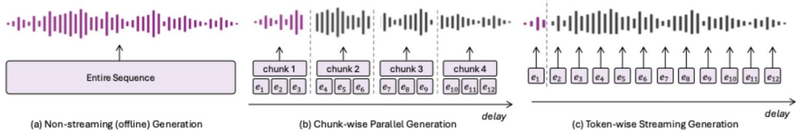


Fig 1. Comparison of generation strategies

Motivations

1. **Enabling fully Real-Time Duplex Communication**
   Recent speech language models (e.g., GPT-4o) require audio generators capable of true fully-duplex, low-latency streaming, motivating a generative framework that can operate at frame-level timescales.
2. **Extending the potential of Conditional Flow Matching**
   CFM-based decoders exhibit strong high-fidelity generation capabilities, yet existing designs are limited to non-streaming, full-sequence inference. There is a clear need to adapt these advantages to real-time scenarios.
3. **Accelerating Streaming Application**
   Real-time TTS, streaming voice conversion, interactive agents, and speech language models.

Challenges

1. **Real-time latency limits of existing CFM**
   Multi-step iterative sampling and chunk-wise generation prevent true frame-synchronous streaming, making conventional CFM unsuitable for low-latency real-time applications.
2. **Quality degradation when reconstructing high-resolution waveforms from low-bitrate tokens**
   Neural codecs compress speech into discrete low-bitrate tokens, but lack intermediate temporal refinement.

Contributions

1. **Streaming Flow Matching (SFM) for token-wise real-time generation**
   - Proposing a novel streaming generative model that leverages self-conditioned context to estimate multi-time vector fields, enabling token-wise streaming generation.
2. **Scale-DiT Architecture**
   - Regularizes representation by modeling and scaling residual–feature differences.
   - Improves stability and generalization without increasing parameter size, enhancing DiT-based audio generation.
3. **Streaming-optimized waveform generation**
   - Replaces STFT/iSTFT with a linear-reshape transformation suited for real-time systems.
   - Two-stage training (SFM pre-training + adversarial fine-tuning) achieves high-fidelity 24 kHz waveform reconstruction.

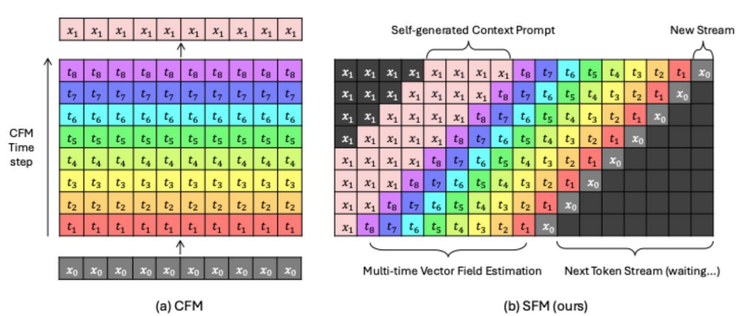


Fig 2. Comparison of (a) CFM and (b) proposed SFM

Methods

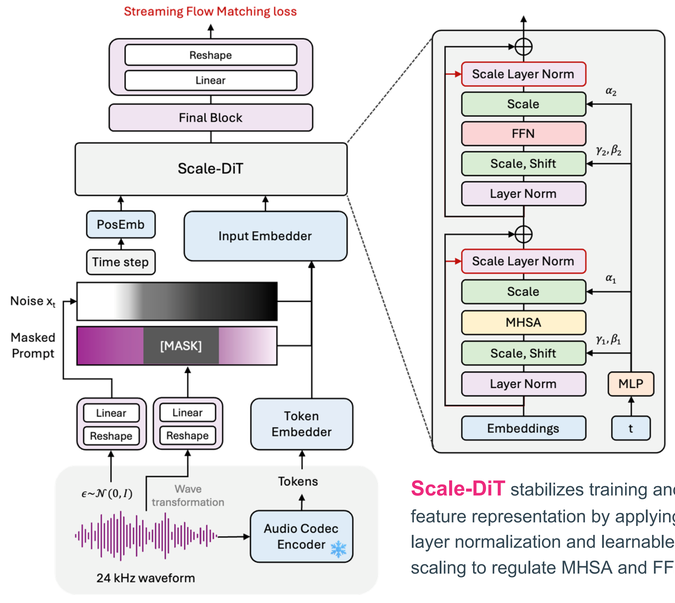
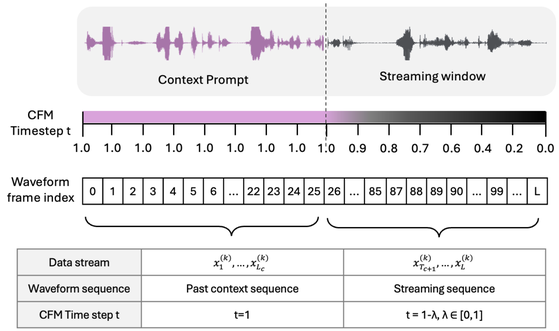


