LEARNING PERSISTENT AND EDITABLE VOICE PROFILES WITH GENERATIVE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing controllable text-to-speech systems can design voices or modify reference speakers, but these capabilities are often mediated by implicit voice representations distributed across conditioning and generation paths without exposing a persistent state that can be compiled, recovered or edited. In this paper, we introduce GRAPE-TTS, a framework for **G**enerative **R**etrieval and **A**coustic **P**rofile **E**diting, which represents each voice as a persistent acoustic profile comprising a short sequence of residual-quantization identifiers. A supervised residual vector-quantization codec maps speaker embeddings into profiles, and an assembler converts each profile into acoustic conditioning for a TTS backbone. All voice operations are then acts of generative retrieval over this shared state. Extensive experiments demonstrate that GRAPE-TTS achieves strong performance across voice design, cloning, and editing, while preserving the naturalness and intelligibility of the backbone. These results establish explicit acoustic profiles as a unified and direct representation for controllable speech synthesis.

1 INTRODUCTION

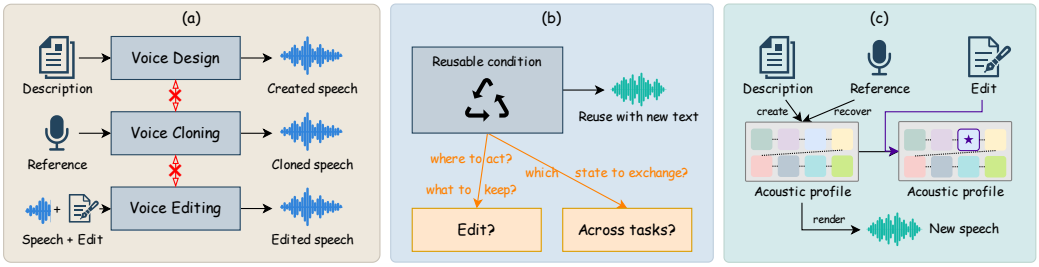

Figure 1: **Motivation for a shared acoustic profile.** (a) Existing voice capabilities are typically realized through task-specific pipelines, leaving the underlying voice object implicit. (b) Consequently, there is no unified way to directly save, reuse, or transfer the operated voice state across tasks. (c) These limitations motivate the acoustic profile: an explicit and reusable representation that serves as a common object for voice creation, recovery, modification, and reuse cross tasks.

Text-to-speech (TTS) systems have evolved from generating speech for a fixed speaker toward modeling the speaker's style and prosodic factors that determine how an utterance is expressed. They increasingly need to support not only the synthesis of an utterance, but also the creation and continued use of a particular voice. Voice cloning transfers a voice from recorded speech, voice design creates a voice from specifications, and voice editing changes selected features of an existing voice. These capabilities are closely related in applications that require a character or personal voice to remain recognizable while speaking new text or undergoing a controlled modification.

Existing systems approach these requirements through several forms of voice conditioning. Speaker embeddings transfer information from speaker verification to multispeaker synthesis (Jia et al., 2018), acoustic prompts and compact global tokens provide conditions for cloning (Wang et al., 2025b; Hu et al., 2026). Description-conditioned systems model the variation specified by a text prompt (Leng et al., 2024; Hai et al., 2026), while editing methods modify speaker representations

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

or regenerate speech under an instruction (Sheng et al., 2025; Yan et al., 2025). Recent frameworks also support multiple voice operations within a common architecture (Hai et al., 2026; Shen et al., 2026). These advances establish strong capabilities in voice matching, instruction following and expressive generation.

However, supporting several capabilities and exposing an editable representation are different design objectives. Although resuable voice conditions already exist (Wang et al., 2025b; Hu et al., 2026), reuse alone does not specify where a requested attribute change should act, or which parts of the condition should remain fixed. A representation learned for speaker discrimination or acoustic reconstruction is not automatically organized into coordinates suitable for selective intervention. Likewise, connecting operations through rendered speech is not the same as allowing them to exchange and modify the same intermediate representation.

This reveals a gap in controllable TTS that a system may successfully produce and modify voices but fail to explicitly determine or migrate voice itself. We study a shared persistent acoustic profile, in which a short sequence of codebook identifiers decides what can be generated from a voice description, recovered from reference speech, and modified before being mapped to synthesis conditioning. Also, the persistent profile can be retained independently of a particular target text and reused. Figure 1 illustrates this distinction between voice capabilities and the object on which they operate.

We introduce **GRAPE-TTS** for **G**enerative **R**etrieval and **A**coustic **P**rofile **E**diting. In this framework, a residual-quantization codec learns compact profiles from speaker embeddings with trait supervision organizing selected levels and smoothing within speakers for reducing assignment variation. A generator predicts vectors that are retrieved against the frozen codebooks, while a reference registrar aligns observed embeddings before profile encoding. Identifier replacement provides an operation for editing. Finally, an assembler maps a profile to an acoustic carrier. The same profile space is shared across operations while cloning, design and editing use differently trained assembler weights. Figure 2 connects these choices to operations on the profile.

Our experiments evaluate multi-attribute voice design, cloning, and targeted editing. The reported design results provide strong attribute matching and instruction compatibility, with candidate selection contributing substantially to the result. Cloning exposes a tradeoff between a common profile and reference fidelity. Editing is evaluated separately for discrete intervention and continuous carrier control. Ablation studies further support the contributions of our architecture.

Our contributions can be summarized as follows:

- **We introduce a state based formulation of controllable TTS.** Motivated by the observation that voice is typically realized through implicit representations distributed across task-specific conditioning and generation approaches, we formulate voice design, cloning and editing as unified operations of a shared and persistent voice state.
- **We design the shared state as a discrete acoustic profile for generative retrieval.** We propose GRAPE-TTS, in which a voice is considered as residual-quantized sequence over embedding space of speakers. A deterministic assembler maps each profile to acoustic conditioning, while profile generation, encoding and profile intervention all operate on the same directly addressable profile.
- **We empirically characterize capabilities of the explicit acoustic profiles.** Experiments across voice design, cloning and editing demonstrate strong attribute control, effective profile recovery from aligned references, and improved identity preservation under targeted profile editing. GRAPE-TTS additionally preserves the intelligibility of the TTS backbone and achieves high naturalness through the voice conversion.

2 Related Work

2.1 Voice Representations

Speaker embeddings provide a compact conditioning signal for multispeaker TTS (Jia et al., 2018; Desplanques et al., 2020). Reference-conditioned synthesis uses acoustic context, including codec representations and mel-spectrogram prompts. F5-TTS and ZipVoice illustrate flow matching syn-

thesis conditioned on reference speech (Chen et al., 2025b; Zhu et al., 2025). These mechanisms allow a target voice to be transferred without training a separate synthesizer for each speaker. Spark-TTS uses global tokens which are fixed in size alongside time-varying semantic tokens to support generation (Wang et al., 2025b). Qwen3-TTS supports voice cloning and voice design (Hu et al., 2026). GRAPE-TTS studies how an identifier sequence can be organized for generation, recovery and a defined intervention with separate mappings to acoustic conditioning.

2.2 Voice Design and Editing

Description-conditioned TTS maps verbal specifications to voice characteristics. PromptTTS 2 explicitly models variation that is not fully determined by a description (Leng et al., 2024). Editing methods differ in their intervention space. VoxEditor uses speakers' representation modification guided by text (Sheng et al., 2025), whereas Step-Audio-EditX supports expressive and iterative audio editing (Yan et al., 2025). VoiceDesigner and FireRedTTS3 further integrate voice generation and editing in unified frameworks (Hai et al., 2026; Shen et al., 2026). Our emphasis is a retained profile that the operations can exchange directly together with an explicit distinction between changing that profile and controlling its acoustic realization.

2.3 Voice Cloning

Voice cloning aims to synthesize new text while preserving the vocal identity of a reference speaker. Zeroshot approaches infer the target voice from reference speech without speaker-specific parameter updates. Encoder methods transfer representations learned for speaker verification to multispeaker synthesis, enabling conditioning on speakers unseen during training (Jia et al., 2018). Acoustic prompting methods instead condition generation on a sequence of reference speech features. VALL-E uses neural codec tokens as acoustic prompts for conditional language modeling (Chen et al., 2025a). GRAPE-TTS builds on speaker representations but formulates cloning as recovery into the same structured profile space used for generation and targeted profile editing. The distinction concerns how the recovered representation is organized and operated on a voice condition stored.

2.4 Structured Quantization

VQ-VAE learns discrete latent representations (van den Oord et al., 2017). SoundStream and EnCodec apply residual vector quantization to acoustic sequences (Zeghidour et al., 2021; Défossez et al., 2023). SpeechTokenizer introduces semantic supervision to organize linguistic and acoustic information across residual levels (Zhang et al., 2024). WavTokenizer instead explores low rate acoustic tokenization with a single quantizer (Ji et al., 2025). In contrast to an utterance-level acoustic token sequence, a GRAPE-TTS profile quantizes a speaker embedding. Trait labels encourage selected levels to contain predictable attribute information, but do not establish complete acoustic disentanglement.

2.5 Generative Retrieval

Generative retrieval predicts structured identifiers instead of retrieving an object merely through a conventional similarity search. TIGER uses residual-quantized semantic identifiers for recommendation (Rajput et al., 2023). GRAPE-TTS adapts this strategy to a voice representation, where a conditional generator samples level vectors, and codebook retrieval produces a valid identifier sequence. Here, the semantics of an identifier concerning structures related to voice. Reference recovery and editing access the same address space without invoking the description generator.

3 Method

3.1 Preliminaries

Acoustic Objects. Let x denote a waveform, s the text to synthesize, and τ description of the desired voice. A frozen ECAPA-TDNN encoder produces a unit embedding $e = E(x) \in \mathbb{R}^{192}$ (Desplanques et al., 2020). We write $\mathcal{N}(v) = v/\|v\|_2$ and $\mathrm{sim}(e, e') = e^\top e'$ for normalized embeddings. The codec C maps e to an acoustic profile $\mathbf{c}$, and its decoder D_C maps a profile back to

reconstructed speaker space. The profile, selected code vectors, and standardized embedding are

$$\mathbf{c} = (c_0, \ldots, c_{Q-1}) \in \mathcal{C} = \{0, \ldots, K-1\}^Q, \quad b_q = B_q[c_q], \quad u = (e - \mu_e) \oslash \sigma_e, \tag{1}$$

where $Q = 8$, $K = 2048$, $B_q \in \mathbb{R}^{K \times 128}$, and the normalization statistics are fixed after training. A profile is independent of the target text, and its inferred value may still vary across reference recordings.

