DUALREF-TTS: A LABEL-FREE DUAL-REFERENCE FRAMEWORK FOR INDEPENDENT STYLE-TIMBRE CONTROL IN ZERO-SHOT SPEECH SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot text-to-speech (TTS) systems have achieved significant progress in both speaker similarity and speech naturalness. Building on these advances, recent work has further explored more independent control over emotion and speaking style through a variety of paradigms. However, most existing approaches still rely on human-defined style taxonomies, which limits their ability to achieve reliable dual-reference control, scalable label-free learning, and robust style control for in-the-wild scenarios. We introduce DualRef-TTS, a label-free dual-reference paradigm for independent style–timbre control in zero-shot speech synthesis, where style is learned as a continuous, reference-driven, and speaker-transferable variable, rather than as a predefined label, abstract control vector, or prompt-derived description. First, attribute perturbation constructs attribute-purified style references for label-free style supervision. Second, Branch-Aware Positional Conditioning (BAPC) combines branch embeddings with branch-wise positional offsets to distinguish the timbre reference and style reference, thereby reducing cross-branch interference. Building on this architecture, we further adopt a mean-flow-based acoustic generation method to reduce inference steps while preserving synthesis quality. Finally, Direct Preference Optimization (DPO)-style post-training further aligns the generation distribution with style alignment, semantic accuracy, and prosodic naturalness. Experimental results demonstrate that the proposed method is comparable to state-of-the-art (SOTA) systems across standard metrics, while introducing a label-free dual-reference paradigm that avoids intermediate representations and prompt-based control. This supports more flexible and independent control of timbre and style in open-domain speech synthesis. Audio samples are available at: `https://dualref-tts.github.io/`

1 INTRODUCTION

Conventional zero-shot text-to-speech (TTS) is predominantly formulated as single-reference voice cloning (Anastassiou et al., 2024; Chen et al., 2024; 2025; Liao et al., 2024; Wang et al., 2023; Le et al., 2023), where a prompt utterance jointly specifies speaker timbre and speaking style. While this formulation is effective for joint modeling of timbre and style, it is inherently mismatched with scenarios that require independent control over these two factors. In practice, users may wish to preserve the timbre of one speaker while transferring the style of another, especially when the target style is subtle, highly personalized, or difficult to verbalize. Recent work has explored more flexible control over speaking style and timbre through a variety of paradigms, including disentanglement-based modeling (Ju et al., 2024), discrete style labels (Kang et al., 2023; Li et al., 2025a), continuous control variables (Cho et al., 2024), natural language prompts (Du et al., 2025; Hu et al., 2026; Wu et al., 2025b), and reference-based transfer (Zhang et al., 2025c; Zhou et al., 2026; Li et al., 2025b). Nevertheless, a unified, label-free, and broadly applicable style representation remains missing.

One line of research attempts to promote style-related factorization through feature-level supervision. Early methods such as StyleTTS2 (Li et al., 2023b) and NaturalSpeech3 (Shen et al., 2023; Ju et al., 2024) use prosodic cues including fundamental frequency (F0), energy, and duration (Inoue et al., 2025) to supervise style-related branches or encoders. While these signals are informative, they are intrinsically mixed: besides style variation, they also encode speaker identity information,

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.

such as phonation patterns and timbre characteristics. During training, models can exploit such mixed cues as shortcuts, which often leads to timbre leakage in the learned style representation and weakens cross-speaker style transfer (Cho et al., 2025).

Another line of work focuses on designing more flexible control interfaces, exposing style through discrete labels (Li et al., 2025a; Kang et al., 2023), disentangled continuous variables (Cho et al., 2024), natural-language prompts (Hu et al., 2026; Wu et al., 2025b; Du et al., 2025), or reference-based control (Li et al., 2023b; Zhang et al., 2025c). These paradigms improve controllability to varying degrees, but they still represent style indirectly through human-defined taxonomies, abstract affective coordinates, linguistic descriptions, or compressed proxy representations. Discrete labels impose a closed control space (Xing et al., 2025); abstract continuous variables are difficult for users to interpret and manipulate; natural-language prompts often struggle to specify subtle or highly customized styles (Yang et al., 2024; Vyas et al., 2023); and even reference-based systems may project rich continuous style cues into predefined categories or compact control codes before generation. Consequently, existing methods fail to model style as a label-free, continuous, and reference-guided variable, limiting effective timbre–style separation, open-set controllability, and scalable learning from unlabeled speech. Throughout this paper, we use style as a broad term for expressive speech characteristics beyond textual content and speaker identity.

Accordingly, we propose DualRef-TTS, a label-free dual-reference framework for zero-shot TTS that enables independent control of style and timbre. Our framework is designed around two challenges. First, style-matched reference audio is difficult to obtain during training. In natural speech corpora, it is rare to observe an auxiliary utterance that preserves the target speaking style while varying other factors such as timbre and content; even segments from the same utterance may drift in style over time. To address this issue, we introduce an attribute perturbation strategy that constructs attribute-purified style references by perturbing non-style factors, including timbre, F0, and content. Second, even with improved style references, a shared generative backbone may still re-entangle timbre, target generation state, and style. We therefore design Branch-Aware Positional Conditioning (BAPC), which combines branch embeddings with branch-wise positional offsets to organize the timbre, target, and style streams before fusion in the shared backbone, mitigating cross-branch source information leakage. Based on this dual-reference framework, we adopt mean-flow modeling to reduce inference steps (Geng et al., 2025; Lipman et al., 2022; Zheng et al., 2025; Mehta et al., 2024) without compromising synthesis quality, and further introduce a Direct Preference Optimization (DPO)-style post-training objective to improve style similarity, semantic accuracy, and prosodic naturalness (Zhang et al., 2025a; Rafailov et al., 2023; Wallace et al., 2024). Experiments show that DualRef-TTS remains competitive with state-of-the-art (SOTA) systems on standard objective and subjective metrics while enabling zero-shot generation with independent control over style and timbre. Our main contributions are summarized as follows:

- We formulate zero-shot TTS with independent style–timbre control as a label-free dual-reference generation problem, where style is modeled as a continuous, reference-driven, and speaker-transferable variable rather than a predefined label, abstract control vector, or prompt-derived description.
- We introduce an attribute perturbation strategy that constructs attribute-purified style references from unpaired training utterances, enabling large-scale label-free style learning without manual style supervision or discrete intermediate style bottlenecks.
- We propose BAPC, a lightweight conditioning scheme that augments the timbre, target, and style streams with branch embeddings and branch-wise positional offsets, providing a structural bias for more stable dual-reference control within a unified generative backbone.
- Building on this dual-reference architecture, we combine mean-flow acoustic generation with DPO-style post-training to improve inference efficiency, style fidelity, semantic accuracy, and prosodic naturalness.