Fig 3. Overall architecture of StremFlow

Scale-DiT stabilizes training and improves feature representation by applying adaptive layer normalization and learnable residual scaling to regulate MHSA and FFN updates.



Data stream	$x_1^{(k)}, \ldots, x_{t_c}^{(k)}$	$x_{t_{c+1}}^{(k)}, \ldots, x_L^{(k)}$
Waveform sequence	Past context sequence	Streaming sequence
CFM Time step t	t=1	t = 1-λ, λ ∈ [0,1]

Fig 4. Causal Nosing

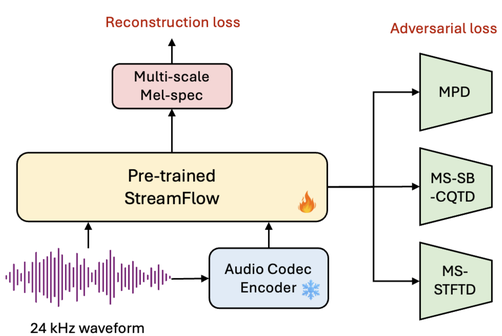


Fig 5. Fine-tuning with Adversarial Training

Streaming Flow Matching

Training: Apply causal noising by fixing past context and masking the streaming window to learn strictly causal representations.

Inference: Prepend newly generated frames and advance the window to enable continuous frame-aligned streaming via in-context learning.

Waveform Transformation

We replace STFT/iSTFT with a linear-reshape transformation that removes large receptive fields and future-frame dependency, enabling efficient causal streaming. The reshaped waveform is projected into and back from the Scale-DiT feature space without extra computation during sampling.

Adversarial Training

For high-fidelity audio, the pre-trained StreamFlow is further refined using adversarial discriminators (MPD, MS-STFTD, MS-SB-CQTD) combined with multi-scale STFT losses.

Experiments

Encodec Token Reconstruction Results

Model	Streaming	Params.	M-STFT ↓	PESQ ↑	Period ↓	V/UV ↑	UTMOS ↑	MOS ↑
GT	-	-	-	-	-	-	3.423	4.07±0.02
Vocos [46]	✗	7M	1.074	3.051	0.086	0.957	3.100	3.98±0.02
MBD [13]	✗	411M	1.612	2.645	0.108	0.946	3.300	3.95±0.03
RFWave [32]	✗	18M	1.280	3.020	0.078	0.957	2.988	3.99±0.02
StreamFlow	✗	170M	0.997	3.473	0.080	0.957	3.450	4.03±0.02
Encodec [9]	✓	15M	1.170	2.643	0.112	0.941	2.542	3.74±0.03
StreamFlow-Tiny	✓	11M	1.111	3.027	0.107	0.947	3.060	3.99±0.02
StreamFlow-Small	✓	44M	1.072	3.207	0.096	0.950	3.206	3.99±0.03
StreamFlow-Base	✓	175M	1.061	3.335	0.102	0.948	3.325	4.03±0.02