Residual Quantization. For latent $z = f_{\text{enc}}(u)$ and initial residual $r_0 = z$, residual vector quantization (RVQ) gives

$$c_q = \underset{0 \le k < K}{\arg\min} \|r_q - B_q[k]\|_2^2, \qquad r_{q+1} = r_q - B_q[c_q], \qquad \hat{z} = \sum_{q=0}^{Q-1} B_q[c_q]. \tag{2}$$

The decoder reconstructs $\hat{u} = f_{\text{dec}}(\hat{z})$ and $D_C(\mathbf{c}) = \mathcal{N}(\mu_e + \sigma_e \odot \hat{u})$. Residual ordering does not assign a particular attribute to a level on its own.

Profile Operations and Realization. Generation returns $\mathbf{c}_{\text{d}} = \mathcal{R}_B(G_\theta(\tau, \epsilon))$, where G_θ generates continuous level vectors and $\mathcal{R}_B$ retrieves their nearest codebook entries. Reference recovery returns $\mathbf{c}_{\text{r}} = C(R_\psi(e))$, where R_ψ is a normalized residual registrar. An editor S_ω proposes an identifier replacement within $\mathcal{C}$. The assembler configuration $a \in \{\text{main}, \text{clone}\}$ determines a fixed map A_a from profiles to PCA carrier coefficients $X \in \mathbb{R}^{768 \times 64}$. Inverse PCA yields a mel carrier $M_a(\mathbf{c}) \in \mathbb{R}^{768 \times 100}$. Vocos V renders that carrier, while a frozen TTS backbone $\mathcal{T}$ can use its mel representation and a fixed prompt string s_0 to generate new content

$$\hat{x}_{\text{car}} = V(\Phi_\gamma(M_a(\mathbf{c}))), \qquad \hat{y} = \mathcal{T}(s; \Phi_\gamma(M_a(\mathbf{c})), s_0, \xi). \tag{3}$$

The profile is $\mathbf{c}$ here, and Φ_γ applies carrier controls $\gamma = (d_{\text{p}}, \rho)$, with default $\gamma_0 = (0, 1)$, and ξ denotes backbone randomness. Reproducing a controlled rendering requires retaining the assembler configuration and controls as well as the profile.

3.2 Learning Structured Acoustic Profiles

A useful intermediate representation must compress speaker information without leaving its intervention entirely unspecified. We therefore combine reconstruction of speaker embeddings with supervision at selected residual levels. A MLP maps u to a latent, and a mirrored decoder reconstructs u from the sum of retrieved vectors.

Structure for Targeted Intervention. Classification heads on levels 0, 1, 2, and 3 predict gender, age, pitch, and speaking rate category. Levels 4 to 7 are codes of residual details, and have no direct classification target. Missing labels are masked when training. These assignments encourage coordinates associated to attributes.

Stability across References. To reduce assignment changes associated with different recordings of a speaker, we align their soft assignment distributions using symmetric KL divergence. The distributions are obtained from negative squared residual-to-codebook distances, scaled by a detached batch median. This smoothing is distinct from the registrar, and it regularizes the learned codec, whereas the registrar corrects a reference embedding before quantization.

Training. The reconstruction is $\mathcal{L}_{\text{rec}} = 1 - \cos(u, \hat{u}) + \text{MSE}(u, \hat{u})$. With groups g for records, the objective is

$$\mathcal{L}_C = \sum_g w_g \big(\mathcal{L}_{\text{rec},g} + \mathcal{L}_{\text{commit},g} + \mathcal{L}_{\text{trait},g}\big) + \mathcal{L}_{\text{smooth}}, \tag{4}$$

where $(w_{\text{synth}}, w_{\text{utt}}, w_{\text{spk}}) = (0.5, 0.3, 0.2)$. Commitment penalizes the distance to selected detached code vectors, and trait supervision uses masked cross entropy. Codebooks are updated by exponential moving averages. After this stage, the codec and its codebooks are frozen for future operations.

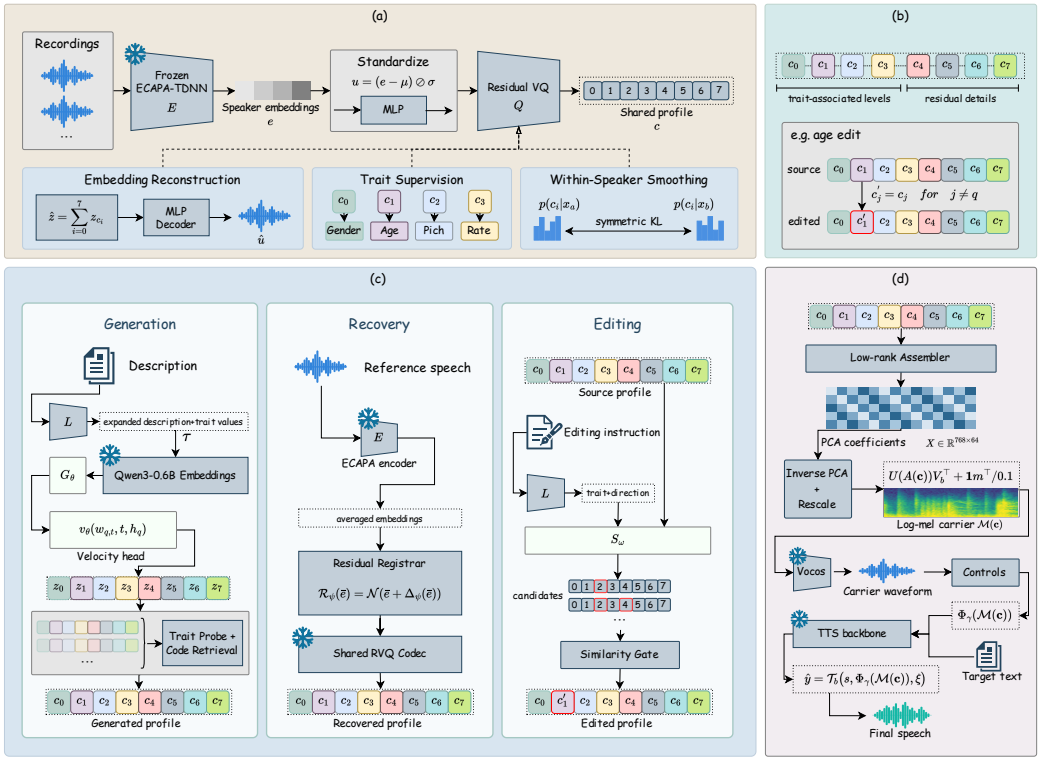

Figure 2: **Core framework of GRAPE-TTS around the shared profile.** (a) A codec learns residual identifiers from frozen speaker embeddings using reconstruction, trait supervision and within-speaker smoothing. (b) Identifiers form the profile, function as core of the profile, first several codes decide trait information, the other following levels retain residual details without direct trait labels. (c) Conditioned continuous generation followed by codebook retrieval, reference registration followed by encoding, and level replacement return the same type of profile. (d) A specific additive assembler reconstructs a mel carrier. Pitch and rate controls modify this carrier without changing identifiers. Vocos renders the carrier directly, while ZipVoice consumes the carrier mel and a fixed prompt string to synthesize target text.

3.3 Voice Operations over Shared Profiles

Conditioned Generation. A description can specify several attributes without uniquely determining a voice. We therefore model a conditional distribution over profiles. Frozen Qwen3-0.6B embeddings (Qwen Team, 2025) provide the description prefix to a causal transformer. A learned null voice prefix is used for generation with description only. With fixed vector normalization $\tilde{b}_q = \mathcal{Z}(b_q)$, the model factorizes

$$p_\theta(\tilde{b}_0, \ldots, \tilde{b}_{Q-1} \mid \tau) = \prod_{q=0}^{Q-1} p_\theta(\tilde{b}_q \mid \tilde{b}_{<q}, \tau). \tag{5}$$

Each context h_q conditions a shared flow matching velocity head (Lipman et al., 2023). During training, preceding target vectors are teacher forced and

$$\begin{aligned} w_{q,t} &= t\epsilon_q + (1-t)\tilde{b}_q, \quad t \sim \mathcal{U}[0,1], \quad \epsilon_q \sim \mathcal{N}(0, I), \\ \mathcal{L}_G &= \mathbb{E}\left[\frac{1}{Q}\sum_q \mathrm{MSE}\Big(v_\theta(w_{q,t}, t, h_q), \epsilon_q - \tilde{b}_q\Big)\right]. \end{aligned} \tag{6}$$

Sampling integrates from noise at $t = 1$ to $t = 0$ with 16 Euler steps per level. A completed *continuous* vector is passed to the next autoregressive position. Only after all eight vectors have

been generated, normalization is inverted and retrieval is performed

$$c_q = \arg\min_{0 \le k < K} \|\mathcal{Z}^{-1}(\tilde{\bar{b}}_q) - B_q[k]\|_2^2. \tag{7}$$

This ordering avoids conflating continuous autoregressive sampling with feeding quantized codes back into the generator.

Candidate Selection. The multi-candidate configuration alternates two generator outputs and proposes $N = 6$ profiles. Requested structured traits are automatically scored on rendered carriers. The design path additionally considers temporal variants for requested fast or slow speech. Selection uses the requested attribute matches, and the retained rendering must include the selected temporal setting as well as the profile.

Reference Recovery. Reference recovery aligns a measured speaker embedding before applying the same codec. For reference segments $x_1, \ldots, x_L$, the final sequence passes through averaging, residual correction, normalization, and quantization

$$\bar{e} = \frac{1}{L}\sum_{\ell=1}^{L} E(x_\ell), \qquad R_\psi(\bar{e}) = \mathcal{N}(\bar{e} + \Delta_\psi(\bar{e})), \qquad \mathbf{c}_{\mathrm{r}} = C(R_\psi(\bar{e})). \tag{8}$$

The registrar is a residual MLP with zero-initialized output. The training procedure pairs degraded and normalized reference embeddings with normalized speaker means e_m, using

$$\mathcal{L}_R = 1 - \mathbb{E}[\cos(\hat{e}, e_m)] + \mathrm{MSE}(\hat{e}, e_m). \tag{9}$$

Code agreement is monitored but does not enter this loss. Augmentation includes cropping, noise, resampling, and gain changes. The cloning path uses A_{clone}, whereas editing a recovered profile uses A_{main}. Comparisons before and after an edit must therefore hold the assembler fixed.