2 Related work

Modern zero-shot TTS uses text and short reference speech. Diffusion or flow-based generators (Le et al., 2023; Peng et al., 2025), neural codec language models (Du et al., 2024; 2025; Chen et al., 2024; Liao et al., 2024; Borsos et al., 2023; Wang et al., 2023), and masked codec transformers

(Wang et al., 2024) improve naturalness and speaker generalization. Reference speech provides speaker cues for cloning (Casanova et al., 2022; Qin et al., 2023), but the standard single-reference formulation uses the same utterance to specify both timbre and speaking style, coupling the two factors at the conditioning level(Zhang et al., 2025c; Li et al., 2025b).

Style-controllable TTS systems mainly differ in the interface used to expose style to the generator. Existing systems condition generation on discrete emotion or style labels (Li et al., 2025a; Xing et al., 2025; Deng et al., 2025), continuous affective vectors (Cho et al., 2024), natural-language prompts (Hu et al., 2026; Wu et al., 2025b; Vyas et al., 2023; Yang et al., 2024; Huang et al., 2025; Wu et al., 2025a), or speech references (Zhou et al., 2026; Li et al., 2023b). Appendix A.3 summarizes these interfaces, generation mappings, and representative systems. Labels are easy to supervise but impose a closed style space; continuous vectors are more flexible but hard for users to interpret; prompts improve interaction but may underspecify subtle or personalized styles, and speech references are intuitive but often compressed into intermediate representations. Thus, existing interfaces still tend to represent style through human-defined categories or proxy control codes, rather than open-set, label-free, and speaker-transferable reference control.

In addition to control interfaces, disentangled representation learning uses prosodic and structural constraints to improve style control. F0, energy, and duration provide expressive cues (Ju et al., 2024; Inoue et al., 2025; Ren et al., 2020), but also encode speaker identity, phonation, and content. Other methods introduce dedicated style branches and factorization (Cho et al., 2025; Lu et al., 2023), together with adversarial and identity-suppression methods (Deng et al., 2024; Du et al., 2021; Chou et al., 2024) further reduce identity leakage but do not provide a clean, reusable, directly controllable, cross-speaker style reference. Singing Voice Conversion (SVC) (Geng et al., 2026; Liu, 2024; Li et al., 2023a; Bai et al., 2024) and factorized acoustic generation (Hsu et al., 2019; Gengembre et al., 2024; Xie et al., 2024; Choi et al., 2021) indicate that timbre, pitch, and content can be manipulated relatively independently. Constant-Q Transform (CQT)-based modeling improves pitch representation (Gu et al., 2024), while speech disentanglement(Chan et al., 2022; Liu et al., 2023) further supports independent content control. Although mainly developed for conversion or editing, these methods motivate attribute perturbation for cleaner label-free style references.

3 METHOD

3.1 OVERVIEW

As shown in Figure 1, we explore a label-free dual-reference zero-shot TTS framework. Let x denote the input text, y_{tim} a timbre reference utterance, and y_{sty} a style reference utterance. The goal is to synthesize speech $\hat{y}$ such that its linguistic content matches x, its speaker identity follows y_{tim}, and its speaking style is transferred from y_{sty}. Formally, the generation process can be written as

$$\hat{y} = G_\theta(x, y_{\mathrm{tim}}, y_{\mathrm{sty}}). \tag{1}$$

Compared to conventional single-reference zero-shot TTS and style-controllable systems that rely on labels, latent vectors, or prompts (Kang et al., 2023; Zhou et al., 2026; Cho et al., 2024; Hu et al., 2026; Du et al., 2025), our formulation explicitly separates timbre control from style control. We model style as a continuous reference-conditioned variable that can be transferred independently across speakers. This design demonstrates a stronger alignment with real-world usage, where the target is directly controlled by a style reference utterance.

However, this setting introduces two major challenges. First, under label-free training, it is difficult to obtain style supervision that is sufficiently style-focused and stable. Second, even with dual-reference inputs, a shared generative backbone may still re-entangle timbre, target generation state, and style unless their roles are explicitly separated. To address these issues, DualRef-TTS is built upon four components: attribute perturbation for constructing attribute-purified style references, BAPC for organizing timbre, target, and style streams with branch embeddings and positional offsets, mean-flow modeling for efficient acoustic generation, and DPO-style post-training for preference alignment.

A major difficulty in label-free style learning is that style-matched reference audio is difficult to obtain during training. Ideally, training would require an auxiliary utterance that matches the target

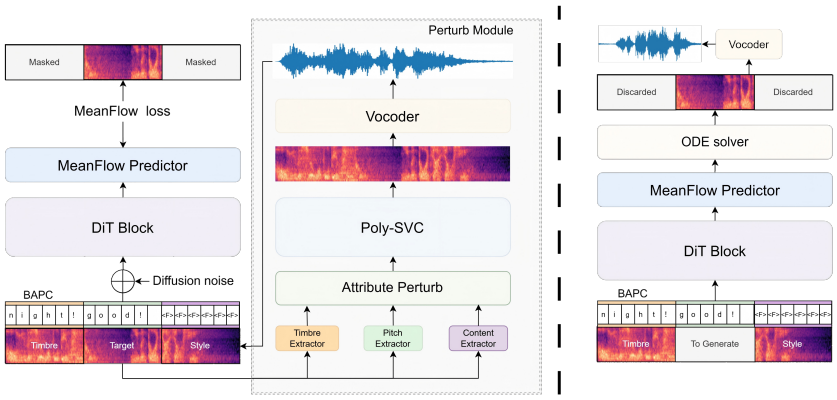

Figure 1: Comprehensive diagram of the DualRef-TTS model. The left shows the training procedure, the right the inference procedure.

speaking style while differing in timbre, linguistic content, and other nuisance factors. In practice, such reference pairs are rarely available in natural speech corpora. Even segmenting multiple regions from the same utterance does not reliably solve this issue, since speaking style may evolve over time and adjacent segments can still exhibit noticeable intra-utterance inconsistency. As a result, naturally available reference pairs are insufficient to support scalable learning of a stable and transferable style variable.

