FreyaTTS Technical Report

Freya Team

July 7, 2026

Project: https://github.com/freyavoiceai
Model: https://huggingface.co/freyavoice/freya-tts
Samples: https://huggingface.co/freyavoice

Abstract

We present FreyaTTS, an open, tokenizer-free Turkish speech foundation model designed for reliable, efficient speech synthesis. FreyaTTS advances mid-resource TTS in three dimensions: (i) **efficiency**, packing competitive Turkish synthesis into a 183.2M-parameter model pretrained from scratch on roughly 1,200 hours of speech; (ii) **Turkish-first, tokenizer-free design**, driving generation from a 92-symbol character vocabulary with no phonemizer or discrete speech tokenizer, so that agglutinative morphology, vowel harmony, and spoken-form numbers are learned end-to-end from audio; and (iii) **non-autoregressive generation**, via a conditional flow-matching Diffusion Transformer that predicts entire utterances in parallel inside the frozen continuous-latent space of the VoxCPM2 AudioVAE (16 kHz encode, 48 kHz decode) (Zhou et al., 2025), with explicit cross-attention conditioning that aligns where E2/F5-style implicit alignment (Eskimez et al., 2024; Chen et al., 2025b) fails at this scale. On **Freya-TR-Eval**, a 495-sentence Turkish benchmark we release, FreyaTTS reaches 8.0 % WER and 3.0 % CER, and a 7-rater listening study places its naturalness above every same-class open system; on intelligibility it outperforms the larger open systems XTTS-v2 (~470M, 11.1 %) and F5-TTS (~336M, 24.3 %). Batched-ODE serving delivers 65.2 audio-seconds per second at batch 8 on one RTX 4090 – 2.6× the 2B VoxCPM2 serving engine at a third of the memory – and the model runs at or faster than real time on a laptop CPU. We release the model weights, training and inference code, and the benchmark under the Apache-2.0 license.

Contents

1 Introduction

Text-to-speech (TTS) has advanced from producing merely intelligible speech toward generating natural, expressive, and controllable audio (Shen et al., 2018; Ren et al., 2020), driven by the large-language-model paradigm of framing synthesis as sequence modeling over discrete audio tokens (Borsos et al., 2023; Chen et al., 2025a) and, more recently, by continuous-latent flow-matching generators (Shen et al., 2023; Le et al., 2023; Chen et al., 2025b). This progress, however, is concentrated in high-resource English and Chinese. Mid-resource languages such as Turkish, for which public corpora such as FLEURS-tr (Conneau et al., 2023) and Common-Voice-tr (Ardila et al., 2020) offer only tens of hours of validated read speech, remain underserved: large multilingual systems (Casanova et al., 2024; Du et al., 2025) treat Turkish as one language among dozens rather than as a first-class target. A second gap is computational. The strongest open systems reach their quality through multi-billion-parameter backbones and 50–95k-hour training runs, an operating point that excludes single-GPU serving and on-device deployment. Turkish additionally stresses the text frontend: agglutinative morphology, vowel harmony, and the spoken-form expansion of numbers, dates, and acronyms make hand-built grapheme-to-phoneme and normalization pipelines brittle and incomplete. These pressures argue for a compact, Turkish-first system that learns pronunciation end-to-end from audio rather than dictating it by rules.

Two modeling paradigms dominate contemporary TTS. Discrete-token systems represent speech as codec tokens and inherit LLM-style scaling and in-context learning (Borsos et al., 2023; Chen et al., 2025a; Du et al., 2025), but quantization discards fine acoustic detail and typically forces a multi-stage pipeline. Continuous-latent systems instead model speech representations directly with denoising or flow-matching objectives, spanning non-autoregressive diffusion models (Shen et al., 2023; Le et al., 2023; Chen et al., 2025b; Eskimez et al., 2024) and diffusion-autoregressive hybrids (Li et al., 2024; Jia et al., 2025). VoxCPM (Zhou et al., 2025) unified these lines through a differentiable finite-scalar-quantization bottleneck (Mentzer et al., 2024) inside a single continuous-latent backbone, and its successor VoxCPM2 scales the design to a 2B-parameter, 48 kHz, 30-language foundation model whose asymmetric AudioVAE encodes at 16 kHz and reconstructs at 48 kHz. Rather than replicate the multi-million-hour training budget such foundation models require, FreyaTTS takes a deliberately resource-frugal route: we treat the VoxCPM2 AudioVAE as a *frozen*, Apache-2.0 continuous-latent codec – a fixed 25 Hz, 64-dimensional latent space – and train a compact 183.2M-parameter model entirely inside it. Freezing the codec decouples our data budget from the codec's, so that a modest from-scratch Turkish pretraining run suffices to reach a strong Turkish-first model.

The second design decision concerns *how* the latent sequence is generated. VoxCPM and VoxCPM2 decode autoregressively, one acoustic patch at a time, atop a MiniCPM-4 backbone (Team et al., 2025). Autoregression is expressive, but it accumulates errors over long horizons and is prone to number and word garbling – a benign glitch in a story reader, a critical failure wherever a misread digit changes meaning. FreyaTTS therefore adopts a non-autoregressive formulation: a conditional flow-matching Diffusion Transformer predicts an entire utterance's latent sequence in parallel over a separately predicted duration, sidestepping left-to-right error accumulation by construction. Conditioning is explicit – cross-attention over ConvNeXt-refined character features from a 92-symbol Turkish character vocabulary, with no byte-pair encoding, phonemizer, or grapheme-to-phoneme step – so the model learns the pronunciation of in-context numbers, currencies, and acronyms directly from audio. A dedicated numeric stress test (section 4.3) sharpens this argument in an unexpected direction: what limits digit rendering in the NAR design is not decoding order but duration allocation, so digit strings are expanded to their spoken form at the text frontend – the standard input contract for character-level synthesis – while isolated bare tokens are handled by post-training and a thin inference layer (sections 3.5 and 4). A two-stage post-training recipe then specializes the from-scratch pretrained prior into a single-voice production model.

The main contributions of FreyaTTS are as follows.

1. **A tokenizer-free non-autoregressive Turkish TTS model.** FreyaTTS is a 183.2M-parameter conditional flow-matching Diffusion Transformer that synthesizes 48 kHz Turkish speech in the frozen continuous-latent space of the VoxCPM2 AudioVAE, from a 92-symbol character vocabulary with no phonemizer, grapheme-to-phoneme frontend, or discrete speech tokenizer – to our knowledge, the first openly released tokenizer-free NAR TTS model pretrained from scratch as a Turkish-first system (Pratap et al., 2024; Casanova et al.,

2024). We release both the pretrained prior and the single-voice production model.

2. **A cross-attention-versus-implicit-alignment finding at mid-resource scale.** The implicit-alignment recipe of E2 TTS and F5-TTS (Eskimez et al., 2024; Chen et al., 2025b) failed to align at our data budget, plateauing near 1.06 and emitting non-lexical babble, whereas explicit cross-attention over ConvNeXt-refined character features aligned cleanly on the same data and descended to 0.918, supporting the hypothesis that *at mid-resource scale, explicit cross-attention conditioning is markedly more sample-efficient than implicit alignment* (section 4.4).

3. **A pretrain-to-SFT production-hardening recipe.** A full-parameter *voice-lock* stage on 76.5 hours of a single consented speaker writes the voice into the weights, collapsing cross-generation F_0 standard deviation from 74.9 to 5.0 Hz, and a *short-utterance-coverage* stage of 48,424 forced-aligned one- and two-word segments repairs the isolated-token failures a sentence-only corpus structurally cannot.