Profile Editing. To change a designated coordinate without regenerating the entire profile, we restrict candidates to

$$\mathcal{C}_q(\mathbf{c}) = \{\mathbf{c}' : c'_j = c_j \text{ for all } j \neq q\}. \tag{10}$$

The editor receives code tokens, level information and an attribute direction condition and predicts replacement logits. Its offline teacher scans the designated codebook and scores decoded candidate embeddings. Search-derived target codes train the target level classifier, while non-target positions receive identity targets

$$\mathcal{L}_S = \mathcal{L}_{\mathrm{target}} + \mathcal{L}_{\mathrm{keep}}. \tag{11}$$

The training set contains operation pairs spanning pitch, rate, and age targets.

3.4 Deterministic Profile Realization

A profile needs fixed interpretation as acoustic conditioning without making its identifiers depend on each new synthesis text. We implement that interpretation through an additive, low-rank assembler and explicitly separate it from controls that modify the realized carrier.

Assembly. Let $D_X = 768 \cdot 64$, rank $r_A = 128$, $F_{a,q} \in \mathbb{R}^{D_X \times r_A}$, and $H_{a,q} \in \mathbb{R}^{K \times r_A}$. The assembler first predicts normalized coefficients $\bar{X} = \bar{A}_a(\mathbf{c})$

$$\mathrm{vec}(\bar{A}_a(\mathbf{c})) = x_{a,\mathrm{base}} + \sum_q F_{a,q}\big(H_{a,q}[c_q] - \bar{H}_{a,q}\big), \quad \bar{H}_{a,q} = \frac{1}{K}\sum_k H_{a,q}[k]. \tag{12}$$

The actual coefficient output is $A_a(\mathbf{c}) = \sigma_a \odot \bar{A}_a(\mathbf{c}) + \mu_a$. For profiles differing only at q, the normalized change is

$$\mathrm{vec}(\bar{A}_a(\mathbf{c}') - \bar{A}_a(\mathbf{c})) = F_{a,q}\big(H_{a,q}[c'_q] - H_{a,q}[c_q]\big). \tag{13}$$

Denormalization further scales this change by σ_a. This algebraic locality explains the compatibility of additive assembly with code replacement, and it does not establish trait disentanglement after vocoding or synthesis.

Carrier. With fixed PCA basis $V_b \in \mathbb{R}^{100\times 64}$ and mean m, the carrier is

$$M_a(\mathbf{c}) = \frac{A_a(\mathbf{c})V_b^\top + \mathbf{1}m^\top}{0.1}. \tag{14}$$

Vocos renders the carrier directly (Siuzdak, 2024). In the downstream ZipVoice, its mel representation is supplied directly as the prompt feature sequence with the fixed string s_0, and there is no required carrier-waveform re-encoding between the assembler and ZipVoice (Zhu et al., 2025). The requested text s controls the generated segment. Determinism here concerns the profile-to-carrier map for fixed parameters and controls.

Training. Teacher waveforms are converted to 100-bin mel sequences, cropped or padded to 768 frames, and projected to PCA coordinates. Training combines normalized coefficient MSE, coefficient cosine distance, factor regularization, a mismatched-profile penalty, and similarity objectives on both direct carriers and a ZipVoice-rendered branch

$$\mathcal{L}_A = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\cos} + \Omega_A + \mathcal{L}_{\text{cons}} + \mathcal{L}_{\text{sim,car}} + \lambda_f(\nu)\mathcal{L}_{\text{sim,flow}}. \tag{15}$$

The main and cloning assemblers have the same architecture but different training distributions, with the cloning model warm-started and adapted to real speech records.

4 Experiments

4.1 Experimental Setup

Environment. GRAPE-TTS is implemented in PyTorch, with inference on one 48 GB NVIDIA RTX Pro 5000 Blackwell GPU. Each profile contains eight residual identifiers with 2,048 entries per codebook, and generation uses 16 Euler steps per level. Detailed training and inference settings are given in Appendix A.

Data. Training combines real speaker representations, 80,415 synthetic speech samples from Qwen3-TTS (Hu et al., 2026), and profile-carrier supervision. Emilia English speech (He et al., 2024) additionally supports the cloning assembler. Evaluation includes 1,724 design cases, 1,200 selected InstructTTSEval cases per system (Huang et al., 2025), 1,067 cloning cases from ELDERLY67, Seed-zh, and Seed-en, and 200 editing speakers. The instruction set contains 400 cases each for acoustic parameter specification (APS), descriptive style directive (DSD), and role play (RP). Editing uses distinct enrollment and scoring recordings. The multi-text study additionally renders five fixed texts for each of these 200 speakers. Data construction and training exposure are specified in Appendix B.

Baselines. We compare with six baselines for design and instruction following, fifteen for cloning, and two baselines plus one signal processing reference for editing. System assignments are listed in Appendix C. Baselines use one candidate per input due to their architectures. Our main design configuration uses six profile candidates. The budget study keeps the policy fixed, changing only the allowed budget.

Metrics. Design uses gender, pitch, rate and age accuracy averaged over the applicable subsets. The same probes support automated selection and attribute scoring. Instruction following uses valid response compatibility under a local Qwen2.5-Omni-7B judge (Xu et al., 2025) with the InstructTTSEval. Speaker similarity (SIM) is normalized ECAPA cosine similarity. Editing uses similarity to the unedited system output (id_{self}). Pitch and rate use linear relative difference error, $\text{RDE} = |\widehat{r} - r^\star|/r^\star$. Age uses hit rate and the change in target probability, ΔP_{ya}. UTMOS estimates naturalness (Saeki et al., 2022).

4.2 Experimental Results

Voice Design and Instruction Following. We test whether conditioned generation and codebook retrieval produce profiles with the requested voice characteristics. Table 1 shows the highest reported accuracy on all four attributes, with gains of 15.6 and 33.4 percentage points over the strongest

Table 1: Comparison of voice design and instruction following results (%).

	Voice design				Instruction following			
System	Gender	Pitch	Rate	Age	APS	DSD	RP	Avg.
Qwen3-TTS-vd	95.8	54.1	62.0	36.9	99.7	85.3	69.3	84.8
FireRedTTS3	95.8	49.5	44.4	33.0	99.7	83.0	66.9	83.2
Spark-TTS	94.8	79.3	71.7	41.1	**100.0**	81.2	59.4	80.2
Higgs Audio v2	77.8	34.0	42.4	36.4	99.7	79.8	59.3	79.6
Higgs Audio v3	57.4	77.3	36.1	44.2	99.5	76.9	60.9	79.1
Parler-TTS	49.3	33.7	37.7	41.5	**100.0**	64.3	50.0	71.4
Ours	**96.6**	**82.9**	**87.3**	**77.6**	**100.0**	**91.7**	**85.3**	**92.3**

Table 2: Comparison of voice cloning.

System	ELDERLY67	Seed-zh	Seed-en	Macro	UTMOS
F5-TTS	0.446	**0.758**	0.658	**0.621**	3.074
IndexTTS2	0.465	0.726	0.610	0.600	3.049
Qwen3-TTS	0.368	0.724	0.599	0.564	3.720
Higgs Audio v3	0.355	0.657	0.563	0.525	3.550
CosyVoice3	0.443	0.736	0.581	0.586	3.471
Chatterbox	0.363	0.519	0.605	0.496	3.113
XTTS-v2	0.356	0.627	0.492	0.492	3.103
MOSS-TTSD	0.440	0.727	0.626	0.598	3.371
VoxCPM	0.416	0.698	0.622	0.579	3.144
FireRedTTS3	0.463	0.659	0.587	0.570	3.344
ZipVoice(backbone)	0.409	0.750	**0.666**	0.608	3.222
Ours	**0.580**	0.611	0.561	0.584	3.184

baseline for rate and age respectively. Instruction compatibility averages 92.3%, with the largest differences on DSD and RP. These results support control through generated profiles under the reported selection and rendering configuration. The ablation below separates the contribution of candidate budget from the single-proposal result.

Voice Cloning. We test recovery of a reference voice into the same profile space used for design. GRAPE-TTS obtains 0.580 on ELDERLY67, above the strongest baseline. These results demonstrate reference-based use of the common cloning task while exposing a fidelity gap for the decrease from backbone to our method in Seed-zh and Seed-en. Appendix D.2 provides results of all fifteen baselines.

Voice Editing. We test targeted modification at age identifier replacement and pitch/rate control of the realized carrier. Age editing attains a 29.5% hit rate and $\mathrm{id}_{\mathrm{self}} = 0.8848$, showing substantial retention of the system's unedited voice while modifying the designated coordinate (Table 3). Carrier control yields rate RDE 0.001 and pitch RDE 0.025. The ablation below tests how candidate validation affects both age attainment and preservation.

Profile Reuse. We additionally synthesize new texts from retained profiles. Similarity to recordings averages 0.72 for unedited profiles and 0.63 for edited profiles. These scores establish voice reuse and indicate that editing has no major effect destroying identity.

Table 3: Comparison of voice editing results.

Model	$\mathrm{id}_{\mathrm{self}} \uparrow$	Pitch RDE $\downarrow$	Rate RDE $\downarrow$	Age hit (%) $\uparrow$
StepAudioEditX	0.717	–	0.198	23.5
FireRedTTS-3	0.664	**0.022**	0.28	–
DSP reference	0.684	0.025	≈ 0	–
Ours	**0.885**	0.025	**0.001**	**29.5**

Table 4: Ablation comparisons.

(a) Candidate budget for generative retrieval

Configuration	Gender	Pitch	Rate	Age
One candidate	93.2	57.0	65.6	56.0
Six candidates	96.6	82.9	87.3	77.6

(b) Joint identity gating for age intervention

Configuration	Age hit (%)	ΔP_{ya}	$\mathrm{id}_{\mathrm{self}}$	n
Both gates enabled	29.5	$+0.0685$	0.8848	200
Both gates disabled	0.5	-0.0442	0.0994	200

4.3 Ablation Study

Selecting Profiles. Holding the settings fixed, increasing the proposal budget from one to six candidates improves gender, pitch, rate, and age accuracy by 3.4, 25.9, 21.7, and 21.6 percentage points (Table 4a). This supports candidate routing as a major contributor to the attribute control.