To address this issue, we introduce an attribute perturbation strategy that constructs an attribute-purified style reference from each training utterance. As shown in Figure 2, rather than requiring manually curated style pairs or style annotations, we explicitly perturb non-style factors in the target speech and resynthesize a reference utterance that preserves the dominant speaking style while suppressing interfering cues. This allows the model to learn style from ordinary training data while maintaining a reference-based supervision interface (Lu et al., 2023; Choi et al., 2021). Given a training utterance y, we first extract its timbre, content, and pitch features:

$$f_{\text{tim}} = E_{\text{tim}}(y), \quad f_{\text{con}} = E_{\text{con}}(y), \quad f_{\text{pit}} = E_{\text{pit}}(y). \tag{2}$$

We then apply targeted perturbations to these non-style factors to obtain $\tilde{f}_{\text{tim}}$, $\tilde{f}_{\text{con}}$, and $\tilde{f}_{\text{pit}}$, and reconstruct an attribute-purified style reference through a singing voice conversion (SVC)-based resynthesis module (Geng et al., 2026):

$$\tilde{y}_{\text{sty}} = \Phi(\tilde{f}_{\text{tim}}, \tilde{f}_{\text{con}}, \tilde{f}_{\text{pit}}). \tag{3}$$

3.2 Attribute-Purified Style Reference Construction

Timbre perturbation We replace the original speaker-related representation (Hsu et al., 2019) with timbre features sampled from other speakers, thereby explicitly disrupting the shortcut correlation between style and speaker identity.

Pitch perturbation We adopt a CQT-based pitch representation and apply both global (±12 semitones) and local (±4 semitones) pitch shifts to mitigate residual speaker information, including gender and habitual prosody patterns.

Content perturbation We perturb the content representation through temporal smoothing, feature shuffling, and cross-sample replacement to prevent the style branch from capturing semantic and local phonetic information.

After perturbation and resynthesis, the resulting utterance still preserves the dominant speaking style, while the influence of timbre, content, and pitch-related shortcuts is substantially reduced. This makes $\tilde{y}_{\text{sty}}$ a more suitable supervisory signal for learning style as an independent variable. Importantly, our method constructs a reference-based style signal by substantially mitigating the influence of non-style factors, thereby preserving consistency between training-time supervision and inference-time control.

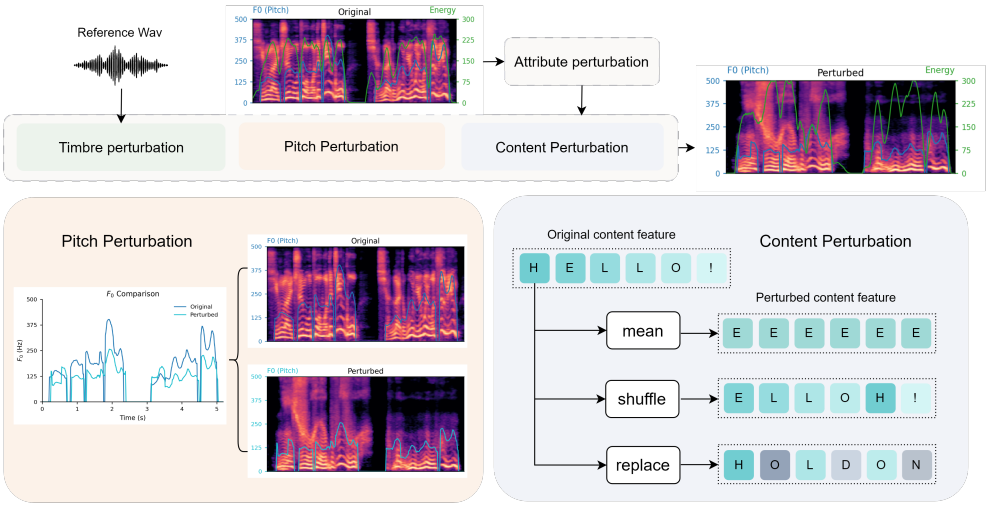

Figure 2: Attribute Perturbation framework.

3.3 Branch-Aware Positional Conditioning

Constructing a more purified style reference is necessary but not sufficient for independent style-timbre control. Even with an improved style condition, the shared generative backbone still needs to jointly process the timbre reference, the target generation state, and the style reference. Without explicit structural constraints, these inputs may still interfere with one another in latent space, leading to ambiguous factor attribution and incomplete disentanglement. To address this issue, we introduce BAPC, which combines branch embeddings with branch-wise positional offsets. The branch embeddings indicate whether a token belongs to the timbre reference, the style reference, or the target generation state, while the positional offsets assign different branches to separate positional coordinate ranges. This design mitigates cross-branch source information leakage before shared-backbone fusion.

BAPC uses branch embeddings to identify the timbre reference branch, the style reference branch and the target generation branch structurally. Specifically, we construct three branch-wise conditional representations:

$$h_{\text{tim}} = E_{\text{a}}(y_{\text{tim}}) + e(c_{\text{tim}}), \quad h_{\text{sty}} = E_{\text{a}}(\tilde{y}_{\text{sty}}) + e(c_{\text{sty}}), \quad h_{\text{tar}} = z_t + e(c_{\text{tar}}). \tag{4}$$

where $E_{\text{a}}(\cdot)$ denotes the acoustic encoder, z_t is the noisy target state at time step t, and $e(c)$ denotes the embedding of the branch-specific identifier.

In addition to branch embeddings, we further introduce branch-wise positional shift to explicitly separate the three streams in latent space. Let $p(\cdot)$ denote the positional encoding function. For branch $b \in \{\text{tim}, \text{tar}, \text{sty}\}$, the i-th token representation is written as

$$\bar{h}_b(i) = h_b(i) + p(i + \Delta_b), \tag{5}$$

where Δ_b is the branch-specific positional offset.