4. **A public benchmark and an efficiency-oriented evaluation.** We release **Freya-TR-Eval**, a 495-sentence, domain-neutral Turkish benchmark, and evaluate under an identical band-matched Whisper WER/CER protocol: FreyaTTS (WER 8.0 %, CER 3.0 %) outperforms open systems several times its size, including XTTS-v2 and F5-TTS. Batched-ODE serving reaches 65.2 audio-seconds per second at batch 8 on a single RTX 4090 – 2.6× the 2B VoxCPM2 serving engine at a third of the memory – and the model runs in real time on a laptop CPU.

The remainder of this report is organized as follows. Section 2 situates FreyaTTS within large-scale and low-resource TTS foundation models and continuous-latent flow-matching generation. Section 3 presents the FreyaTTS system: the tokenizer-free non-autoregressive architecture over the frozen VoxCPM2 latent space, the flow-matching and duration objectives, the from-scratch Turkish pretraining stage, and the two-stage voice-lock and short-utterance-coverage fine-tuning recipe. Section 4 reports the Turkish evaluation against openly available systems together with inference- and serving-efficiency measurements. Section 5 discusses limitations and future directions.

2 Related Work

We review prior work along the two axes that situate FreyaTTS: the generative paradigm it instantiates (non-autoregressive flow matching over continuous latents; section 2.1) and the regime it targets (efficient, mid-resource Turkish synthesis; section 2.2).

2.1 Non-Autoregressive Flow-Matching and Continuous-Latent TTS

Contemporary TTS divides into two broad families according to how speech is represented: as sequences of *discrete* tokens emitted by a neural audio codec, or as *continuous* acoustic features rendered by a denoising or flow-matching model. FreyaTTS belongs to the latter; we review both to make the distinction precise.

Discrete-token codec language models. The dominant paradigm casts synthesis as autoregressive language modeling over discrete acoustic tokens produced by a neural audio codec (Défossez et al., 2022; Kumar et al., 2023), thereby inheriting the scaling behavior and in-context learning of LLMs. VALL-E (Chen et al., 2025a) showed that a decoder-only LM over residual vector-quantized tokens performs zero-shot voice cloning by prompt continuation, and the CosyVoice family (Du et al., 2024, 2025) pairs an LM that predicts supervised semantic tokens with a separate flow-matching decoder that restores acoustic detail. This route is powerful but pays two prices: quantization irreversibly discards fine-grained acoustic information, and the semantic-token / acoustic-decoder split fragments high-level planning from waveform rendering into stages that cannot be optimized jointly end to end.

Non-autoregressive flow matching over continuous features. A complementary family dispenses with discrete tokens and models continuous acoustic features directly and in parallel, under a diffusion or conditional-flow-matching objective. NaturalSpeech 2 (Shen et al., 2023) applies latent diffusion over a neural-codec latent with explicit duration and pitch predictors, and Voicebox (Le et al., 2023) casts TTS as text-conditioned flow matching over mel-spectrogram frames guided by an external forced alignment and a duration model. This alignment-and-duration template has a longer lineage in diffusion TTS: Grad-TTS (Popov et al., 2021) pairs a score-based diffusion decoder with monotonic-alignment search and a duration predictor, and Matcha-TTS (Mehta et al., 2024) recasts that decoder under optimal-transport conditional flow matching; it is the closest architectural relative to FreyaTTS, from which we differ by generating a *frozen VoxCPM2 latent* rather than a mel-spectrogram and by conditioning through cross-attention rather than monotonic-alignment search. Other continuous and diffusion systems relax the mel target differently: StyleTTS2 (Li et al., 2023) adds style diffusion and adversarial training, NaturalSpeech 3 (Ju et al., 2024) factorizes a codec latent for attribute-wise diffusion, and end-to-end VAE-GAN models in the VITS (Kim et al., 2021) family (on which the MMS Turkish baseline is built (Pratap et al., 2024)) jointly learn alignment and waveform in one network. E2-TTS (Eskimez et al., 2024) then removed both the duration predictor and the alignment module: it pads the character sequence with filler tokens to the target mel length and lets a single flat Transformer learn the monotonic text-to-audio correspondence *implicitly* through self-attention. F5-TTS (Chen et al., 2025b) keeps this filler-padded implicit-alignment recipe while adding a ConvNeXt text refiner and sway-sampling inference, and lightweight successors such as ZipVoice (Zhu et al., 2025) push the same flow-matching formulation toward compact, fast models. Implicit alignment is architecturally elegant, but, as we show in section 2.2, it becomes learnable only at very large data scale, a constraint that directly shaped our design.

Diffusion-autoregressive continuous-latent hybrids. A third family autoregresses over continuous latents while rendering each step with a local diffusion head, combining LLM-style long-range planning with continuous acoustic fidelity (Meng et al., 2025). The mechanism originates in image generation: MAR (Li et al., 2024) replaced the categorical softmax of a token-based autoregressive model with a small per-token diffusion loss, demonstrating that vector quantization is not required for autoregressive generation. DiTAR (Jia et al., 2025) ported this to speech through a Local Diffusion Transformer that denoises each patch of continuous latents conditioned on autoregressive context, and VoxCPM (Zhou et al., 2025) realized a semantic-acoustic hierarchy inside a single continuous-latent backbone using a differentiable semi-discrete FSQ bottleneck (Mentzer et al., 2024) atop a MiniCPM-4 language model (Team et al., 2025), training end to end without any external discrete tokenizer. These hybrids inherit the expressiveness of LLMs but retain an autoregressive decode loop, with its attendant exposure bias and error accumulation over long or number-dense utterances.

Positioning FreyaTTS. FreyaTTS sits in the non-autoregressive flow-matching family of Matcha-TTS, Voicebox, E2-TTS, and F5-TTS, but departs from all of them in both its representation and its alignment. Rather than generating mel-spectrograms, it operates over the *continuous latent of a frozen, pretrained AudioVAE*, the 25 Hz, 64-dimensional, 16 kHz-in / 48 kHz-out codec of VoxCPM2 (Zhou et al., 2025), reused unmodified under its Apache-2.0 license and never trained. This decouples representation learning from generative modeling: the diffusion transformer need only learn a text-to-latent map, and 48 kHz super-resolution output comes for free from the frozen decoder. Unlike the diffusion-autoregressive route of DiTAR and VoxCPM, FreyaTTS discards the hierarchical TSLM/FSQ/RALM stack and the autoregressive decode loop entirely, predicting an utterance duration and then denoising all latent frames in parallel with a single 183.2 M-parameter flow-matching transformer (a 32-step Euler ODE at inference). And unlike the implicit alignment of E2/F5, it conditions generation through *explicit cross-attention* from latent-frame queries onto character features: the choice that makes training converge at our data scale (section 2.2). Finally, it is tokenizer-free on both ends: no discrete audio codec, and a 92-symbol character vocabulary with no phonemizer or grapheme-to-phoneme frontend, so that the pronunciation of numbers, acronyms, and code-switched terms is learned end to end from audio rather than dictated by a rule-based frontend.

2.2 Low-Resource and Turkish TTS

The mid-resource multilingual gap. Large-scale TTS is overwhelmingly trained on English- and Chinese-dominant read and audiobook corpora such as Emilia (He et al., 2024), and its dominant zero-shot benchmarks, for example Seed-TTS-Eval (Anastassiou et al., 2024), are likewise English/Chinese-centric. Massively multilingual systems extend nominal coverage to dozens or hundreds of languages (Casanova et al., 2024; Du et al., 2025; Zhu et al., 2026), and Turkish is typically present among them, but a single multilingual checkpoint spreads capacity across all of its languages and is trained on generic read speech. Turkish specifically occupies a *mid-resource* band: public read-speech corpora such as FLEURS-tr (Conneau et al., 2023) and Common-Voice-tr (Ardila et al., 2020) supply only on the order of tens of hours, far better served than genuinely low-resource languages yet an order of magnitude below English or Chinese, and with no specialization to any target conversational domain. Neither the tens-of-thousands-of-hours recipes of the large multilingual models nor the few-shot adaptation tricks of truly low-resource work map cleanly onto this regime, which is the operating point FreyaTTS is built for.