Reference Alignment. In the earlier paired registrar comparison, removal reduces macro cloning SIM from 0.584 to 0.499, with the largest reduction on Seed-en. This supports correcting the reference embedding before quantization.

Profile Editing. We disable both the decoded space and rendered space identity gates while retaining level-1 search on the same 200 source cases. Age hit falls from 29.5% to 0.5%, and $\mathrm{id}_{\mathrm{self}}$ from 0.885 to 0.0994 (Table 4b). These results support identity-constrained candidate validation as part of successful profile editing.

5 Conclusion

We presented GRAPE-TTS, a framework which makes a residual-quantized acoustic profile the common representation. Trait supervision organizes selected profile levels, and an additive assembler gives each profile a fixed interpretation as acoustic conditioning under a specified configuration. Generative retrieval constructs profiles from descriptions, while reference enrollment recovers the same representation from speech. Profile interventions and continuous carrier controls support targeted attribute editing. A deterministic assembler connects these operations to a frozen TTS backbone, separating voice operations and speech realization. Experiments demonstrate improved voice design and instruction following capability, retained cloning performance, and improved identity consistency relative to unedited speech. These results support acoustic profiles as a shared path for controllable speech synthesis. However, there are still limits including reduced fidelity in zeroshot cloning, backbone dependent tradeoffs between naturalness and limited identity preservation. Future work will prioritize real variation within speakers, and carrier-aware assembler targets. Hybrid profiles with a protected identity residual channel offer a direction for improving source preservation. These measured boundaries provide an actionable basis for engineering more faithful persistent voice control.

AI USE STATEMENT

In this work, Qwen3-TTS (Hu et al., 2026) was used to synthesize speech samples for model pretraining. Qwen2.5-Omni-7B (Xu et al., 2025) was used solely as a predefined evaluator as described in the experimental section. ChatGPT (OpenAI et al., 2024) was used to polish the writing of manuscript and to organize and refactor code written entirely by the authors. We did not use generative AI tools for literature retrieval or discovery, research ideation, experimental design or execution, drafting substantive sections of the paper, or proving mathematical claims. Generative AI was not used to develop the proposed method, generate the experimental implementation, or produce the scientific conclusions. All AI assisted outputs were reviewed by the authors. AI assisted writing was manually reviewed and edited, and AI assisted code was checked against the original code. We take full responsibility for the content of the paper.

ETHICS STATEMENT

This work follows the principles of the ICLR Code of Ethics. We use publicly available speech datasets and synthetic speech generated with publicly available models. All datasets and models are used in accordance with their respective licenses and usage conditions.

REPRODUCIBILITY STATEMENT

This work complies with the ICLR standards of reproducibility. We provide detailed descriptions of the model architecture, training, inference, dataset and evaluation in the main text and appendix. Pretrained components, baselines and metric implementations are documented to support independent reproduction and verification of the reported results. The anonymized training and evaluating codes are available at `https://anonymous.4open.science/r/GRAPE-TTS-D502`

REFERENCES

Zhiqi Ai, Meixuan Bao, Zhiyong Chen, Zhi Yang, Xinnuo Li, and Shugong Xu. VoxAging: Continuously Tracking Speaker Aging with a Large-Scale Longitudinal Dataset in English and Mandarin. In *Interspeech 2025*, pp. 3648–3652, 2025. doi: 10.21437/Interspeech.2025-57.

Boson AI. Higgs Audio V2: Redefining Expressiveness in Audio Generation. `https://github.com/boson-ai/higgs-audio`, 2025. GitHub repository. Release blog available at `https://www.boson.ai/blog/higgs-audio-v2`.

Boson AI. Higgs audio v3 tts: Conversational speech for voice ai from boson ai. https://huggingface.co/bosonai/higgs-audio-v3-tts-4b, 2026.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In *Interspeech 2024*, pp. 4978–4982, 2024. doi: 10.21437/Interspeech.2024-2016.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718, 2025a. doi: 10.1109/TASLPRO.2025.3530270.

Yang Chen, Hui Wang, shiyao wang, Junyang Chen, Jiabei He, Jiaming Zhou, Xi Yang, Yequan Wang, Yonghua Lin, and Yong Qin. Seniortalk: A chinese conversation dataset with rich annotations for super-aged seniors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL `https://openreview.net/forum?id=QzWPUEuMKU`.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In

Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.313. URL https://aclanthology.org/2025.acl-long.313/.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. VoxCeleb2: Deep Speaker Recognition. In *Interspeech 2018*, pp. 1086–1090, 2018. doi: 10.21437/Interspeech.2018-1929.

Gabriel Clark, Sofian Mejjoute, Mohamed Osman, George Close, and Beren Millidge. Zonos2 technical report, 2026. URL https://arxiv.org/abs/2606.24320.

Alain de Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 04 2002. ISSN 0001-4966. doi: 10.1121/1.1458024. URL https://doi.org/10.1121/1.1458024.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=ivCd8z8zR2. Featured Certification, Reproducibility Certification.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Interspeech 2020*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models, 2024. URL https://arxiv.org/abs/2412.10117.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025. URL https://arxiv.org/abs/2505.17589.

Jiarui Hai, Karan Thakkar, Ke Chen, Yunyun Wang, Jiaqi Su, Rithesh Kumar, Mounya Elhilali, and Zeyu Jin. Voicedesigner: Text-to-voice generation and editing via unified diffusion modeling and data augmentation, 2026. URL https://arxiv.org/abs/2608.13613.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 885–890, 2024. doi: 10.1109/SLT61566.2024.10832365.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, Xinyu Zhang, Pei Zhang, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-tts technical report, 2026. URL https://arxiv.org/abs/2601.15621.

Kexin Huang, Qian Tu, Liwei Fan, Chenchen Yang, Dong Zhang, Shimin Li, Zhaoye Fei, Qinyuan Cheng, and Xipeng Qiu. Instructttseval: Benchmarking complex natural-language instruction following in text-to-speech systems, 2025. URL https://arxiv.org/abs/2506.16381.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, zehan wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, YIDI JIANG, Qian Chen, Siqi Zheng, and Zhou Zhao. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 93809–93826, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ea1f5f0878d43ff4fb8bf64ef4a2326c-Paper-Conference.pdf.

Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 4485–4495. Curran Associates Inc., 2018.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. Parler-tts. https://github.com/huggingface/parler-tts, 2024.

Yichong Leng, ZHifang Guo, Kai Shen, Zeqian Ju, Xu Tan, Eric Liu, Yufei Liu, Dongchao Yang, leying zhang, Kaitao Song, Lei He, Xiangyang Li, sheng zhao, Tao Qin, and Jiang Bian. Prompttts 2: Describing and generating voices with text prompt. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 57672–57688, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fed1ea8dcc2a13f3835cc854e8c8294c-Paper-Conference.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Hulikal Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Tran, Jonah Samost, Maciej Kula, Ed Chi, and Maheswaran Sathiamoorthy. Recommender systems with generative retrieval. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10299–10315. Curran Associates, Inc., 2023. doi: 10.52202/075280-0452. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/20dcab0f14046a5c6b02b61da9f13229-Paper-Conference.pdf.

Resemble AI. Chatterbox-TTS. https://github.com/resemble-ai/chatterbox, 2025. GitHub repository.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022, 2022. URL https://arxiv.org/abs/2204.02152.

Feiyu Shen, Kun Xie, Yichen Wu, Ziqi Dai, Yichen Han, Junjie Li, Xuelong Geng, Fenglong Xie, Lei Xie, Xu Tang, and Yao Hu. Fireredtts3: Unified speech generation and editing with semantically enriched speech representations, 2026. URL https://arxiv.org/abs/2608.17492.

Zheng-Yan Sheng, Li-Juan Liu, Yang Ai, Jia Pan, and Zhen-Hua Ling. Voice attribute editing with text prompt. *IEEE Transactions on Audio, Speech and Language Processing*, 33:1641–1652, 2025. doi: 10.1109/TASLPRO.2025.3557193.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Interspeech 2021*, pp. 2756–2760, 2021. doi: 10.21437/Interspeech.2021-755.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 25719–25733, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6db0903efdfe9b1bbafb015c10990b78-Paper-Conference.pdf.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6309–6318, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Hui Wang, Jiaming Zhou, Jiabei He, Haoqin Sun, and Yong Qin. Wildelder: A chinese elderly speech dataset from the wild with fine-grained manual annotations. *arXiv preprint arXiv:2510.09344*, 2025a.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens, 2025b. URL https://arxiv.org/abs/2503.01710.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Chao Yan, Boyong Wu, Peng Yang, Pengfei Tan, Guoqiang Hu, Yuxin Zhang, Xiangyu, Zhang, Fei Tian, Xuerui Yang, Xiangyu Zhang, Daxin Jiang, and Gang Yu. Step-audio-editx technical report, 2025. URL https://arxiv.org/abs/2511.03601.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 30:495–507, November 2021. ISSN 2329-9290. doi: 10.1109/TASLP.2021.3129994. URL https://doi.org/10.1109/TASLP.2021.3129994.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 31798–31818, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/86d1ab582afb247ccaa84bec4a7e24f7-Paper-Conference.pdf.

Yuqian Zhang, Donghua Yu, Zhengyuan Lin, Botian Jiang, Mingshu Chen, Yaozhou Jiang, Yiwei Zhao, Yiyang Zhang, Yucheng Yuan, Hanfu Chen, Kexin Huang, Jun Zhan, Cheng Chang, Zhaoye Fei, Shimin Li, Xiaogui Yang, Qinyuan Cheng, and Xipeng Qiu. Moss-ttsd: Text to spoken dialogue generation, 2026. URL https://arxiv.org/abs/2603.19739.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*, 2025a.

Yixuan Zhou, Guoyang Zeng, Xin Liu, Xiang Li, Renjie Yu, Ziyang Wang, Runchuan Ye, Weiyue Sun, Jiancheng Gui, Kehan Li, Zhiyong Wu, and Zhiyuan Liu. Voxcpm: Tokenizer-free tts for context-aware speech generation and true-to-life voice cloning, 2025b. URL https://arxiv.org/abs/2509.24650.

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. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025. doi: 10.1109/ASRU65441.2025.11433841.