The resulting branch-wise representations are then combined with text conditioning and fed into a shared DiT backbone (Peebles & Xie, 2023) for acoustic generation. Here, text conditioning provides the linguistic content constraint, while BAPC distinguishes the three acoustic sources through branch embeddings and branch-wise positional offsets, namely the timbre reference, target generation state, and style reference. This design does not guarantee complete statistical disentanglement by itself; rather, it provides a lightweight structural bias that reduces source ambiguity when multiple acoustic streams are processed by a shared backbone.

3.4 Mean-Flow Acoustic Generation

After establishing the dual-reference conditional formulation, we further consider generation efficiency. Standard diffusion or flow-matching based acoustic generation often requires many inference

Table 1: Objective experimental results on style-controllable zero-shot TTS.

Model	WER (%) ↓	DNSMOS ↑	T-SIM ↑	S-SIM ↑	ΔT-SIM	ΔS-SIM	RTF ↓	Interface
F5-TTS	2.053	3.907	0.742	-	-	-	0.097	✗
StyleTTS2	1.744	3.732	0.463	0.552	0.393	-0.033	**0.033**	A
Vevo2	3.647	3.885	0.623	0.425	0.449	-0.176	0.183	A
IndexTTS2	0.953	3.933	0.724	0.451	0.483	-0.154	0.622	A
CosyVoice3	1.224	3.922	0.744	0.442	-	-	0.854	T
Qwen3-TTS	1.034	3.896	0.654	0.327	-	-	0.904	T
Ours (CFM)	**0.941**	3.908	**0.751**	0.603	-	-	0.101	A
Ours (MF)	0.984	**3.946**	0.736	**0.624**	**0.522**	**0.078**	0.049	**A**

steps(Lipman et al., 2022; Mehta et al., 2024; Popov et al., 2021), which limits the efficiency of zero-shot TTS in real-world applications. To reduce inference cost while preserving synthesis quality, we adopt mean-flow modeling (Geng et al., 2025) as the training objective for acoustic generation.

Instead of predicting the instantaneous velocity field, mean-flow modeling estimates the average transport direction over a time interval (r, t). Given the noisy target state and the conditional representations, the DiT backbone predicts the mean-flow field as

$$\hat{u}_\theta = F_\theta\big(\bar{h}_{\mathrm{tim}}, \bar{h}_{\mathrm{tar}}, \bar{h}_{\mathrm{sty}}, x, r, t\big)\,, \tag{6}$$

where r and t are the start and end of the time interval, with $r < t$. The corresponding training objective is defined as

$$\mathcal{L}_{\mathrm{mf}} = \mathbb{E}\left[\|\hat{u}_\theta - u^*\|_2^2\right], \tag{7}$$

where u^* denotes the target mean-flow field.

Because mean-flow models interval-level transport behavior rather than highly local instantaneous dynamics, the resulting trajectories are smoother and more numerically stable during integration. This allows the model to achieve high-quality generation with fewer diffusion steps. In our framework, mean-flow is introduced as an efficiency-oriented acoustic generation objective, rather than as the core disentanglement mechanism itself.

3.5 Preference-Aligned Post-Training

Although the mean-flow objective provides an effective reconstruction target, it does not explicitly optimize the perceptual criteria that matter most in style-controlled TTS, such as style similarity, semantic accuracy, and prosodic naturalness. To further align the model output with these goals, we introduce a DPO-style post-training (Rafailov et al., 2023) that reshapes the generation distribution through preference learning (Wallace et al., 2024). Given a condition tuple $c = (x, y_{\mathrm{tim}}, y_{\mathrm{sty}})$, we construct preference pairs (y^+, y^-), where y^+ is preferred over y^- according to a target criterion.

Following the standard DPO formulation, we optimize

$$\mathcal{L}_{\mathrm{DPO}} = -\log\sigma\Big(\beta\big[(\log p_\theta(y^+ \mid c) - \log p_\theta(y^- \mid c)) - (\log p_{\mathrm{ref}}(y^+ \mid c) - \log p_{\mathrm{ref}}(y^- \mid c))\big]\Big), \tag{8}$$

where p_θ denotes the current model and p_{ref} denotes the reference model.

We use three complementary preference sources, targeting style alignment, semantic accuracy, and prosodic naturalness. These preference signals improve style fidelity, content correctness, and overall perceptual quality. Further DPO details, including training configuration and preference-pair construction, are provided in Appendix F.1.

4 Experiments

4.1 Experimental Setup

Datasets DualRef-TTS is trained on filtered public TTS datasets and open-source expressive speech corpora (He et al., 2024; Ji et al., 2024; Lotfian & Busso, 2017; Busso et al., 2008). A

multi-stage pipeline removes samples with unreliable transcriptions, poor acoustic quality, unstable pronunciation, or weak style expressiveness, yielding approximately 6,000 hours of speech; dataset statistics and filtering details are provided in Appendices A.1 and A.2. Objective evaluation uses a style-timbre-text grid comprising 16 style references, 8 timbre references, and 8 Chinese/English prompts. Subjective evaluation uses a more challenging open-style set of 26 style references, each paired with male and female timbre references and Chinese and English prompts; the complete style descriptions are provided in Appendix B.2. These grids assess style transfer under fixed timbre and timbre preservation under fixed style, including the cross-gender and cross-lingual settings analyzed in Appendix D.

Experimental Setup We train the model for approximately 1M steps on two NVIDIA A100 GPUs with a batch size of 32, using AdamW with a learning rate of 9.5×10^{-5} and a cosine annealing schedule. We then perform DPO post-training for two epochs on 500 paired samples with a learning rate of 1×10^{-7}. Additional model and inference details are provided in Appendices A.4 and A.5.