Why implicit alignment fails at mid-resource scale. The elegance of E2-TTS and F5-TTS (no duration model, no alignment module) is bought with data: these systems learn implicit alignment only at very large scale, on the order of 50–95 k hours of speech (Eskimez et al., 2024; Chen et al., 2025b), exemplified by the ∼95 k-hour Emilia corpus (He et al., 2024) used to train F5-TTS. We find this requirement to be load-bearing rather than incidental. On our Turkish corpus, at our mid-resource budget, a faithful E2/Voicebox-style implicit-alignment variant, the character sequence filler-padded to the latent length so that self-attention must discover the alignment, failed to align: the flow-matching loss plateaued near 1.06 and the model emitted fluent-sounding but non-lexical *babble*, a posterior collapse in which the decoder ignores the conditioning text. Replacing implicit alignment with explicit cross-attention resolved this at the same data scale, consistent with the broader observation of MegaTTS 3 (Jiang et al., 2025) that alignment guidance stabilizes latent-diffusion synthesis on difficult inputs. This sample-efficiency observation, that at mid-resource scale explicit cross-attention conditioning markedly outperforms implicit alignment, is one of the central findings developed in sections 3 and 4; section 4.4 states its scope.

Conversational and short-utterance synthesis. Conversational Turkish has a distribution largely absent from the audiobook corpora above: it is dense with numbers, dates, and amounts that must be voiced digit-accurately, it code-switches into foreign names, and it is punctuated by very short backchannels and acknowledgments that carry conversational turn-taking. Two properties of this register shaped our approach. First, long numeric strings are a demanding case for either decoding paradigm: autoregressive decoding can compound errors across them, while parallel synthesis must commit to a duration for them up front; section 4.3 measures this trade-off head-to-head instead of assuming it away. Second, read-speech corpora contain essentially no ultra-short utterances (our single-speaker fine-tuning corpus has zero clips of two words or fewer), so isolated acknowledgments and bare numbers are out of distribution and degrade without targeted coverage. FreyaTTS therefore pretrains from scratch on a mid-resource corpus of multi-speaker Turkish and is then post-trained into a single consented voice with dedicated short-utterance data, converting a general-purpose Turkish synthesizer into a reliability-hardened conversational one; we detail this curriculum in section 3.

3 Methodology

3.1 Overview

FreyaTTS is a tokenizer-free, non-autoregressive (NAR) text-to-speech model realized as a conditional flow-matching Diffusion Transformer (DiT) over the continuous latent space of a *frozen* VoxCPM2 AudioVAE. It shares VoxCPM's core premise, modeling speech in a continuous latent space rather than over discrete codec tokens (Zhou et al., 2025), but departs sharply from its generative backbone. VoxCPM and VoxCPM2 render audio autoregressively: a MiniCPM-4 language model (Team et al., 2025) predicts semi-discrete FSQ semantic states (Mentzer et al., 2024) that a local diffusion head decodes patch by patch. We instead discard the entire

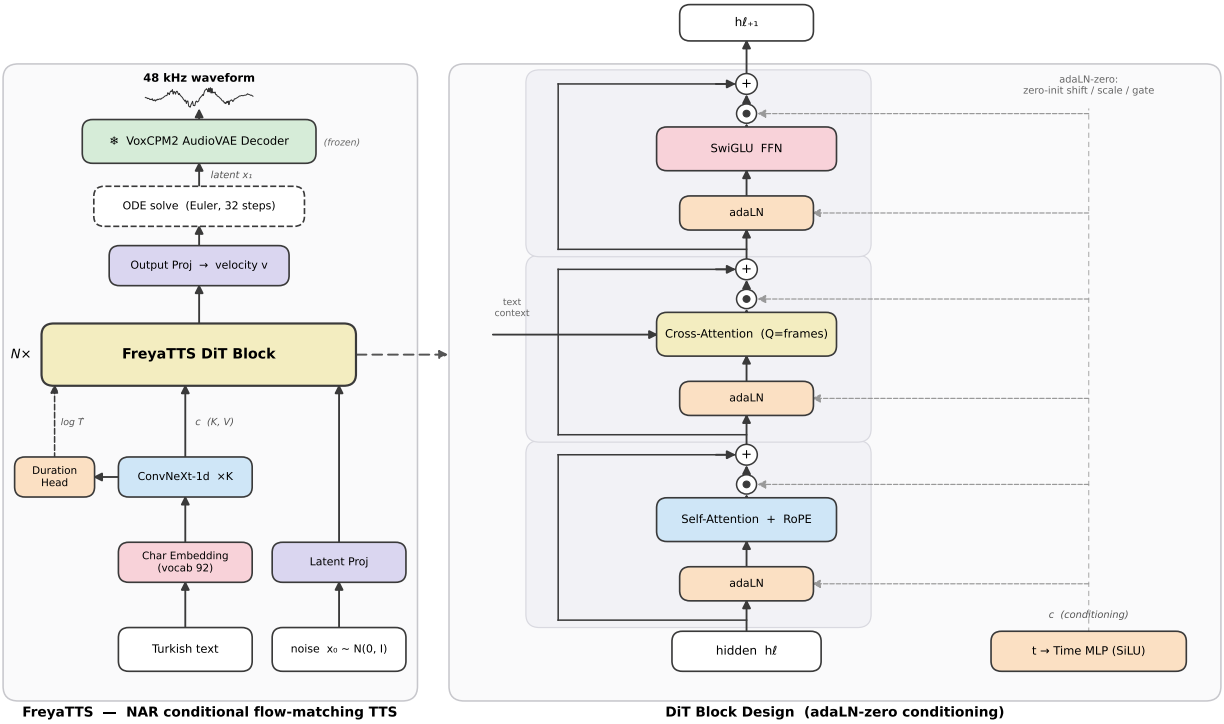


Figure 1: Overall architecture of FreyaTTS. Character-level Turkish text (a 92-symbol vocabulary; no byte-pair encoding, phonemizer, or grapheme-to-phoneme frontend) is embedded and refined by a ConvNeXt-1d encoder into a character-feature sequence c, which (i) drives a duration head that predicts the total latent length $\hat{T}$ and (ii) serves as the key/value memory for the cross-attention layers of a non-autoregressive Diffusion Transformer (DiT). Conditioned on the flow-matching timestep t through adaLN-zero modulation, the DiT denoises a length-$\hat{T}$ sequence of 64-dimensional latent frames drawn from Gaussian noise; the resulting clean latents are rendered to 48 kHz audio by the *frozen* VoxCPM2 AudioVAE decoder. The autoregressive TSLM/FSQ/RALM backbone of VoxCPM2 is discarded; only its AudioVAE is reused.

autoregressive stack (the text-semantic LM, the FSQ bottleneck, the residual acoustic LM, and the local diffusion transformer) and keep only the AudioVAE, replacing the backbone with a single NAR DiT that denoises all latent frames of an utterance in parallel (fig. 1).

This choice is dictated by reliability. Conversational Turkish is dense with digit strings, dates, and amounts, and autoregressive latent prediction can accumulate error across such sequences, garbling numbers mid-utterance. Predicting the whole latent sequence in parallel removes the left-to-right dependency (section 4.3 measures numeric fidelity directly, including where the NAR design introduces its own bottleneck); it also decouples the number of *sequential* generation steps from utterance length (a fixed ODE-step budget replaces a length-proportional decode loop, though each step still attends over all T frames) and pairs naturally with flow matching (Le et al., 2023; Chen et al., 2025b; Anastassiou et al., 2024), in contrast to the multi-stage LM-plus-vocoder pipelines of discrete-token systems (Du et al., 2025).