A Additional Method and Implementation Details

A.1 Shared Representation

Let x denote speech, s the text to synthesize, and τ a voice specification. The frozen ECAPA-TDNN encoder supplies unit speaker embeddings $e = E(x) \in \mathbb{R}^{192}$ (Desplanques et al., 2020). We use

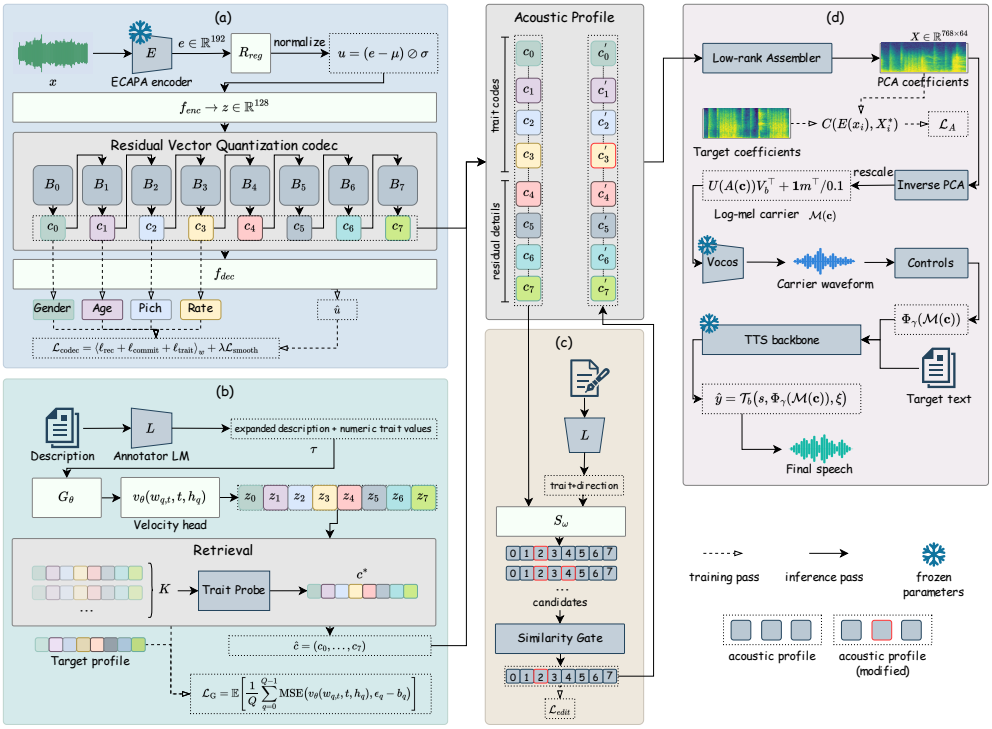

Figure 3: **Overall training and inference pass in GRAPE-TTS.** (a) A frozen speaker encoder E extracts continuous embeddings. The registrar R_{reg} learns alignment to speaker means, and the RVQ codec Q learns an acoustic profile $\mathbf{c} = (c_0, \ldots, c_7)$. (b) The autoregressive generator G learns to generate profiles from voice specifications through flow matching with codec targets. At inference, candidate profiles are selected using trait probes and an identity consistency proxy. (c) The trait editor S predicts target code replacements. Refinement produces a stored profile $\mathbf{c}'$, with remaining identifiers held unchanged if not being targets of modification. (d) A learned low rank assembler A maps the profile to carrier PCA coefficients. Inverse PCA and a frozen vocoder yield an acoustic carrier, with optional proxy backbone model for speech synthesis.

$\mathcal{N}(v) = v/\|v\|_2$. The codec C maps an embedding to

$$\mathbf{c} = C(e) = (c_0, \ldots, c_7) \in \mathcal{C}, \qquad \mathcal{C} = \{0, \ldots, 2047\}^8, \qquad b_q = B_q[c_q] \in \mathbb{R}^{128}. \tag{16}$$

The profile is independent of the requested synthesis text, but its recovery can vary across reference recordings. Table 5 distinguishes the objects.

All operations share the codec and codebook address space. We use A_{main} for design/editing and A_{clone} for cloning. They have the same architecture, but different training distributions and potentially different coefficient normalizers. These correspond to A_{de} and A_{cl} in earlier notation.

A.2 Trait-Supervised Residual Quantization

Normalization. Fixed embedding statistics define

$$u = (e - \mu_e) \oslash \sigma_e, \qquad z = f_{\text{enc}}(u), \qquad \hat{u} = f_{\text{dec}}(\hat{z}). \tag{17}$$

The encoder has hidden width 256 and output width 128 and the decoder mirrors this mapping. Starting with $r_0 = z$, the eight residual assignments are

$$\begin{aligned} c_q &= \operatorname*{arg\,min}_{0 \le k < 2048} \|r_q - B_q[k]\|_2^2, \quad b_q = B_q[c_q], \quad r_{q+1} = r_q - b_q, \\ \hat{z} &= \sum_{q=0}^{7} b_q, \qquad D_C(\mathbf{c}) = \mathcal{N}(\mu_e + \sigma_e \odot \hat{u}). \end{aligned} \tag{18}$$

This supplies a compact encoding of the speaker embedding and the vector targets for generation.

Table 5: Principal mathematical objects.

Object	Dimension / range	Role
e, u	$\mathbb{R}^{192}$	Unit speaker embedding and standardized embedding
z, b_q	$\mathbb{R}^{128}$	Codec latent and selected level vector
$\mathbf{c}$	$\{0,\dots,2047\}^8$	Shared acoustic profile
h_q	$\mathbb{R}^{512}$	Causal context for vector generation
$\bar{X}$, X	$\mathbb{R}^{768\times 64}$	Normalized prediction and denormalized PCA coefficients
$M_a(\mathbf{c})$	$\mathbb{R}^{768\times 100}$	Reconstructed log-mel carrier
a	main / clone	Assembler checkpoint and its normalizer
γ	(d_{p}, ρ)	Mel-bin and temporal factor
s, s_0	Text strings	Requested content and fixed prompt convention

Codebook Maintenance. With stop-gradient operator sg, training uses

$$b_q^{\mathrm{ST}} = r_q + \mathrm{sg}(b_q - r_q), \qquad r_{q+1} = r_q - \mathrm{sg}(b_q^{\mathrm{ST}}), \qquad \ell_{\mathrm{commit}} = \sum_q \mathrm{MSE}(r_q, \mathrm{sg}(b_q)). \tag{19}$$

Codebooks use EMA decay 0.99. Assignment counts are Laplace-smoothed before forming the residual averages. Entries with tracked usage below 2 in a 1,000-step maintenance window are restarted from a sampled batch residual. These updates are disabled during the paired smoothing pass. The configuration specifies the deployed 2,048-entry books.

Attribute Levels. Table 6 gives the supervision allocation. Each head H_q^{trait} is $128 \to 64 \to C_q$ with ReLU, acting on b_q^{ST}. Age classes follow the implementation order child, elderly, middle-aged, teen, and young adult. Missing labels of data are masked.

Table 6: Levels of trait supervision.

Level	Attribute	Classes	Label source
0	Gender	2	Corpus metadata / synthetic annotation
1	Age	5	Corpus metadata / synthetic annotation
2	Pitch	7	Synthetic ordinal instruction labels
3	Speaking rate	7	Synthetic ordinal instruction labels
4-7	Residual detail	–	No direct classification target

For labeled examples $\mathcal{I}_{g,q}$ at level q in group g,

$$\mathcal{L}_{\mathrm{trait},g} = \sum_{q=0}^{3} \frac{1}{|\mathcal{I}_{g,q}|} \sum_{i\in\mathcal{I}_{g,q}} \mathrm{CE}(H_q^{\mathrm{trait}}(b_{i,q}^{\mathrm{ST}}), y_{i,q}), \tag{20}$$

where an empty labeled set contributes zero.

Grouped Reconstruction. Each step draws independent batches of 512 synthetic, 512 real, and 512 real-speaker-mean representations. For each group,

$$\mathcal{L}_{\mathrm{rec},g} = \frac{1}{512} \sum_{i\in\mathcal{B}_g} \left[1 - \cos(u_i, \hat{u}_i) + 0.5\,\mathrm{MSE}(u_i, \hat{u}_i)\right]. \tag{21}$$

With $(w_{\mathrm{synth}}, w_{\mathrm{utt}}, w_{\mathrm{spk}}) = (0.5, 0.3, 0.2)$,

$$\mathcal{L}_C = \sum_g w_g \left(\mathcal{L}_{\mathrm{rec},g} + 0.25\mathcal{L}_{\mathrm{commit},g} + 0.05\mathcal{L}_{\mathrm{trait},g}\right) + 0.1\mathcal{L}_{\mathrm{smooth}}. \tag{22}$$

Reconstruction concerns the standardized speaker embedding. The codec is fixed before training its dependent generation, editing, and assembly paths.

Within-Speaker Smoothing. A separate batch contains 256 pairs consisting of 128 explicit clean/degraded pairs and 128 randomly paired recordings from the same speaker. For $a \in \{1, 2\}$ and residual level q, let

$$\begin{aligned} D^{(a)}_{i,q,k} &= \|r^{(a)}_{i,q} - B_q[k]\|_2^2, \\ s_q &= \mathrm{sg}\Big(\mathrm{med}_{i,a,k}\, D^{(a)}_{i,q,k}\Big), \qquad p^{(a)}_{i,q} = \mathrm{softmax}(-D^{(a)}_{i,q}/s_q), \\ \mathcal{L}_{\mathrm{smooth}} &= \frac{1}{256Q} \sum_{i=1}^{256} \sum_{q=0}^{Q-1} \mathcal{D}_{\mathrm{sym}}(p^{(1)}_{i,q}, p^{(2)}_{i,q}). \end{aligned} \tag{23}$$

$\mathcal{D}_{\mathrm{sym}}$ denotes the implementation's symmetric KL penalty formed from both directional divergences.

A.3 Autoregressive Generation and Codebook Retrieval

Description Representation. The frozen Qwen3-0.6B encoder supplies hidden states $H(\tau) \in \mathbb{R}^{L\times 1024}$ (Qwen Team, 2025). With attention mask m_i, pooling and prefix projection are

$$d_\tau = \frac{\sum_i m_i H_i(\tau)}{\sum_i m_i}, \qquad D_\tau = \mathrm{reshape}_{4\times 512}(W_{\mathrm{d}} d_\tau + b_{\mathrm{d}}). \tag{24}$$

The description extractor uses maximum length 384 without an additional prompt template. A source-prefix projects eight 128-dimensional code vectors to width 512 with learned embeddings. Description-only generation uses a learned null prefix of the same length. During training, the description prefix is zeroed with probability 0.1.