4.2 Baselines

We compare DualRef-TTS with six representative zero-shot and controllable TTS systems spanning both autoregressive (AR) and non-autoregressive (NAR) paradigms. F5-TTS (Chen et al., 2025) serves as a strong flow-matching-based NAR zero-shot baseline with high naturalness and timbre generalization. StyleTTS2 and IndexTTS 2 (Li et al., 2023b; Zhou et al., 2026) represent reference-guided TTS systems that condition generation on speech references for expressive synthesis. Vevo2 (Zhang et al., 2025b) represents a hybrid AR and flow-matching framework with disentangled reference-based control over style and timbre. CosyVoice3 and Qwen3-TTS (Du et al., 2025; Hu et al., 2026) represent large-scale instruction-based TTS systems that control style and emotion through natural-language prompts. Detailed descriptions and instruction prompts are provided in Appendices A.3 and B.2.

4.3 Evaluation Metrics

Since objective metrics do not always correlate well with human perception, we adopt both objective and subjective evaluations (Cooper et al., 2024). This is particularly important when diverse speaking styles introduce large acoustic variations that may degrade the reliability of open-source evaluation models.

Objective Metrics For objective evaluation, we report WER for content accuracy, DNSMOS (Reddy et al., 2021) for acoustic naturalness, S-SIM for style similarity, T-SIM for timbre similarity, and RTF for inference efficiency. WER is computed using Whisper (Radford et al., 2023) for English and Paraformer (Gao et al., 2022) for Chinese, while T-SIM is computed with a pretrained speaker verification model (Chen et al., 2023). Since no standardized representation exists for style similarity, we compute S-SIM using a pretrained emotion recognition model (Bott et al., 2024) and treat it as a proxy metric. Appendix B.4 discusses its interpretation and limitations.

Subjective Evaluation We conduct a human evaluation with 24 bilingual Chinese-English participants. Each sample is rated on a 5-point scale for subjective style similarity (SS-SIM), subjective timbre similarity (ST-SIM), and overall quality (S-MOS). We report mean scores with 95% confidence intervals. The full protocol is provided in Appendix B.3.

5 Results

5.1 Main Results

Objective Results Table 1 summarizes the objective results on controlled TTS, where MF denotes MeanFlow, and A and T in the Interface column denote audio-reference and text-based style control, respectively. ΔT-SIM is computed as the T-SIM to the timbre reference minus that to the style reference, while ΔS-SIM is computed as the S-SIM to the style reference minus that to the timbre reference. Our MeanFlow model achieves the best DNSMOS and S-SIM while remaining competitive in WER and T-SIM. Compared with the closely related dual-reference system Vevo2,

Table 2: Subjective scores for open-style settings

Model	S-MOS ↑	ST-SIM ↑	SS-SIM ↑
StyleTTS2	3.90 ± 0.09	3.12 ± 0.07	3.20 ± 0.06
Vevo2	3.93 ± 0.05	3.42 ± 0.06	3.13 ± 0.06
IndexTTS2	3.92 ± 0.08	3.56 ± 0.06	3.11 ± 0.05
Qwen3-TTS	3.82 ± 0.09	3.66 ± 0.08	2.86 ± 0.05
CosyVoice3	3.88 ± 0.07	3.62 ± 0.07	2.92 ± 0.06
Ours	$\mathbf{3.95 \pm 0.08}$	$\mathbf{3.68 \pm 0.06}$	$\mathbf{3.92 \pm 0.08}$

Table 3: Same-backbone disentanglement comparison.

Style Representation	WER (%) ↓	DNSMOS ↑	T-SIM ↑	S-SIM ↑	ΔT-SIM ↑	ΔS-SIM ↑
StyleTTS2	1.428	3.883	0.526	0.543	0.055	0.011
IndexTTS2	1.164	**3.920**	0.732	0.471	0.413	-0.162
Vevo2-CS (same-utt)	166.588	3.856	0.653	**0.641**	0.476	0.106
Vevo2-CS (non-overlap)	2.823	3.874	0.637	0.495	0.468	-0.126
Vevo2-Prosody	26.513	3.801	0.632	0.541	0.455	-0.052
Ours (AP)	**0.941**	3.908	**0.751**	0.603	**0.532**	**0.069**

DualRef-TTS substantially improves both generation quality and reference attribution, achieving the highest ΔT-SIM of 0.522 and the only positive ΔS-SIM of 0.078 among reference-based systems. StyleTTS2 and IndexTTS2 obtain competitive timbre similarity but show smaller or negative style attribution margins, indicating residual style–timbre coupling. Prompt-based systems, including CosyVoice3 and Qwen3-TTS, preserve content and timbre well but achieve lower style similarity, suggesting that text instructions remain limited for subtle and personalized style transfer. Within our framework, the CFM variant achieves better WER and T-SIM, whereas MeanFlow provides higher DNSMOS and S-SIM with substantially lower latency. We therefore adopt MeanFlow as the main operating point of DualRef-TTS, since it provides the best overall balance among style fidelity, perceptual quality, disentanglement, and efficiency.

Subjective Results As shown in Table 2, DualRef-TTS achieves the highest SS-SIM, ST-SIM, and S-MOS among all systems. These improvements indicate that reference-driven style control provides more effective open-set style transfer while preserving speaker identity. The results further support that the proposed dual-reference design reduces style-timbre interference and supports more independent control.

5.2 Controlled Same-Backbone Disentanglement Comparison

To isolate the effect of style representation from architecture and training, all variants share the same CFM DiT backbone, BAPC interface, training data, training epochs, and all hyperparameters; only the style representation is replaced by alternatives derived from StyleTTS2, IndexTTS2, and Vevo2. Evaluation uses the same 2,048-utterance swapping grid as Table 1.

As shown in Table 3, our attribute-perturbation (AP) representation achieves the best WER of 0.941 and T-SIM of 0.751, together with the largest ΔT-SIM of 0.532 and a positive ΔS-SIM of 0.069. StyleTTS2 obtains a positive style margin, but its lower T-SIM indicates residual speaker information in the style representation. IndexTTS2 preserves timbre similarity well, yet produces a negative ΔS-SIM, suggesting that its generated style remains coupled with the timbre reference. Vevo2-CS (same-utt) achieves apparently strong style attribution but suffers from extremely high WER due to content copying. After removing content overlap, Vevo2-CS and Vevo2-Prosody show degraded style similarity and negative ΔS-SIM, indicating incomplete independent style control. Overall, AP provides the strongest valid timbre–style disentanglement under an identical backbone.