Formulation. Let $x_1 \in \mathbb{R}^{T\times 64}$ be the clean latent sequence produced by the frozen AudioVAE encoder and $c = \text{TextEnc}(y)$ the character-feature memory of the input text y. We adopt a conditional flow-matching objective with a linear probability path, which interpolates on a straight line between a Gaussian prior $x_0 \sim \mathcal{N}(0, I)$ and

the data x_1 and trains a velocity field v_θ to transport one to the other:

$$x_t = (1-t)\,x_0 + t\,x_1, \qquad x_0 \sim \mathcal{N}(0, I), \quad t \sim \mathcal{U}[0,1],$$

$$\mathcal{L} = \underbrace{\mathbb{E}_{t,x_0,x_1}\left[\left\| m \odot \left(v_\theta(x_t, t, c) - \overbrace{(x_1 - x_0)}^{v^\star}\right)\right\|_2^2\right]}_{\text{masked flow-matching}} + \lambda_{\text{dur}}\left(\log \hat{T} - \log T\right)^2. \tag{1}$$

Here $v^\star = x_1 - x_0$ is the target velocity, which is *constant* along the straight path so that a single network evaluation supervises the entire trajectory; $m \in \{0,1\}^T$ masks padding frames, and the flow-matching term is averaged over the $\|m\|_1$ real frames; $\hat{T}$ is the length predicted by the duration head (section 3.3); and $\lambda_{\text{dur}} = 0.1$. Because x_0 and x_1 are sampled independently rather than optimally coupled, this is the linear-path, *rectified-flow* instance of conditional flow matching (Lipman et al., 2023; Tong et al., 2024), not a minibatch optimal-transport coupling. There is no adversarial loss, no autoregression, and no discrete-token cross-entropy: only the masked flow-matching regression and the auxiliary duration term. Because the model consumes raw characters and no discrete speech tokenizer intervenes, the pronunciation of in-context numbers, currencies, and acronyms is learned end-to-end from audio rather than delegated to a text frontend; isolated tokens are harder and are addressed by post-training and inference-time guards (section 3.5).

3.2 Frozen VoxCPM2 AudioVAE Latent Space

FreyaTTS operates entirely inside the latent space of the VoxCPM2 AudioVAE, which we reuse *frozen* (Apache-2.0) and never train. Its encoder maps a 16 kHz waveform to 64-dimensional latent frames at 25 Hz (one frame per 40 ms), while its decoder reconstructs at 48 kHz. This 16 kHz-in / 48 kHz-out asymmetry provides implicit super-resolution: even though our single-speaker fine-tuning corpus is 16 kHz telephony-narrowband (section 3.5), the decoder renders full 48 kHz output.

Reusing a strong pretrained codec is what makes training a 183.2M-parameter generator from scratch at a mid-resource data budget (section 3.4) tractable: high-fidelity waveform reconstruction is already solved and held fixed, so all model capacity and training data are spent on the text-conditional generative prior. Modeling continuous latents rather than discrete tokens preserves fine acoustic detail that quantization discards (Shen et al., 2023; Le et al., 2023) and keeps the pipeline tokenizer-free end to end. We stress that we retain only the *continuous* AudioVAE: the semi-discrete FSQ bottleneck (Mentzer et al., 2024) that gives VoxCPM its hierarchical semantic-acoustic factorization lives inside the autoregressive backbone we remove and plays no role in FreyaTTS. Finally, whereas VoxCPM2 groups P=4 frames into 160 ms patches to shorten its autoregressive loop, our NAR DiT has no such constraint and operates directly on the native 25 Hz grid, so the sequence length T is simply the number of 40 ms frames in the utterance.

3.3 The FreyaTTS Diffusion Transformer

The velocity field v_θ is a DiT that takes the noisy latent sequence $x_t \in \mathbb{R}^{T\times 64}$, the timestep t, and the character memory c, and predicts a velocity at every frame. Table 1 lists the configuration.

Text encoder and duration head. The text encoder embeds the character-level Turkish vocabulary (92 symbols in total: the 29-letter alphabet including ç, ğ, ı, ö, ş, ü, together with digits, punctuation, and whitespace) and refines it with four ConvNeXt-1d blocks (Chen et al., 2025b; Liu et al., 2022) into the feature sequence c, shared by the cross-attention memory and the duration head. Operating on raw characters means that Turkish orthographic phenomena (vowel harmony, the dotted/dotless ı distinction, and the spoken form of acronyms) are absorbed into the weights rather than handed to a grapheme-to-phoneme frontend. Digit strings are the deliberate exception: their orthographic and phonetic lengths diverge so sharply that we expand them to spoken form at the text frontend, a bottleneck section 4.3 isolates in the duration predictor rather than in the acoustic model; isolated acronyms and bare tokens remain a documented failure mode addressed in sections 3.5 and 4. Since generation is

Table 1: FreyaTTS configuration. The AudioVAE is reused frozen from VoxCPM2; only the DiT and text encoder (183.2M parameters) are trained.

Component	Setting
Latent space (frozen AudioVAE)	64-dim @ 25 Hz; 16 kHz encode / 48 kHz decode
Text vocabulary	character-level Turkish, 92 symbols (no BPE, no G2P)
Text encoder	ConvNeXt-1d $\times 4$
DiT width / depth / heads	d=640 / 16 / 10
DiT feed-forward	SwiGLU, hidden 2048
Self-attention position encoding	RoPE
Time conditioning	adaLN-zero (9-way modulation)
Text conditioning	cross-attention (frame queries $\rightarrow$ character keys/values)
Duration objective weight λ_{dur}	0.1 (pretraining)
Trainable parameters	183.2M
Inference solver	32-step Euler ODE

non-autoregressive, the target length must be known before denoising: a small MLP over the mask-averaged mean of c regresses $\log \hat{T}$, supervised by the auxiliary term of eq. (1) and used to size the noise tensor at inference.

DiT block. Each of the 16 layers updates the frame hidden states $h \in \mathbb{R}^{T \times d}$ (d=640) through three residual sub-layers (RoPE self-attention, text cross-attention, and a SwiGLU feed-forward network), each modulated by the flow-matching timestep through adaLN-zero:

$$
\begin{aligned}
(\beta_1, \gamma_1, g_1,\ \beta_2, \gamma_2, g_2,\ \beta_3, \gamma_3, g_3) &= \mathrm{MLP}_{\text{zero}}\big(\mathrm{emb}(t)\big), \\
h &\leftarrow h + g_1 \odot \mathrm{SelfAttn}_{\text{RoPE}}\big(\gamma_1 \odot \mathrm{LN}(h) + \beta_1\big), \\
h &\leftarrow h + g_2 \odot \mathrm{CrossAttn}\big(\gamma_2 \odot \mathrm{LN}(h) + \beta_2,\ c\big), \\
h &\leftarrow h + g_3 \odot \mathrm{SwiGLU}\big(\gamma_3 \odot \mathrm{LN}(h) + \beta_3\big),
\end{aligned}
\tag{2}
$$

where $\mathrm{emb}(t)$ is a sinusoidal timestep embedding and $\odot$ applies per-channel shift/scale/gate broadcast over frames. The nine modulation vectors are produced by a per-layer MLP whose final projection is *zero-initialized* (adaLN-zero): every block therefore begins as the identity and departs from it only as training warrants, a stabilizer standard in diffusion transformers (Peebles & Xie, 2023; Li et al., 2024; Jia et al., 2025). Self-attention carries rotary position embeddings (Su et al., 2024) over the frame axis; in contrast to the NoPE residual LM of VoxCPM2 (Kazemnejad et al., 2023), we retain RoPE because our sequence is generated in one shot and benefits from explicit positional structure. In CrossAttn the queries are the modulated frame states while the keys and values are the character features c: this is the sole pathway through which text conditions acoustics.