Causal Sequence. The decoder-only transformer processes a begin-of-voice token, eight voice-prefix tokens, and eight prediction positions. Let $\tilde{b}_q = \mathcal{Z}(b_q)$ use the generator's fixed vector normalizer, with $\tilde{b}_{-1} = 0$. Prediction inputs and contexts are

$$p_q = \pi_q + f_{\mathrm{in}}(\tilde{b}_{q-1}), \qquad (h_0, \ldots, h_7) = \mathrm{Last}_8\left(T_\theta[D_\tau, \mathrm{BOV}, \mathbf{v}, \mathbf{p}]\right). \tag{25}$$

π_q is a learned query. Sampling feeds previous completed continuous samples. The factorization is

$$p_\theta(\tilde{b}_0, \ldots, \tilde{b}_7 \mid \tau, \mathbf{v}) = \prod_{q=0}^{7} p_\theta(\tilde{b}_q \mid \tilde{b}_{<q}, \tau, \mathbf{v}). \tag{26}$$

Per-Level Flow Matching. A shared velocity head conditions on the level context h_q, time t, and noisy vector. Conditional flow matching (Lipman et al., 2023) uses

$$\begin{aligned} w_{q,t} &= t\epsilon_q + (1-t)\tilde{b}_q, \qquad t \sim \mathcal{U}[0,1], \quad \epsilon_q \sim \mathcal{N}(0, I), \\ \mathcal{L}_G &= \mathbb{E}\left[\frac{1}{Q}\sum_{q=0}^{Q-1} \mathrm{MSE}\big(v_\theta(w_{q,t}, t, h_q), \epsilon_q - \tilde{b}_q\big)\right]. \end{aligned} \tag{27}$$

The targets are frozen codec vectors in the generator's normalized space. Sampling integrates from $t = 1$ to 0 with 16 Euler steps

$$w_{q,t_{j+1}} = w_{q,t_j} - \frac{1}{16}\widetilde{v}_\theta(w_{q,t_j}, t_j, h_q), \qquad t_j = 1 - j/16. \tag{28}$$

The unconditional context zeroes description tokens while retaining the prefix and history. Only after all continuous vectors have been generated are they inverse-normalized and mapped to identifiers

$$c_q = \mathcal{R}_q(\bar{b}_q) = \underset{0\le k<2048}{\arg\min} \|\bar{b}_q - B_q[k]\|_2^2, \qquad \bar{b}_q = \mathcal{Z}^{-1}(\bar{\tilde{b}}_q). \tag{29}$$

Thus, discrete retrieval does not feed quantized vectors back into the continuous autoregressive sampler. The resulting address can combine entries without identifying an existing stored recording (Rajput et al., 2023).

Algorithm 1 Continuous autoregressive sampling followed by codebook retrieval

Require: Description τ, generator G_θ, normalizer $\mathcal{Z}$, codebooks B
Ensure: Profile $\mathbf{c}$

Encode τ, initialize the null voice prefix and empty continuous history
for $q = 0, \dots, 7$ **do**
  Form conditional context h_q^{c} and zero-description context h_q^{u}
  $w \leftarrow \epsilon_q$, with $\epsilon_q \sim \mathcal{N}(0, I_{128})$
  for $j = 0, \dots, 15$ **do**
    $t \leftarrow 1 - j/16$; compute the conditional velocity
    Apply scale-2.5 guidance for $0.2 \le t \le 0.8$, with rescale 0.7
    $w \leftarrow w - \widetilde{v}_\theta(w, t, h_q)/16$
  end for
  Append w to the continuous history
end for
Apply $\mathcal{Z}^{-1}$ to the complete vector sequence
for $q = 0, \dots, 7$ **do**
  $c_q \leftarrow \arg\min_k \|\bar{b}_q - B_q[k]\|_2^2$
end for
return $(c_0, \dots, c_7)$

Candidate Generation and Selection. The architecture considers six profile proposals with the same codebook space. With structured requested labels, candidates are automatically scored by their number of matching attributes

$$n^\star = \arg\max_{n \le N} \sum_{a \in \mathcal{A}_\tau} \mathbf{1}[\widehat{\ell}_a(\hat{x}_n) = \ell_a^\tau]. \tag{30}$$

The probe path also evaluates temporal factors $\{1.0, 1.2, 1.4\}$ for fast requests, $\{1.0, 0.85, 0.70\}$ for slow requests, and $\{1.0\}$ for middle-rate requests. The original sampling permits early termination on complete attribute agreement.

A.4 Reference Registration

For reference segments $x_1, \dots, x_L$, the ordering is

$$\bar{e} = \frac{1}{L} \sum_{\ell=1}^{L} E(x_\ell), \qquad R_\psi(\bar{e}) = \mathcal{N}(\bar{e} + \Delta_\psi(\bar{e})), \qquad \mathbf{c}_{\mathrm{clone}} = C(R_\psi(\bar{e})). \tag{31}$$

Reference waveforms are resampled to 16 kHz for ECAPA. The residual MLP operates in 192-dimensional space with hidden width 1,024 and three residual layers. Zero-initializing the output starts from an identity correction. Codec standardization uses its stored μ_e, σ_e after the registrar's final unit normalization.

Training and Augmentation. Training uses normalized augmented embeddings e_a, normalized speaker means e_m, and $\hat{e} = \mathcal{N}(e_a + \Delta_\psi(e_a))$, minimizing

$$\mathcal{L}_R = 1 - \mathbb{E}_b[\cos(\hat{e}_b, e_{m,b})] + 2\,\mathrm{MSE}(\hat{e}, e_m). \tag{32}$$

A.5 Carrier Realization

Low-Rank Parameterization. Let $D_X = 768 \cdot 64$, rank $r_A = 128$, $F_{a,q} \in \mathbb{R}^{D_X \times r_A}$, and $H_{a,q} \in \mathbb{R}^{2048 \times r_A}$. The centered code tables produce normalized coefficients $\bar{A}_a(\mathbf{c})$:

$$\begin{aligned} \bar{H}_{a,q} &= \frac{1}{2048} \sum_{k=0}^{2047} H_{a,q}[k], \\ \mathrm{vec}(\bar{A}_a(\mathbf{c})) &= x_{a,\mathrm{base}} + \sum_{q=0}^{7} F_{a,q}\big(H_{a,q}[c_q] - \bar{H}_{a,q}\big), \\ A_a(\mathbf{c}) &= \sigma_a \odot \bar{A}_a(\mathbf{c}) + \mu_a. \end{aligned} \tag{33}$$

In code, the two factors are `U[q]` and `V[q, code]`. Centering gives zero contribution over entries. The trainable base corresponds to mean-carrier initialization. Separate coefficient statistics may be refitted when adapting the main assembler to the real cloning distribution.

Exact Coefficient-Space Effect of an Intervention. For profiles differing only at q, all other contributions cancel

$$\mathrm{vec}\big(\bar{A}_a(\mathbf{c}') - \bar{A}_a(\mathbf{c})\big) = F_{a,q}\big(H_{a,q}[c'_q] - H_{a,q}[c_q]\big), \tag{34}$$

which gives

$$\|\bar{A}_a(\mathbf{c}') - \bar{A}_a(\mathbf{c})\|_F \le \|F_{a,q}\|_2 \, \|H_{a,q}[c'_q] - H_{a,q}[c_q]\|_2. \tag{35}$$

These statements hold in normalized coefficient space. Actual coefficient changes are additionally scaled by σ_a in Equation 33.

PCA Carrier. Teacher waveforms yield 100-channel log-mel features at 24 kHz, cropped or zero-padded to 768 frames. The PCA basis $V_b \in \mathbb{R}^{100\times 64}$ and mean $m \in \mathbb{R}^{100}$ were fitted on 2,072 carrier waveforms. Forward and inverse transforms are

$$X^\star = (0.1M^\star - \mathbf{1}m^\top)V_b, \qquad M_a(\mathbf{c}) = \frac{A_a(\mathbf{c})V_b^\top + \mathbf{1}m^\top}{0.1}. \tag{36}$$

$X^\star$ is in denormalized PCA coordinates. Normalized training targets are $\bar{X}^\star = (X^\star - \mu_a) \oslash \sigma_a$.

Training. For normalized predictions $\bar{X}_b$ and targets $\bar{X}_b^\star$,

$$\begin{aligned} \mathcal{L}_{\mathrm{mse}} &= \mathrm{MSE}(\bar{X}, \bar{X}^\star), \\ \mathcal{L}_{\cos} &= 1 - \mathbb{E}_b[\cos(\mathrm{vec}\,\bar{X}_b, \mathrm{vec}\,\bar{X}_b^\star)], \\ \Omega_A &= \mathrm{mean}(H^2) + \mathrm{mean}(F^2), \\ \mathcal{L}_{\mathrm{cons}} &= \mathbb{E}_b\left[\max\{0, \cos(\bar{A}_a(\mathbf{c}_{\pi(b)}), \bar{X}_b^\star) - 0.5\}\right]. \end{aligned} \tag{37}$$

Cosines flatten coefficient matrices. π is the implemented batch permutation. The consistency term penalizes excessive similarity for a mismatched input. The ZipVoice similarity uses carrier mel as prompt features, generates new speech, vocodes it, and scores against the stored embedding. Write their losses as $\mathcal{L}_{\mathrm{sim,car}} = 1 - \mathrm{SIM}_{\mathrm{car}}$ and $\mathcal{L}_{\mathrm{sim,flow}} = 1 - \mathrm{SIM}_{\mathrm{flow}}$. Only assembler parameters are updated through the frozen differentiable rendering components.

Rendering. The frozen `vocos-mel-24khz` model directly renders $\Phi_\gamma(M_a(\mathbf{c}))$ (Siuzdak, 2024). Its linguistic content follows the carrier. The ZipVoice receives carrier mel directly with fixed prompt string s_0 and target text s (Zhu et al., 2025).

B Data Construction and Metrics

The registry combines real speech with synthetic examples carrying attribute annotations. Real speech sources include AISHELL-3 (Shi et al., 2021), VoxCeleb2 (Chung et al., 2018), SeniorTalk (Chen et al., 2026), VoxAging (Ai et al., 2025), and WildElder (Wang et al., 2025a). The selected recordings contribute utterance embeddings together with available source metadata. English speech from Emilia (He et al., 2024) additionally supplies training pairs for the cloning assembler.