To further verify whether the objective disentanglement differences are perceptually observable, we conduct a subjective evaluation under the same-backbone setting. As shown in Table 4, Ours (CFM) achieves the highest ST-SIM and SS-SIM among all variants, while maintaining competitive

Table 4: Subjective evaluation on style-controllable TTS.

System	S-MOS ↑	ST-SIM ↑	SS-SIM ↑
StyleTTS2	3.90 ± 0.07	3.23 ± 0.06	3.12 ± 0.06
IndexTTS2	$\mathbf{3.95 \pm 0.09}$	3.62 ± 0.07	3.03 ± 0.05
Vevo2-CS (non-overlap)	3.88 ± 0.09	3.50 ± 0.06	3.06 ± 0.05
Vevo2-Prosody	3.28 ± 0.07	3.42 ± 0.06	3.14 ± 0.06
Ours (CFM)	3.93 ± 0.08	$\mathbf{3.69 \pm 0.06}$	$\mathbf{3.86 \pm 0.07}$

S-MOS. These results are consistent with the objective T-SIM and S-SIM trends, indicating that the improved timbre–style separation is perceptually observable rather than only reflected by automatic metrics.

Table 5: Ablation analysis of key components in the MeanFlow-based framework.

Model Variant	WER (%) ↓	DNSMOS ↑	T-SIM ↑	S-SIM ↑
Base	1.483	3.911	0.713	0.598
Base w/o BAPC	2.084	3.897	0.664	0.633
Base w/o perturbations	129.461	**3.953**	0.213	**0.893**
– w/o pitch perturbation	1.889	3.878	0.462	0.612
– w/o content perturbation	172.548	3.075	0.504	0.529
– w/o timbre perturbation	13.051	3.352	0.223	0.581
Full (Base+DPO)	**0.984**	3.946	**0.736**	0.624

5.3 Ablation on the MeanFlow Backbone

Table 5 presents ablations on the MeanFlow backbone. Removing BAPC increases WER from 1.483 to 2.084 and reduces T-SIM from 0.713 to 0.664, while slightly increasing S-SIM. This trade-off suggests possible timbre leakage from the style reference, indicating that BAPC provides a useful structural bias for balancing style transfer and timbre preservation rather than merely maximizing style similarity. Additional analyses are provided in Appendix C.

For the attribute perturbation module (Du et al., 2021; Gengembre et al., 2024), the ablation results reveal a clear division of roles among the three components. Timbre perturbation provides the strongest suppression of speaker identity, reducing T-SIM from 0.713 to 0.223, while pitch perturbation mainly removes the remaining speaker-related cues with only a limited effect on style. By contrast, content perturbation contributes most to the loss of style similarity, as removing it drops S-SIM from 0.598 to 0.529. This confirms that attribute perturbation is necessary to prevent the model from directly following the style reference and to achieve disentangled dual-reference control.

For DPO post-training, our results demonstrate that DPO yields a substantial reduction in WER and consistent improvements across DNSMOS, T-SIM and S-SIM metrics. These results validate that preference alignment via DPO effectively boosts overall generation quality, which further enhances the structural advantages brought by BAPC conditioning and attribute purification modules.

6 Conclusion

We introduced DualRef-TTS, a label-free dual-reference framework for zero-shot TTS that enables more independent control of style and timbre. The framework combines attribute perturbation, BAPC, MeanFlow-based acoustic generation, and DPO post-training to support controllable cross-speaker style transfer while preserving semantic accuracy, timbre consistency, and synthesis quality. Experiments show that DualRef-TTS remains competitive with strong baselines on objective and subjective metrics, while providing a more direct and flexible form of style control than approaches based on style labels, prompt descriptions, or compressed intermediate representations. These results support modeling style as a label-free, continuous, and reference-guided variable for open-set zero-shot speech synthesis. Limitations and responsible-use considerations are discussed in Appendices G.1 and G.2.

AI USE STATEMENT

We used generative AI tools, including OpenAI Codex, to assist with literature discovery, software-code development and debugging, and LaTeX typesetting and formatting. Candidate references suggested by AI tools were independently checked by the authors against the original publications before being included in the manuscript. AI-assisted code was manually reviewed, executed, and tested by the authors, and the resulting experimental outputs were independently verified. The final research design, analysis, interpretation, scientific claims, and conclusions were determined and verified by the authors. We take full responsibility for the final content of this work, including all text, code, claims, results, references, and artifacts.

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.

Bingsong Bai, Fengping Wang, Yingming Gao, and Ya Li. Spa-svc: Self-supervised pitch augmentation for singing voice conversion. *arXiv preprint arXiv:2406.05692*, 2024.

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.

Thomas Bott, Florian Lux, and Ngoc Thang Vu. Controlling emotion in text-to-speech with natural language prompts. *arXiv preprint arXiv:2406.06406*, 2024.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International conference on machine learning*, pp. 2709–2720. PMLR, 2022.

Chak Ho Chan, Kaizhi Qian, Yang Zhang, and Mark Hasegawa-Johnson. Speechsplit2. 0: Unsupervised speech disentanglement for voice conversion without tuning autoencoder bottlenecks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6332–6336. IEEE, 2022.

Sanyuan Chen, Shujie Liu, Long Zhou, Yanqing Liu, Xu Tan, Jinyu Li, Sheng Zhao, Yao Qian, and Furu Wei. Vall-e 2: Neural codec language models are human parity zero-shot text to speech synthesizers. *arXiv preprint arXiv:2406.05370*, 2024.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. An enhanced res2net with local and global feature fusion for speaker verification. *arXiv preprint arXiv:2305.12838*, 2023.

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, 2025.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. *arXiv preprint arXiv:2406.07803*, 2024.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, and Seong-Whan Lee. Diemo-tts: Disentangled emotion representations via self-supervised distillation for cross-speaker emotion transfer in text-to-speech. *arXiv preprint arXiv:2505.19687*, 2025.

Hyeong-Seok Choi, Juheon Lee, Wansoo Kim, Jie Lee, Hoon Heo, and Kyogu Lee. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. *Advances in Neural Information Processing Systems*, 34:16251–16265, 2021.