Explicit cross-attention over implicit alignment. The tokenizer-free NAR systems E2 TTS (Eskimez et al., 2024) and F5-TTS (Chen et al., 2025b) dispense with an alignment module: the character sequence is padded with filler to the latent length T, concatenated channel-wise with the noisy latents, and self-attention is left to discover the text–acoustic alignment. This is elegant but data-hungry: both were trained on 50–95k hours. At our mid-resource target budget (section 3.4) we found the implicit-alignment variant fails to align: the flow-matching loss plateaus near 1.06 and the decoder emits unintelligible babble (posterior collapse). Routing text through dedicated cross-attention layers instead, so that every frame can sparsely and directly attend to the relevant characters, in the spirit of alignment-and-duration diffusion decoders (Popov et al., 2021; Mehta et al., 2024; Shen et al., 2023), proved markedly more sample-efficient and aligned cleanly at the same budget. We regard this as a central design finding of FreyaTTS; section 4.4 details the comparison and its scope.

3.4 Pretraining from Scratch

The DiT and text encoder are trained from random initialization (the AudioVAE contributes no gradients) on an internal multi-speaker corpus of 881k Turkish clips totaling ~1,200 hours, pre-encoded offline into VoxCPM2

latents so that the frozen encoder never runs during training, and spanning number-string reads, name-list reads, conversational sentences, and general speech. We optimize eq. (1) in bf16 with AdamW (learning rate 5×10^{-4}, weight decay 0.01, gradient clipping at 1.0, 2,000-step linear warmup then cosine decay) for 150k steps at batch size 64, roughly 11 epochs over the corpus, in a single training run with a fixed seed. Training runs on H100 and H200 GPUs with a memory-safe loader over the pre-encoded fp16 latents.

Loss trajectory and behavior. The flow-matching loss descends monotonically, from 2.38 at step 200 through 1.29 at 1.5k, 1.036 at 20k, and 0.995 at 38k to 0.918 at 150k, while the duration term settles near 0.013. Re-transcribing generations with Whisper-large-v3 (Radford et al., 2023), intelligible and well-aligned Turkish emerges from ~20k steps, with spot checks showing in-context numbers read in their full spoken form; section 4.3 later quantifies numeric fidelity systematically, including the failure mode spot checks miss. The pretrained model has, however, *no speaker identity*: with no speaker field in the corpus and no speaker conditioning in the network, the prior x_0 seeds a fresh speaker each generation, so the median F0 spans 70–347 Hz across regenerations of the same text, with a standard deviation of 74.9 Hz and a 60% gender-flip rate. This is expected behavior for an unconditioned prior and directly motivates the voice-lock post-training below.

3.5 Post-training

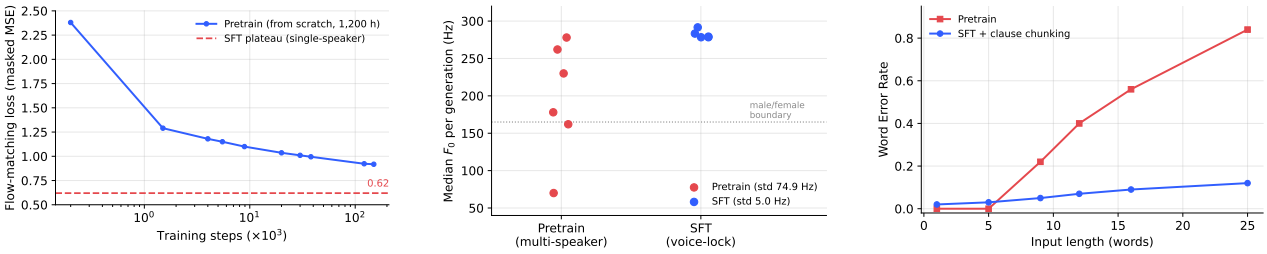
Two supervised fine-tuning stages turn the speaker-agnostic pretrained model into a deployable single-voice product.

SFT v1: single-speaker voice lock. Initialized from the 150k-step checkpoint, we full-parameter fine-tune (the AudioVAE still frozen) on an internal single-speaker corpus: 56,295 clips totaling 76.5 hours of a single consented professional voice talent, recorded at 16 kHz over a narrowband channel matching the deployment domain. Training uses multi-GPU DDP, bf16, and learning rate 1×10^{-4} with cosine decay; the flow-matching loss falls from 0.918 to ≈0.62 within roughly one epoch. The speaker identity is written into the weights: cross-generation F0 standard deviation collapses from 74.9 to 5.0 Hz and content similarity reaches 0.923, while domain acronyms stabilize, pronounced as words rather than spelled letter by letter. We select **step 1000**: continued training over-fits and destabilizes the voice, with the voice-lock standard deviation drifting from 5.0 back to 11 Hz by 3k steps.

SFT v2: short-utterance coverage. The single-speaker corpus contains *no* clips of two words or fewer, which we identified as the structural root cause of a collapse on isolated acknowledgments, bare numbers, and lone acronyms. We therefore continue fine-tuning from the v1 step-1000 checkpoint on the same corpus augmented with 48,424 short single-speaker segments (one- and two-word phrases, numbers, and acknowledgments) that we extracted from the same speaker by MMS forced alignment (torchaudio `MMS_FA` with a Turkish-to-roman character map). This continuation stage uses learning rate 5×10^{-5} and λ_{dur}=0.2, and we again select step 1000. Isolated acknowledgments are rendered once, cleanly, rather than entering a repetition loop; in-context numbers are correct; and full-sentence quality and the locked voice are preserved. **SFT v2 is the shipped production model.** A thin inference wrapper completes the system: clause-level chunking for long inputs, a duration floor for very short ones, and a voicing-based retry on the rare degenerate draw.

3.6 Inference

Given input text, the text encoder produces c and the duration head predicts the latent length $\hat{T}$. We draw $x_0 \sim \mathcal{N}(0, I)$ of length $\hat{T}$ and integrate the flow-matching ODE $\mathrm{d}x/\mathrm{d}t = v_\theta(x_t, t, c)$ from t=0 to t=1 with a fixed-step Euler solver (32 steps), evaluating the DiT once per step with c supplied through cross-attention. The resulting clean latent sequence is decoded by the frozen AudioVAE into a 48 kHz waveform. Because generation is non-autoregressive, the whole sequence is denoised in parallel: the number of *sequential* solver steps is fixed at 32 regardless of utterance length (each step still attending over all T frames), and there is no left-to-right error accumulation. Unlike the autoregressive sampler of VoxCPM2, which leans on classifier-free guidance, sway



(a) Training loss. **(b)** Voice-lock (F_0). **(c)** WER vs. length.

Figure 2: Left: flow-matching loss over pretraining (log-scale steps) with the single-speaker SFT plateau. **Center:** median F_0 across regenerations of a fixed prompt; the multi-speaker pretrained model scatters across an octave (std 74.9 Hz) whereas the voice-locked SFT model is stable (std 5.0 Hz). **Right:** word error rate as a function of input length on a length-swept probe set; long-horizon drift in the pretrained model is mitigated, though not removed, by inference-time clause chunking.

sampling (Chen et al., 2025b), and CFG-Zero* (Fan et al., 2025), we find a plain 32-step deterministic Euler integrator sufficient; the inference wrapper re-draws x_0 only on the rare voicing-check failure. Empirically the shipped FreyaTTS model attains a mean per-utterance real-time factor of ≈0.14 on an H100, which we detail in section 4.6.

4 Experiments and Results

We evaluate FreyaTTS along five axes: main results against the field of openly-available, sub-billion-parameter Turkish TTS systems on a single held-out benchmark (sections 4.1 and 4.2); numeric fidelity under digit-dense input (section 4.3); the cross-attention-versus-implicit-alignment comparison (section 4.4); speaker consistency and the voice-lock stage (section 4.5); and inference efficiency (section 4.6). Our central comparative result is that on natural, everyday Turkish, a 183M model trained from scratch outperforms every larger system in the openly available sub-billion-parameter Turkish field, and all weaker same-class models, carrying its target voice in its weights rather than in a reference prompt.