Utterance and Speaker Representations. For training utterance $x_{s,j}$ from speaker s, the frozen speaker encoder produces a normalized embedding $e_{s,j}$. The speaker representation is

$$e_{s,j} = E(x_{s,j}), \qquad \bar{e}_s = \mathcal{N}\left(\frac{1}{n_s}\sum_{j=1}^{n_s} e_{s,j}\right), \qquad \mathcal{N}(v) = \frac{v}{\|v\|_2}. \tag{38}$$

Here n_s is the number of training recordings. The registry retains both utterance records and mean of speakers. This supports reconstruction of voice representations and learning from variation across recordings of the same speaker. The embedding normalization statistics are shared by the codec, registrar and downstream profile construction.

Synthetic Speech. Qwen3-TTS in its voice design configuration generates 80,415 speech samples for model pretraining (Hu et al., 2026). Synthetic examples associate a waveform with its synthesis text and requested voice attributes. The generation annotations provide supervised trait labels for the corresponding registry entries. Speech sample counts and derived registry counts are tracked separately because they describe different stages of data construction. Table 7 summarizes training resources.

Table 7: Training resources.

Resource	Description	Size or coverage
Synthetic	Generated waveforms	80,415 samples
Registrar	Reference pairs	48,797 pairs
Emilia-EN	Assembler training	120,000 records
Editor supervision	Search-derived profile pairs	41,000 pairs
Carrier PCA	Carrier waveforms	2,072 waveforms

B.1 Speaker Similarity and Quality

Speaker Similarity. We use the frozen ECAPA-TDNN speaker encoder shared with profile construction (Desplanques et al., 2020). Its normalized output is denoted by $E(x)$. For synthesized speech $\hat{y}_i$ and recording x_i^{h} of the target speaker, cloning similarity is

$$\mathrm{SIM}_i = E(\hat{y}_i)^\top E(x_i^{\mathrm{h}}). \tag{39}$$

The reference and the scoring recording serve distinct roles. Higher SIM indicates greater speaker similarity in the ECAPA embedding space.

Predicted Naturalness. We use UTMOS to estimate perceptual speech quality (Saeki et al., 2022). This pretrained predictor assigns a scalar score to each waveform. Higher scores indicate higher predicted perceptual quality. UTMOS values retain their native scale and are reported for the speech configuration identified in the corresponding result table.

Intelligibility. A frozen Whisper recognizer transcribes generated speech (Radford et al., 2023). Let $\mathcal{H}(\hat{y}_i)$ be its hypothesis and η the text normalization used by the scoring pipeline. For target text s_i, the utterance character error rate is

$$\mathrm{CER}_i = 100 \times \frac{d_{\mathrm{edit}}\big(\eta(\mathcal{H}(\hat{y}_i)), \eta(s_i)\big)}{|\eta(s_i)|}, \tag{40}$$

where d_{edit} counts character substitutions, deletions, and insertions. CER is reported as a percentage with lower values indicating closer agreement with the requested text. The main comparisons estimate speech after generation.

B.2 Voice Design

Attribute Predictions. A frozen probing model supplies predictions for gender, pitch, speaking rate and age. It is composed of a gender classifier, an age predictor, and pitch and rate probes over seven ordinal categories. The age labels are child, teen, young adult, middle-aged, and elderly. Label mappings align probe outputs with the structured attribute targets associated with each evaluation case. Writing f_a for the probe and π_a for its decision and mapping rule, the predicted target label is

$$\widehat{\ell}_{i,a} = \pi_a\big(f_a(\hat{y}_i)\big). \tag{41}$$

The probe vocabulary and its task mapping determine which attribute distinctions are evaluated.

Attribute Control Accuracy. Let $\mathcal{I}_{a,g}$ contain the annotated cases contributing to attribute a in evaluation group g, and let $\ell^\star_{i,a}$ be the requested label. Accuracy is

$$\mathrm{Acc}_{a,g} = \frac{100}{|\mathcal{I}_{a,g}|} \sum_{i \in \mathcal{I}_{a,g}} \mathbf{1}\Big[\widehat{\ell}_{i,a} = \ell^\star_{i,a}\Big]. \tag{42}$$

Candidate Selection. GRAPE-TTS considers up to six generated profiles. Selection uses attribute matching counts on rendered carriers. The same probe set is used for candidate selection and attribute scoring. The single candidate setting is reported separately. Candidate rendering and rate variants are described in Appendix A.

Instruction Following. We evaluate a selected 1,200 case subset of InstructTTSEval (Huang et al., 2025) for each system. Acoustic-Parameter Specification (APS) evaluates explicit requirements for acoustic attributes. Descriptive-Style Directive (DSD) evaluates free-form descriptions of vocal style. Role-Play (RP) evaluates the vocal realization of a specified character or situation. These categories cover increasingly contextual forms of voice instructions.

Automated Judgment. Qwen2.5-Omni-7B evaluates the generated speech using the official judgement templates and the binary compatibility rubric (Xu et al., 2025; Huang et al., 2025). A compatible output satisfies the principal requested style attributes without a clear conflict. An incompatible output conflicts with a key requirement or deviates from the requested overall style. Unspecified attributes are unconstrained. The judge is used for evaluation and is separate from the trait probes. All systems are scored with the same configuration.

Let $J_i \in \{0, 1\}$ denote the parsed judgment and $\mathcal{I}_t$ the cases for instruction category t. Category scores and the reported average are

$$\mathrm{Score}_t = \frac{100}{|\mathcal{I}_t|} \sum_{i \in \mathcal{I}_t} J_i, \qquad \mathrm{Avg}_{\mathrm{inst}} = \frac{\mathrm{Score}_{\mathrm{APS}} + \mathrm{Score}_{\mathrm{DSD}} + \mathrm{Score}_{\mathrm{RP}}}{3}. \tag{43}$$

B.3 Voice Editing

Identity consistency. For edited output $\hat{y}_i^{(t)}$ and unedited output $\hat{y}_i^{(0)}$, we use

$$\mathrm{id}_{\mathrm{self},i} = E(\hat{y}_i^{(t)})^\top E(\hat{y}_i^{(0)}). \tag{44}$$

The main editing table reports $\mathrm{id}_{\mathrm{self}}$, which measures consistency with the voice realized by each system before editing.

Pitch and Speaking Rate. Pitch is estimated with YIN and a RMS-based filter (de Cheveigné & Kawahara, 2002). The waveform estimate is the median over the retained frames. Let $F(x)$ denote this estimate, $T(x)$ the measured duration, and $x_i^{(0)}$ the unedited reference used for the requested transformation. The realized pitch and rate ratios can be expressed as

$$\widehat{r}_{\mathrm{p},i} = \frac{F(\hat{y}_i^{(t)})}{F(x_i^{(0)})}, \qquad \widehat{r}_{\mathrm{r},i} = \frac{T(x_i^{(0)})}{T(\hat{y}_i^{(t)})}. \tag{45}$$

The requested ratios are $2^{\delta_i/12}$ for $\delta_i \in \{-2, +2\}$ semitones and $\rho_i \in \{0.8, 1.25\}$ for speaking rate. Relative difference error (RDE) measures disagreement between the realized change and its target under the task scorer.

Age Control. The age operation requests a young adult voice. For the 200 assigned age editing cases, the hit rate is

$$\mathrm{Hit}_{\mathrm{age}} = 100 \times \frac{1}{200} \sum_{i \in \mathcal{I}_{\mathrm{age}}} \mathbf{1}\left[\widehat{\ell}_{i,\mathrm{age}} = \text{young adult}\right]. \tag{46}$$

The reported percentage uses the calibrated age predictor. Identity consistency and accuracy are evaluated on the same subset. Systems without the result are marked with a dash.

DSP Reference. The DSP implementation is evaluated with the same setting as the systems. It provides a reference for control precision under the measurement pipeline. Its results depend on the transformation and estimator used in this experiment.

Table 8: Baseline task assignments. C, D, and E denote cloning, design and editing.

System or configuration	C	D	E
F5-TTS (Chen et al., 2025b)	✓		
ZipVoice (Zhu et al., 2025)	✓		
IndexTTS2 (Zhou et al., 2025a)	✓		
MOSS-TTSD (Zhang et al., 2026)	✓		
CosyVoice2 (Du et al., 2024)	✓		
CosyVoice3 (Du et al., 2025)	✓		
VoxCPM (Zhou et al., 2025b)	✓		
Qwen3-TTS cloning (Hu et al., 2026)	✓		
XTTS-v2 (Casanova et al., 2024)	✓		
Higgs Audio v2 (Boson AI, 2025)	✓	✓	
Higgs Audio v3 (Boson AI, 2026)	✓	✓	
Zonos (Clark et al., 2026)	✓		
Chatterbox (Resemble AI, 2025)	✓		
FireRedTTS3 (Shen et al., 2026)	✓	✓	✓
Step-Audio-EditX (Yan et al., 2025)	✓		✓
Qwen3-TTS voice design (Hu et al., 2026)		✓	
Spark-TTS (Wang et al., 2025b)		✓	
Parler-TTS (Lacombe et al., 2024)		✓	
DSP reference			✓

C Baseline Configurations

We evaluate 18 baselines using their official public checkpoints and released inference codes. Table 8 gives the evaluated task assignments. The design baselines also appear in the instruction following comparison.

Voice Design. Design systems receive the voice description and synthesis text using their native entrypoint. Baselines produce one candidate per input. GRAPE-TTS uses the profile proposal budget and candidate selection described in Appendix A.

Voice Cloning. Each baseline receives the reference speech and target text for the cloning case through its native conditioning path. The reference transcript is supplied when required by the released inference procedure. Each cloning baseline is evaluated on 1,067 cases. ZipVoice is included both as a direct reference baseline and as the frozen backbone used by GRAPE-TTS.

Voice Editing. Voice editing baselines receive source speech and the requested attribute modification. Step-Audio-EditX includes the age operation in the evaluated baseline set, while FireRedTTS3 has no reported age-editing result caused by runtime errors while attempting to evaluate. The DSP supplies pitch and rate reference transformations.

D Extended Task Results

D.1 Voice Design and Local Instruction Following

Table 9 retains the original design and local-judge results. Single-candidate's instruction scores were not recorded.