Hsing-Hang Chou, Yun-Shao Lin, Ching-Chin Sung, Yu Tsao, and Chi-Chun Lee. Diffevc: Any-to-any emotion voice conversion with expressive guidance. *arXiv preprint arXiv:2409.03636*, 2024.

Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. A review on subjective and objective evaluation of synthetic speech. *Acoustical Science and Technology*, 45(4):161–183, 2024.

Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang. Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system. *arXiv preprint arXiv:2502.05512*, 2025. doi: 10.48550/arXiv.2502.05512. URL https://arxiv.org/abs/2502.05512.

Yimin Deng, Jianzong Wang, Xulong Zhang, Ning Cheng, and Jing Xiao. Learning expressive disentangled speech representations with soft speech units and adversarial style augmentation. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2024.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

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

Zongyang Du, Berrak Sisman, Kun Zhou, and Haizhou Li. Disentanglement of emotional style and speaker identity for expressive voice conversion. *arXiv preprint arXiv:2110.10326*, 2021.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. *arXiv preprint arXiv:2206.08317*, 2022.

Chen Geng, Meng Chen, Ruohua Zhou, Ruolan Liu, and Weifeng Zhao. Poly-svc: Polyphony-aware singing voice conversion with harmonic modeling. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Nicolas Gengembre, Olivier Le Blouch, and Cédric Gendrot. Disentangling prosody and timbre embeddings via voice conversion. In *INTERSPEECH*, 2024.

Yicheng Gu, Xueyao Zhang, Liumeng Xue, and Zhizheng Wu. Multi-scale sub-band constant-q transform discriminator for high-fidelity vocoder. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10616–10620. IEEE, 2024.

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.

Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5901–5905. IEEE, 2019.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*, 2025.

Sho Inoue, Kun Zhou, Shuai Wang, and Haizhou Li. Hierarchical control of emotion rendering in speech synthesis. *IEEE Transactions on Affective Computing*, 2025.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10301–10305. IEEE, 2024.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. *arXiv preprint arXiv:2403.03100*, 2024.

Minki Kang, Wooseok Han, Sung Ju Hwang, and Eunho Yang. Zet-speech: Zero-shot adaptive emotion-controllable text-to-speech synthesis with diffusion and style-based models. *arXiv preprint arXiv:2305.13831*, 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.

Haoxun Li, Leyuan Qu, Jiaxi Hu, and Taihao Li. Eme-tts: Unlocking the emphasis and emotion link in speech synthesis. *arXiv preprint arXiv:2507.12015*, 2025a.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

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. *Advances in neural information processing systems*, 36:19594–19621, 2023b.

Yinghao Aaron Li, Cong Han, and Nima Mesgarani. Styletts: A style-based generative model for natural and diverse text-to-speech synthesis. *IEEE Journal of Selected Topics in Signal Processing*, 19(1):283–296, 2025b.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

Zhonghua Liu, Shijun Wang, and Ning Chen. Automatic speech disentanglement for voice conversion using rank module and speech augmentation. *arXiv preprint arXiv:2306.12259*, 2023.

Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017.

Hui Lu, Xixin Wu, Zhiyong Wu, and Helen Meng. Speechtriplenet: End-to-end disentangled speech representation learning for content, timbre and prosody. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 2829–2837, 2023.

Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11341–11345. IEEE, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, et al. Vibevoice technical report. *arXiv preprint arXiv:2508.19205*, 2025.

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*, pp. 8599–8608. PMLR, 2021.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. Openvoice: Versatile instant voice cloning. *arXiv preprint arXiv:2312.01479*, 2023.

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*, pp. 28492–28518. PMLR, 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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497. IEEE, 2021.

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. *arXiv preprint arXiv:2006.04558*, 2020.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8228–8238, 2024.

Chengyi Wang, Sanyuan Chen, 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. *arXiv preprint arXiv:2301.02111*, 2023.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*, 2024.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025a.

Zhichao Wu, Yueteng Kang, Songjun Cao, Long Ma, Qiulin Li, and Qun Yang. Mpe-tts: Customized emotion zero-shot text-to-speech using multi-modal prompt. *arXiv preprint arXiv:2505.18453*, 2025b.

Yuying Xie, Michael Kuhlmann, Frederik Rautenberg, Zheng-Hua Tan, and Reinhold Haeb-Umbach. Speaker and style disentanglement of speech based on contrastive predictive coding supported factorized variational autoencoder. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pp. 436–440. IEEE, 2024.

Jingyuan Xing, Zhipeng Li, Shuaiqi Chen, Xiaofen Xing, and Xiangmin Xu. Eats-speech: Emotion-adaptive transformation and priority synthesis for zero-shot text-to-speech. In *Proc. Interspeech 2025*, pp. 4358–4362, 2025.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024.

Xueyao Zhang, Yuancheng Wang, Chaoren Wang, Ziniu Li, Zhuo Chen, and Zhizheng Wu. Advancing zero-shot text-to-speech intelligibility across diverse domains via preference alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12251–12270, 2025a.

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu. Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning. *arXiv preprint arXiv:2508.16332*, 2025b.

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, et al. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement. *arXiv preprint arXiv:2502.07243*, 2025c.

Qixi Zheng, Yushen Chen, Zhikang Niu, Ziyang Ma, Xiaofei Wang, Kai Yu, and Xie Chen. Accelerating flow-matching-based text-to-speech via empirically pruned step sampling. *arXiv preprint arXiv:2505.19931*, 2025.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35139–35148, 2026.

A Experimental Setup and Implementation Details

A.1 Training Datasets

As shown in Table 6, the training data used in our experiments consist of a high-quality filtered subset of Emilia and several public emotional speech datasets. These datasets jointly cover Chinese and English. The original Emilia corpus is filtered using multiple criteria, including text-audio alignment accuracy, acoustic quality, speaker pronunciation consistency, and the presence of distinguishable emotional expressions. After filtering and integrating the additional emotional and public speech corpora, the final training corpus comprises approximately 6,000 hours of speech, providing a rich and reliable source of emotional and stylistic variation for model training and controllable generation.

Table 6: Summary of training datasets used in our experiments.