4.1 Experimental Setup

Benchmark. We evaluate on a single, purpose-built benchmark, **Freya-TR-Eval**: 495 natural, domain-neutral Turkish sentences of everyday conversational register (statements, *mI*-questions, and exclamations), length 3–13 words. It merges native Turkish sentence texts sampled from Common Voice 17-tr (Ardila et al., 2020) and CoVoST2-tr (Wang et al., 2021) with everyday-conversational lines generated by `gemini-3.1-pro-preview` under fixed topic and sentence-type quotas (22 everyday topics), filtered for length, full Turkish grapheme coverage (ç ğ ı i ö ş ü), and de-duplication. FreyaTTS is trained only on internal corpora (sections 3.4 and 3.5) disjoint from the benchmark sources, so every item probes generalization; the exact sentence list and the seeded build scripts are released.[1] Because FreyaTTS is a single-voice synthesizer rather than a zero-shot cloner, we synthesize the reference text of each item and score the output; the two reference-prompt baselines (XTTS-v2, F5-TTS) are given one fixed shared reference clip (cloner scores can be sensitive to this choice).

Comparison systems. We position FreyaTTS, the shipped 183.2M single-voice model, against the field of *openly available, sub-billion-parameter* Turkish TTS systems, the models a practitioner can download and self-host: **MMS-TTS-tr** (Pratap et al., 2024) (36M, Meta's VITS-based multilingual TTS), **Coqui GlowTTS-tr** (~28M, flow-based), **Piper-tr** (~16M, VITS, on-device), **SpeechT5-tr** (Ao et al., 2022) (~144M, fine-tuned on Turkish Common Voice), **XTTS-v2** (Casanova et al., 2024) (~470M, multilingual zero-shot cloner), and **F5-TTS-Turkish** (Chen et al., 2025b) (~336M, a flow-matching zero-shot cloner fine-tuned for Turkish). We do *not* compare against multi-billion-

[1] Released as `freyavoice/freya-tr-eval` on HuggingFace.

Table 2: Conversational Turkish TTS, sub-1B open models. All systems evaluated on the full Freya-TR-Eval set (495 everyday, general-purpose Turkish sentences) under an identical protocol: **WER/CER** from Whisper-large-v3 with 8 kHz band-matching (in %), and naturalness **MOS** from a listening study with 7 native Turkish raters (760 ratings, 105–112 per system, 5-point scale, 95% confidence intervals). Best per column in **bold**; systems ordered by size.

System	Params	WER ↓	CER ↓	MOS ↑
Piper (tr, dfki)	16M	**4.4**	**1.1**	3.47 ± 0.22
Coqui GlowTTS (tr)	28M	12.1	3.3	2.53 ± 0.19
MMS-TTS (tr)	36M	6.8	1.7	3.58 ± 0.20
SpeechT5 (tr)	144M	83.4	45.5	1.59 ± 0.14
FreyaTTS (ours)	183.2M	8.0	3.0	3.68 ± 0.22
F5-TTS (tr)	336M	24.3	10.9	3.63 ± 0.23
XTTS-v2 (multi)	470M	11.1	3.9	**3.82** ± 0.19

parameter or closed production systems: the contribution we are measuring is a compact, openly-deployable model, so the relevant reference set is other compact open models of the same class.

Metrics. The primary metrics are **band-matched WER and CER**: every system's output is downsampled to 8 kHz before transcription with Whisper-large-v3 (Radford et al., 2023), so that FreyaTTS's telephony-band 48 kHz output (section 3.2) is not penalized relative to natively wideband baselines, and reference and transcript pass the same Turkish text normalization before scoring; the protocol is applied identically to every system. We also report **real-time factor** (RTF): per-utterance synthesis time over audio duration at batch size 1, including AudioVAE decoding, averaged over the set, on an H100. Naturalness is measured with a **mean opinion score** (MOS) listening study: 7 native Turkish raters scored randomized, system-blind samples of every system on a 5-point scale (760 ratings in total, 105–112 per system); we report per-system means with 95% confidence intervals. **Voice consistency** (section 4.5) is the F_0 standard deviation across regenerations of the same text. **Content similarity** is 1 minus the normalized character-level edit distance between the input text and the Whisper-large-v3 transcript of the synthesis.

4.2 Main Results

FreyaTTS outperforms the larger open systems in its field. Table 2 reports band-matched Whisper WER/CER for every system on the same 495 conversational sentences. FreyaTTS reaches WER 8.0 % and CER 3.0 %, **outperforming both larger open systems in the field**, the zero-shot cloners XTTS-v2 (~470M, WER 11.1 %) and F5-TTS (~336M, WER 24.3 %), each 1.8–2.5× its parameter count, despite their larger backbones and far greater training data: cloning a target voice from a short reference generalizes poorly to arbitrary sentences, whereas FreyaTTS carries a single voice natively in its weights. It also comfortably beats the weaker same-class models (GlowTTS, SpeechT5). Piper (~16M, WER 4.4 %) and MMS-TTS (~36M, WER 6.8 %) are compact phonemizer-driven VITS baselines whose highly regular pronunciation maximizes recognizer agreement; FreyaTTS supplies the single, consistent *learned* voice they lack (section 4.5). The listening study bears this out on naturalness: FreyaTTS reaches MOS 3.68, above every same-class system including the two low-WER VITS baselines (Piper 3.47, MMS-TTS 3.58) and F5-TTS (3.63), and second overall only to the 2.6×-larger XTTS-v2 (3.82); the 95% intervals of the top four systems overlap, so we read the MOS column as a ranking rather than pairwise significance. That a 183M model trained from scratch outperforms every larger open system in this field is the paper's central empirical claim.

Statistical robustness, calibration, and the inference layer. Three checks contextualize the headline numbers. First, a sentence-level bootstrap (10,000 resamples over the 495 items, computed on an independent re-synthesis of the full set) gives WER 7.9% with a 95% confidence interval of [6.8, 9.1] and CER 2.9% [2.4, 3.4], reproducing table 2 within run-to-run noise. Second, two anchors calibrate the protocol's absolute scale: real human recordings (FLEURS-tr test (Conneau et al., 2023)) transcribe at 9.7% WER under the identical band-matched pipeline,

and re-encoding held-out *real* recordings of the target speaker through the frozen AudioVAE alone costs 2.6% WER – the transcription floor of the latent space itself. FreyaTTS's 8.0% therefore sits below the protocol's human-speech operating point and within 5.4 points of its floor. Third, ablating the entire inference layer of section 3.6 (normalization, clause chunking, duration floor, voicing retry) moves WER by at most 0.3 points (paired bootstrap 95% CI of the difference $[-0.3, 0.0]$); the voicing retry fires on 1 of 495 sentences and clause chunking on 7 of 495. The guards are tail insurance against rare failure modes, not a crutch the headline numbers depend on.

4.3 Numeric Fidelity

Conversational Turkish is dense with digit strings, and section 1 motivated the non-autoregressive formulation partly by the tendency of autoregressive decoders to garble them. We test that argument directly rather than leave it as folklore: a numeric stress set of 120 deterministic, number-dense Turkish sentences spanning four categories (numbers in running context, currency amounts, dates and times, and bare standalone figures), scored by transcribing each synthesis under the protocol of section 4.1, parsing spelled-out Turkish number words back to digits, and comparing against the reference digit string – sentence-level exact match plus a digit error rate (DER).