Mimi Token Reconstruction Results

Model	N_q	Bitrate	F (Hz)	CER ↓	WER ↓	M-STFT ↓	PESQ ↑	Period. ↓	V/UV ↑	Pitch ↓	UTMOS ↑
GT	-	-	-	1.12	3.06	-	-	-	-	-	3.862
Mimi	4	550	50	7.42	12.72	1.552	1.657	0.210	0.880	77.575	3.019
StreamFlow	4	550	50	5.22	9.45	1.410	1.584	0.212	0.876	73.891	3.093
Mimi	6	825	75	5.10	9.00	1.426	2.012	0.180	0.901	60.142	3.347
StreamFlow	6	825	75	3.78	6.87	1.272	2.043	0.177	0.906	55.830	3.719
Mimi	8	1100	100	3.05	6.93	1.352	2.266	0.165	0.910	50.686	3.506
StreamFlow	8	1100	100	3.26	6.16	1.217	2.306	0.162	0.915	45.640	3.910

Scalability with respect to model size

Model	Params.	Input Dim.	Hidden	Head	M-STFT ↓	PESQ ↑	Period ↓	V/UV ↑	Pitch ↓	UTMOS ↑
StreamFlow-Tiny	11M	256	1024	4	1.126	3.183	0.097	0.951	27.846	3.550
StreamFlow-Small	44M	512	2048	8	1.099	3.307	0.089	0.955	27.740	3.690
StreamFlow-Base	175M	1024	4096	16	1.088	3.430	0.088	0.955	26.843	3.792

Ablation Study

Model	M-STFT ↓	PESQ ↑	Period ↓	V/UV ↑	Pitch ↓	UTMOS ↑
Streaming (online)						
StreamFlow	1.088	3.430	0.088	0.955	26.843	3.792
w/o In-Context Learning	1.099	3.337	0.088	0.954	26.772	3.748
w/o Adversarial Fine-tuning	1.388	2.669	0.101	0.950	21.099	3.178
w/o SFM Pre-training	1.919	1.153	0.353	0.812	574.55	1.308
w/o Scale-DiT	1.593	2.557	0.099	0.946	27.330	3.033
w/o RoPE	2.049	1.598	0.176	0.898	68.891	1.904
Non-streaming (offline)						
StreamFlow	1.017	3.499	0.085	0.955	29.087	3.888
w/o iSTFT	1.097	3.139	0.082	0.956	22.636	3.619

Further Analysis of Scale-DiT

Model	Training Steps	WER ↓	STOI ↑	PESQ ↑	SPK-SIM ↑	UTMOS ↑
DiT	150k	11.76	0.85	1.71	0.58	2.72
DiT + REPA	150k	8.76	0.86	1.74	0.59	2.85
Scale-DiT	150k	9.68	0.86	1.78	0.59	2.83
Scale-DiT + REPA	150k	8.34	0.87	1.78	0.60	2.92
DiT	300k	9.41	0.87	1.84	0.62	3.04
DiT + REPA	300k	8.17	0.87	1.84	0.64	3.10
Scale-DiT	300k	7.56	0.87	1.90	0.64	3.16
Scale-DiT + REPA	300k	7.18	0.88	1.92	0.65	3.26
DiT	700k	9.59	0.87	1.92	0.64	3.20
DiT + REPA	700k	6.88	0.88	1.92	0.67	3.28
Scale-DiT	700k	6.23	0.88	2.07	0.68	3.45
Scale-DiT + REPA	700k	5.99	0.89	2.09	0.68	3.52

Replacing Mimi decoder with StreamFlow

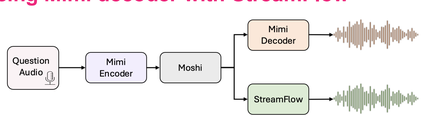


Fig 6. Inference pipeline comparing the Mimi decoder and the proposed StreamFlow

Method	CER ↓	WER ↓	UTMOS ↑
Moshi w/ Mimi decoder	7.59	9.89	3.610
Moshi w/ StreamFlow (Ours)	**7.19**	**9.82**	**3.847**

Demo

Paper