Sub-benchmark Comparisons. Figure 4 reports evaluation results within each design sub-benchmark. The configuration and candidate budgets follow Appendix B. The `psc` results cover the selected ParaSpeechCaps cases with age excluded because this subset has no age targets.

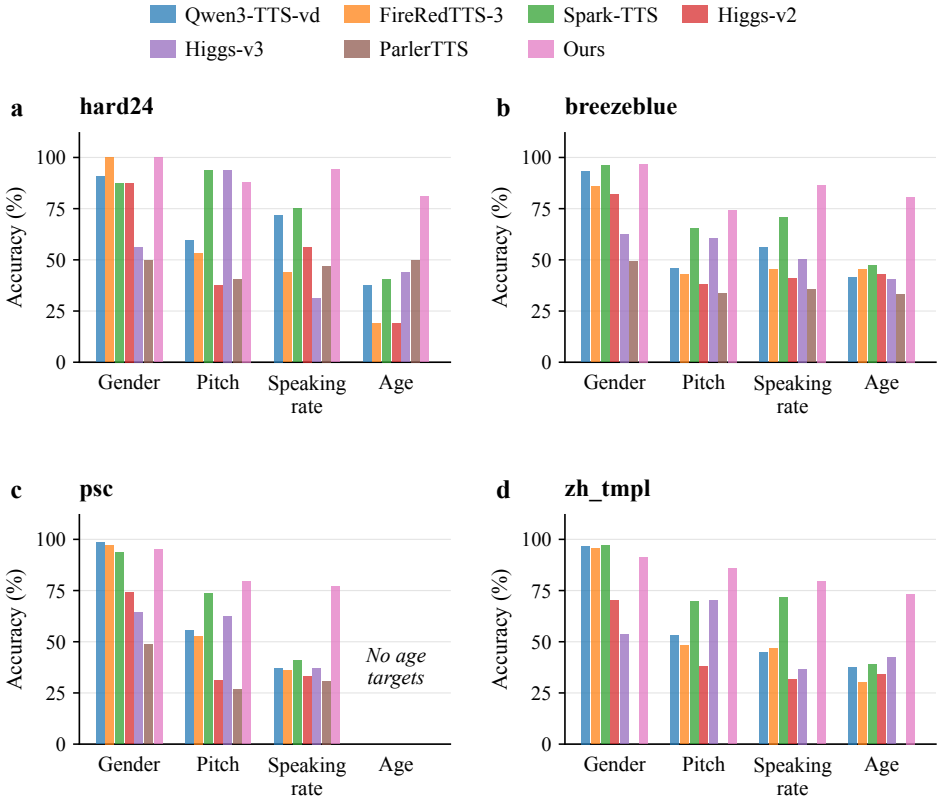

Figure 4: Comparison of voice design accuracy across four sub-benchmarks.

Table 9: Complete comparison of design and local instruction compatibility (%).

	Design				Local instruction judge			
Configuration	Gender	Pitch	Rate	Age	APS	DSD	RP	Avg.
Qwen3-TTS-vd	95.8	54.1	62.0	36.9	99.7	85.3	69.3	84.8
FireRedTTS3	95.8	49.5	44.4	33.0	99.7	83.0	66.9	83.2
Spark-TTS	94.8	79.3	71.7	41.1	100.0	81.2	59.4	80.2
Higgs Audio v2	77.8	34.0	42.4	36.4	99.7	79.8	59.3	79.6
Higgs Audio v3	57.4	77.3	36.1	44.2	99.5	76.9	60.9	79.1
Parler-TTS	49.3	33.7	37.7	41.5	100.0	64.3	50.0	71.4
GRAPE-TTS, $N = 6$	96.6	82.9	87.3	77.6	100.0	91.7	85.3	92.3
GRAPE-TTS, $N = 1$	93.2	57.0	65.6	56.0	–	–	–	–

Table 10: Complete recorded cloning results.

System	ELDERLY67	Seed-zh	Seed-en	Macro	UTMOS
F5-TTS	0.446	0.758	0.658	0.621	3.074
IndexTTS2	0.465	0.726	0.610	0.600	3.049
MOSS-TTSD	0.440	0.727	0.626	0.598	3.371
CosyVoice3	0.443	0.736	0.581	0.586	3.471
VoxCPM	0.416	0.698	0.622	0.579	3.144
Step-Audio-EditX	0.447	0.690	0.588	0.575	3.620
FireRedTTS3	0.463	0.659	0.587	0.570	3.344
CosyVoice2	0.426	0.709	0.556	0.564	3.728
Qwen3-TTS	0.368	0.724	0.599	0.564	3.720
Higgs Audio v2	0.382	0.689	0.604	0.558	3.447
Higgs Audio v3	0.355	0.657	0.563	0.525	3.550
Zonos	0.395	0.590	0.587	0.524	3.165
Chatterbox	0.363	0.519	0.605	0.496	3.113
XTTS-v2	0.356	0.627	0.492	0.492	3.103
ZipVoice (backbone)	0.409	0.750	0.666	0.608	3.222
Ours	**0.580**	0.611	0.561	0.584	3.184

GRAPE-TTS attains the highest accuracy of speaking rate in every reported subset and the highest age accuracy in all three subsets with age annotations. Its pitch accuracy is highest on `breezeblue`, `psc`, and `zh_tmpl`, while Spark-TTS and Higgs-v3 obtain higher pitch accuracy on `hard24`. Gender results vary across subsets. Qwen3-TTS-vd leads on `psc`, and Spark-TTS leads on `zh_tmpl`. These comparisons locate the strongest advantages in speaking rate and age control while identifying attribute-specific differences within the benchmark.

D.2 Cloning Results

Table 10 retains all fifteen cloning baselines from the original experiment table. GRAPE-TTS achieves a macro speaker similarity of 0.584 and UTMOS at 3.184, only damaging minor part of capability of the base model due to the fidelity decrease from PCA pass and 24 kHz vocoder. With specific training, GRAPE-TTS achieves the highest similarity on elderly67 at 0.580, exceeding the strongest baseline by 0.115. These results support the capability of cloning through the shared acoustic profile.

E Extended Ablations

E.1 Assembler Capacity

Training Coverage. The cloning ablations in Table 2 evaluates reference registration and English training coverage. Adding 30,000 Emilia-EN examples improves similarity. Extending the schedule to 120 epochs and increasing assembler capacity provide further reported gains. These comparisons associate improved recovery with training coverage and sufficient optimization of the profile-to-carrier map. The incremental gains refer to their comparisons.

Codebook Size. A separate retraining comparison uses 1024 and 2048 entries per residual level. The two settings yield matching values of 0.2426 and 0.2410, complete agreement of 0.0098 in both cases. The small differences provide no consistent benefit from increased codebook size in this comparison. The deployed profile retains eight levels with 2048 entries each.

E.2 Candidate Selection for Voice Design

The matched design comparison keeps the generation and descriptions fixed while changing the profile candidate budget from $N = 1$ to $N = 6$. As reported in Table 1, the resulting accuracy gains

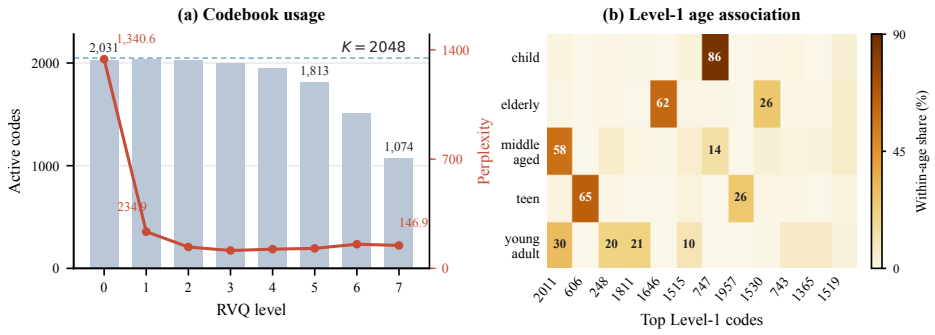

Figure 5: **Codebook usage and age association.** (a) Active code counts (bars, left axis) and perplexity (red curve, right axis) across the eight RVQ levels. The dashed line indicates the capacity of 2,048 entries per codebook. (b) Distribution of selected identifiers within each age class, normalized over the displayed codes separately for each row.

are 3.4 percentage points for gender, 25.9 for pitch, 21.7 for speaking rate, and 21.6 for age. The largest changes occur on attributes with lower accuracy at single candidate. These measurements quantify the benefit of the deployed sampling and selection configuration.

Voice design selection operates on generated profiles using attribute agreement. The InstructTTSEval results in Table 1 provide a complementary assessment through a judge separate from the selection probes.

Residual Calibration Error. The pitch result reports a median final residual gap of 0.013, with larger residuals concentrated near the displacement bound $|d| = 2.2$. The domain summaries show larger residuals for VoxAging than for the other two reported groups. Boundary accumulation is consistent with limited correction range for some speakers. It also motivates examining the pitch estimator and carrier response together when interpreting the residual error.

E.3 Codebook Usage and Attribute Association

We inspect code usage on 287,248 encoded registry rows. Let $p_q(k)$ denote the empirical frequency of identifier k at level q. The active code count and perplexity are

$$A_q = \sum_{k=0}^{K-1} \mathbf{1}[p_q(k) > 0], \qquad \mathrm{PPL}_q = \exp\left(- \sum_{k:p_q(k)>0} p_q(k) \log p_q(k) \right). \tag{47}$$

At level 0, 2,031 of 2,048 entries are active and perplexity is 1,341. The remaining levels have reported perplexities between 114 and 235. Active coverage and effective usage consequently differ across the residual hierarchy, with level 0 exhibiting the broadest distribution.

Conditional age frequencies reveal concentrated level-1 associations. Identifier 747 occurs in 86% of child-labeled rows, identifier 606 in 65% of teen-labeled rows, and identifier 1646 in 62% of elderly-labeled rows. The two adult categories distribute their mass across multiple codes. Level-0 code groups are also strongly associated with gender labels in the reported inspection. These frequencies characterize how annotated traits align with occupied regions of the profile space.

Usage statistics and intervention measurements provide complementary views of the hierarchy. Concentrated code-label associations indicate predictable structure, while the substitution experiments expose the acoustic attributes and identity components that change together. The combined analysis supports a structured voice address with empirically measurable coupling across its coordinates.