The outcome inverts the naive expectation, and we report it in full. Fed raw digit strings, FreyaTTS reaches only 8.3% exact match (DER 0.556), while the ~2B autoregressive VoxCPM2, which decides length as it decodes, reads the same set at 64.2% (DER 0.159). The binding constraint is not decoding order but *duration*: a digit string is orthographically short yet phonetically long (247 is three characters but seventeen in its spoken form), so the character-level duration head systematically under-allocates frames and the rendition truncates. Expanding digits to their spoken form at the text frontend – the standard input contract for character-level synthesis, requiring no retraining – recovers FreyaTTS to 40.8% exact match overall and 73.3% (DER 0.15) on the in-context category. These figures are conservative lower bounds: the recognizer writes digit-voiced outputs as digits but spelled outputs as case-suffixed Turkish number words, which the digit parser recovers imperfectly – a penalty the autoregressive baseline never pays. The takeaway is twofold. Parallel decoding alone does not solve numeric fidelity; explicit duration prediction becomes the bottleneck exactly where orthographic and phonetic length diverge. And with frontend expansion, the compact model recovers most of the gap at 11× fewer parameters. Calibrating the duration predictor on digit-dense text and folding spoken-form number pairs into training are the immediate remedies, and we fold both into the ongoing scaled pretraining run.

4.4 Cross-Attention versus Implicit Alignment

Our single most consequential architectural decision is how text conditions the latent sequence. The tokenizer-free non-autoregressive systems E2 TTS (Eskimez et al., 2024) and F5-TTS (Chen et al., 2025b), in the lineage of Voicebox (Le et al., 2023), use *implicit alignment*: the character sequence is padded with a filler symbol to the latent length T, concatenated channel-wise with the noisy latents, and self-attention is left to discover the text–acoustic alignment. Trained on our corpus, this variant did not learn to align: the flow-matching loss plateaued near 1.06 and the AudioVAE decoded the resulting latents as unintelligible babble. Routing text instead through dedicated cross-attention layers (frame queries attending to ConvNeXt-refined character features (Popov et al., 2021; Mehta et al., 2024; Shen et al., 2023)) resolved the collapse: the model aligned and its loss descended to 0.918 (fig. 2a), yielding intelligible, well-aligned Turkish from ~20k steps. This is a single development-scale comparison, and since the two variants parameterize different networks the loss gap is indicative. It is nonetheless consistent with data scale as the operative constraint: E2 TTS and F5-TTS succeed with implicit alignment at roughly 50–95k hours, one to two orders of magnitude beyond ours. The transferable design lesson: *explicit cross-attention conditioning is markedly more sample-efficient than implicit alignment at mid-resource scale*. Figure 2c further characterizes the resulting model's behavior as a function of utterance length.

Table 3: Inference efficiency on a single RTX 4090 (fp32/bf16 as released, batch size 1, end-to-end including vocoder/VAE decoding; median of 10 runs after 3 warmups on a fixed Turkish prompt set of three length buckets (five sentences each; released with our code)). **TTFT** = time to first audio on the long bucket (22–28 words): native streaming first-chunk for XTTS-v2, first synthesized clause for FreyaTTS, full-utterance latency for the offline systems. **RTF** = synthesis time over audio duration. **Throughput** = aggregate audio-seconds per wall-second over a 45-utterance queue at the best concurrency level (C engines in separate processes). Best per column in **bold**. †Serving-style rows: VoxCPM2 through the nanovllm continuous-batching engine (chunk streaming, 16 concurrent sequences, GPU-memory utilization 0.5), and FreyaTTS through a static-batching ODE server (padded batch, one masked 32-step solve; TTFT column reports its per-batch latency at B=8). All other rows are the systems' released library stacks.

System	Params	VRAM (GB)	$\text{TTFT}_{\text{long}}$ (s)	RTF_{med}	RTF_{long}	Throughput
FreyaTTS (ours)	183M	1.5	0.52	**0.11**	0.10	9.4 (C=4)
F5-TTS (tr)	~336M	**0.8**	0.74	0.13	**0.05**	8.4 (C=2)
XTTS-v2 (multi)	~470M	2.2	0.30	0.33	0.33	3.4 (C=2)
Spark-TTS	~0.5B	5.4	14.0	1.08	1.04	1.4 (C=2)
VoxCPM2	~2B	5.5	4.2	0.37	0.37	1.7 (C=1)
VoxCPM2 (serving engine)†	~2B	11.4	**0.14**	0.15	0.13	24.8 (C=16)
FreyaTTS (batched ODE)†	183M	3.8	0.54	–	–	**65.2** (B=8)

4.5 Speaker Consistency and the Voice-Lock

Because FreyaTTS carries no speaker field and no reference-prompt pathway, the pretrained prior has no notion of speaker identity: the Gaussian prior x_0 seeds a fresh, uncontrolled speaker on every generation. Figure 2b makes this concrete. Regenerating a fixed text many times (a 114-generation voice-consistency probe), the pretrained model's per-generation median F_0 spans roughly 70–347 Hz, with a standard deviation of 74.9 Hz and a 60% rate of crossing the 165 Hz male/female pitch boundary between successive draws. Intelligibility metrics are blind to this (the content is correct on each draw), yet for a deployed single-voice product it is a categorical failure: the listener hears a different person each turn.

The voice-lock stage writes a single identity into the weights. Full-parameter fine-tuning on 76.5 hours of the consented target speaker (SFT-v1; section 3.5) collapses the cross-generation F_0 standard deviation from 74.9 Hz to 5.0 Hz and effectively eliminates the gender flips, while the content-similarity metric of section 4.1 reaches 0.923: the model now emits a deterministic, consistent target voice regardless of the noise draw (fig. 2b), measured on sentence-length in-distribution text. A speaker-verification embedding confirms the lock at the identity level: the ECAPA-TDNN (Desplanques et al., 2020) cosine similarity between FreyaTTS syntheses and the target-speaker centroid is 0.873 ± 0.029, within one standard deviation of the 0.883 ± 0.026 same-speaker ceiling measured on disjoint *real* recordings of the speaker, and far above the 0.140 different-speaker floor. The identity is a property of the parameters, not of a prompt: no reference audio is supplied at inference. We select the step-1000 checkpoint because prolonged fine-tuning over-specializes and destabilizes the voice, with the consistency standard deviation drifting back from 5.0 Hz toward 11 Hz by step 3k. SFT also stabilizes pronunciation: all-caps acronyms are spoken natively as words rather than spelled out, and isolated short utterances, covered by SFT-v2's mined one- and two-word segments (section 3.5), are rendered cleanly. This pretrain-to-SFT contrast, an unconditioned multi-speaker prior turned into a locked single voice, is the mechanism by which the foundation-style generative prior becomes a deployable single-voice product.

4.6 Inference Efficiency

FreyaTTS generates 48 kHz audio non-autoregressively: given the duration head's predicted length $\hat{T}$, a single fixed-step Euler integrator runs 32 ODE steps from noise to a clean latent sequence, which the frozen AudioVAE decodes in one pass (section 3.6). The number of *sequential* solver steps is fixed at 32 regardless of $\hat{T}$, there is no left-to-right error accumulation, and a plain deterministic integrator suffices, without classifier-free guidance or sway sampling. At 183.2M parameters the model fits in 1.5 GB of VRAM.

Table 3 measures FreyaTTS against four larger open systems, including the ~2B VoxCPM2 (Zhou et al., 2025) whose AudioVAE we reuse, the autoregressive Spark-TTS (Wang et al., 2025), XTTS-v2 (Casanova et al., 2024), and F5-TTS (Chen et al., 2025b), each run end-to-end through its released inference stack with default sampling on an RTX 4090. Among the released library stacks, FreyaTTS attains the best medium-sentence RTF (0.11) and the highest throughput (9.4 audio-seconds per wall-second with four concurrent engines) at the second-smallest memory footprint. Against the production-class VoxCPM2 it is 3.2× faster in RTF, 8× faster to first audio on long inputs, 3.7× smaller in VRAM, and 5.5× higher in throughput; against the autoregressive Spark-TTS the gaps widen to roughly 10× in RTF and 27× in TTFT. The optimized F5-TTS runtime is the one system in the same speed class: its fused bf16 kernel path reaches a lower long-form RTF, while FreyaTTS, running unoptimized fp32 eager PyTorch, retains lower TTFT, higher throughput, and, per table 2, a three-times-lower Turkish WER with no reference clip at inference. XTTS-v2's native chunk streaming yields the best library-stack TTFT (0.30 s) at roughly 3× the RTF. We additionally measure VoxCPM2 through a dedicated continuous-batching serving engine (nanovllm) rather than its released library stack: chunk streaming brings its TTFT to 0.14 s and batching lifts throughput to 24.8 audio-seconds per wall-second at 16 concurrent sequences, at 11.4 GB of reserved VRAM. A serving engine is thus worth roughly an order of magnitude on both axes for the 2B model. The same treatment favors the non-autoregressive design even more: a static-batching ODE server for FreyaTTS (pad the batch, run one masked 32-step solve, decode, trim; released with our inference tooling) reaches 65.2 audio-seconds per wall-second at batch size 8 with a 0.54 s per-batch latency and 3.8 GB of VRAM, 2.6× the 2B engine's continuous-batching throughput at one-third of its memory, before any length-bucketing or chunk-streaming refinements. On an H100, FreyaTTS's benchmark-set mean RTF is ≈0.14.

The compact size also admits deployment classes the systems of table 3 do not target. On a laptop CPU (Apple M3, four threads, fp32, no quantization) FreyaTTS synthesizes at or faster than real time: RTF 0.70 on medium-length sentences and 0.94–1.00 on the short and long buckets, with a 5.7 s first-clause TTFT on long inputs. Compiled to Core ML at fp16, the DiT's full 32-step loop renders 4.4 s of audio in 0.205 s on the laptop's neural engine (RTF 0.047) and 0.72 s on CPU only (RTF 0.165), with AudioVAE decoding adding 0.33 s on CPU, for an end-to-end real-time factor of ≈0.12 on consumer Apple silicon. A model competitive with the fastest GPU systems of its field while synthesizing faster than real time on a laptop CPU is an operating point that pairs the foundation-style generative prior with genuinely edge-deployable inference.

5 Conclusion and Future Work

In this work, we presented FreyaTTS, a tokenizer-free, non-autoregressive flow-matching model for Turkish conversational speech synthesis. Built as a compact 183.2M-parameter conditional flow-matching Diffusion Transformer inside the *frozen* 25 Hz continuous-latent space of the VoxCPM2 AudioVAE (Zhou et al., 2025), and driven by a 92-symbol character vocabulary with no phonemizer, grapheme-to-phoneme frontend, or discrete speech tokenizer, FreyaTTS is trained from scratch on Turkish speech at a mid-resource budget and denoises an entire utterance in parallel in a fixed 32-step ODE. A two-stage pretrain-to-SFT recipe, a single-speaker voice lock (F_0 standard deviation 74.9 → 5.0 Hz) followed by short-utterance coverage, converts the speaker-agnostic prior into a reliability-hardened single-voice production model, and development surfaced a transferable design lesson: at mid-resource scale, the implicit-alignment recipe of E2/F5-style models (Eskimez et al., 2024; Chen et al., 2025b) collapsed to babble, whereas explicit cross-attention conditioning aligned cleanly on the same data (section 4.4). Evaluated under a band-matched Whisper WER/CER protocol on the released **Freya-TR-Eval** benchmark, FreyaTTS (WER 8.0 %, CER 3.0 %) outperforms both larger open systems, XTTS-v2 and F5-TTS, at a real-time factor of ≈0.14 and roughly one-tenth their parameter count.

While FreyaTTS delivers strong results for its size, several challenges remain. Two compact phonemizer-driven VITS baselines still attain lower error on clean conversational text; digit-dense input requires spoken-form expansion at the text frontend, because the character-level duration predictor under-allocates frames for compact digit strings (section 4.3); long-form synthesis and isolated out-of-distribution tokens continue to rely on inference-layer mitigations rather than being solved in the weights; the shipped voice inherits a narrowband fidelity ceiling from its telephony-band source; and a controlled replication of the alignment finding remains to be run.

Future work will focus on a jointly learned CTC-monotonic aligner, already prototyped, that targets the residual word-skip and long-horizon-drift failures a single global length prediction cannot prevent; preference-based post-training (Rafailov et al., 2023) from intelligibility and voicing rewards; extending the voice-lock recipe from one speaker to multiple voices and natural-language voice design; and explicit bandwidth extension so that a narrowband-trained speaker can be rendered in true wideband. We hope the open release of FreyaTTS, the pretrained Turkish prior, the Freya-TR-Eval benchmark, and the inference tooling provides a solid foundation for Turkish and low-resource speech research.

Contributors

Ahmet Erdem Pamuk, Ömer Yentür, Ahmet Tunga Bayrak, Yavuz Alp Sencer Öztürk, Mustafa Yavuz.

References

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5723–5738, 2022.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, et al. Common voice: A massively-multilingual speech corpus. In *Language Resources and Evaluation Conference (LREC)*, 2020.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*, 2024.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 2025a.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025b.

Alexis Conneau, Min Ma, Simran Khan, Yu Zhang, et al. FLEURS: Few-shot learning evaluation of universal representations of speech. In *IEEE Spoken Language Technology Workshop (SLT)*, 2023.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech*, 2020.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE spoken language technology workshop (SLT)*, pp. 682–689. IEEE, 2024.

Weichen Fan, Amber Yijia Zheng, Dejia Zhu, Yao Ma, Nikola Liu, Zhangyang Wang, and Dilin Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, et al. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890. IEEE, 2024.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, et al. Ditar: Diffusion transformer autoregressive modeling for speech generation. In *International Conference on Machine Learning*, pp. 27255–27270. PMLR, 2025.

Ziyue Jiang, Yi Ren, Ruiqi Li, Shengpeng Ji, Boyang Zhang, Zhenhui Ye, Chen Zhang, Jionghao Bai, Xiaoda Yang, Jialong Zuo, Yu Zhang, Rui Liu, Xiang Yin, and Zhou Zhao. MegaTTS 3: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. *arXiv preprint arXiv:2502.18924*, 2025.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, et al. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *International Conference on Machine Learning (ICML)*, 2024.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning (ICML)*, 2021.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993, 2023.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36:14005–14034, 2023.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2024.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yaron Lipman, Ricky T Q Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations (ICLR)*, 2023.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-TTS: A fast TTS architecture with conditional flow matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, et al. Autoregressive speech synthesis without vector quantization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1287–1300, 2025.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *The Twelfth International Conference on Learning Representations*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-TTS: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning (ICML)*, 2021.

Vineel Pratap, Andros Tjandra, Bowen Shi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, 2023.

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

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2020.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yichong Leng, Lei He, Tao Qin, Jiang Bian, et al. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. In *The Twelfth International Conference on Learning Representations*, 2023.

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

MiniCPM Team, Chaojun Xiao, Yuxuan Li, Xu Han, Yuzhuo Bai, Jie Cai, Haotian Chen, Wentong Chen, Xin Cong, Ganqu Cui, et al. Minicpm4: Ultra-efficient llms on end devices. *arXiv preprint arXiv:2506.07900*, 2025.

Alexander Tong, Kilian Fatras, Nikolay Malkin, et al. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research (TMLR)*, 2024.

Changhan Wang, Anne Wu, and Juan Pino. CoVoST 2 and massively multilingual speech translation. In *Interspeech*, 2021.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, et al. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, et al. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning. *arXiv preprint arXiv:2509.24650*, 2025.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. *arXiv preprint arXiv:2506.13053*, 2025.

Han Zhu, Lingxuan Ye, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhifeng Han, Weiji Zhuang, Long Lin, and Daniel Povey. Omnivoice: Towards omnilingual zero-shot text-to-speech with diffusion language models. *arXiv preprint arXiv:2604.00688*, 2026.