Dataset	Language	Annotation / Description	Duration	Sampling Rate
Emilia filtered subset	ZH / EN	High-quality diverse-style speech after multi-stage filtering	∼5500 h	24 kHz
TextrolSpeech	EN	Natural text style descriptions with five style factors: gender, pitch, speaking speed, volume, and emotion	∼330 h	24 kHz
IEMOCAP	EN	Dyadic scripted/spontaneous emotional speech with categorical descriptors and activation-valence-dominance annotations	∼12 h	48 kHz
MSP-Podcast	EN	Podcast speech with categorical emotions and valence-arousal-dominance annotations	∼160 h	16 kHz
Total	ZH / EN	–	∼6000 h	–

TextrolSpeech is a large-scale emotional speech dataset, with a total duration of approximately 330 hours. It provides fine-grained descriptions of emotional intensity and textual attributes, enabling more precise controllable speech generation. IEMOCAP contains approximately 12 hours of English emotional speech from 10 speakers and covers a wide range of emotional expressions; it is

Table 7: Taxonomy of style-controllable TTS paradigms under a unified conditional generation framework.

Paradigm	Model Signal	Generation Mapping	Representative Models
Label-guided	Discrete labels l	$[x, t, l] \rightarrow \hat{M}$	ZET-Speech, EME-TTS
Vector-guided	Continuous vector v	$[x, t, v] \rightarrow \hat{M}$	EmoSphere++
Instruction-guided	Natural language prompts p	$[x, t, p] \rightarrow \hat{M}$	CosyVoice3, Qwen3-TTS
Speech-guided	Reference speech r	$[x, t, r] \rightarrow \hat{M}$	StyleTTS2, Vevo, IndexTTS2

widely used for emotional speech modeling and evaluation. MSP-Podcast is a large-scale natural emotional speech dataset collected from real-world podcast scenarios. It contains more than 160 hours of speech from over 1,000 speakers and provides dimensional emotion annotations, which help introduce more natural and open-domain speaking styles.

Overall, the training data used in this study comprise approximately 6,000 hours of speech. The filtered Emilia subset forms the basis of general speech modeling, while the public emotional speech datasets and additional filtered corpora further strengthen the model's ability to capture emotional expression, style variation, and natural speaking patterns.

A.2 Data Processing and Filtering Pipeline

To ensure the quality of training data and the reliability of emotional expression, we construct a multi-stage data processing and filtering pipeline. The pipeline includes audio preprocessing, text-consistency verification, acoustic-quality filtering, emotion and style relevance assessment, and manual review. The goal is to select high-quality samples with accurate text, clear pronunciation, stable acoustic quality, and distinguishable emotional or stylistic expression from large-scale raw speech data.

First, raw speech is standardized through preprocessing, including resampling, volume normalization, silence trimming, and abnormal-duration filtering. Audio samples that are too short, too long, severely truncated, or affected by obvious background noise are removed from the candidate set. Next, an automatic speech recognition model is used to transcribe the audio, and WER or CER is computed between the transcription and the original text to verify content consistency. Samples with obvious deletions, substitutions, insertions, or text-audio mismatch are filtered out to reduce the impact of content errors during training.

For acoustic quality, we combine predicted MOS scores with commonly used acoustic-quality indicators. Specifically, we examine naturalness, clarity, signal-to-noise ratio, silence ratio and abnormal loudness. Samples with clear noise, distortion, clipping, abnormal volume, or unstable pronunciation are removed. This step improves the overall audibility and stability of the training data.

To further ensure the richness and reliability of emotional and stylistic expression, we perform an automatic style-emotion filtering stage based on existing open-source audio understanding models and multi-modal LLM scoring. Specifically, each candidate utterance is analyzed with an audio understanding model to provide audio-conditioned evidence about emotion, prosody, speaking manner, and vocal expression. A multi-modal LLM then assigns scores to the utterance in terms of style expressiveness and emotional richness. Samples with low scores, such as those with flat prosody, weak emotional cues or insufficient expressive variation, are removed from the training pool. This filtering stage helps retain speech samples with clearer and richer style/emotion characteristics, thereby improving the quality of the data used for controllable style and emotion modeling.

After applying the above multi-stage filtering pipeline to the Emilia corpus, we obtain approximately 5,500 hours of high-quality Chinese-English speech with reliable content alignment, stable acoustic quality, and richer emotional or stylistic expression. This filtered Emilia subset serves as the primary training corpus and is further combined with public emotional speech datasets, resulting in approximately 6,000 hours of training data in total.

A.3 Representative Disentangled Control Systems

To complement the discussion in Section 2, Table 7 summarizes representative style-controllable TTS paradigms under a unified conditional generation framework. Existing systems condition generation on discrete emotion or style labels l, continuous affective vectors v, natural-language prompts p, or speech references r. The table further provides their corresponding generation mappings and representative models. Here, x denotes the input text, t the timbre reference speech, and $\hat{M}$ the generated mel-spectrogram.

To ensure representative and fair comparisons, we select baseline systems according to the following principles. First, the baselines should cover different TTS generation paradigms, including autoregressive and non-autoregressive models, so that content accuracy, naturalness, timbre preservation, and style controllability can be compared across different modeling approaches. Second, the selected systems should provide public implementations, pretrained checkpoints, or stable inference interfaces, and should be runnable under a unified evaluation setting. Third, the baselines should exhibit strong performance in content accuracy, speech naturalness, speaker similarity, or style control, making the comparison meaningful. Finally, we prioritize systems that support timbre, emotion, style, or instruction-based control, so as to analyze their behavior under style–timbre control scenarios.

Following these principles, we select F5-TTS, StyleTTS2, IndexTTS2, CosyVoice3, Vevo2 and Qwen3-TTS as the main baseline systems. IndexTTS2, CosyVoice3, Vevo2 and Qwen3-TTS are treated as autoregressive or language-model-based speech generation systems, while F5-TTS and StyleTTS2 are representative non-autoregressive TTS systems.

Although F5-TTS does not directly support the dual-reference timbre–style control setting used in this paper, it is a representative non-autoregressive TTS model, and its DiT and flow-matching modeling paradigm is related to our acoustic generation backbone. We therefore include it as an important reference for comparing content accuracy, speech naturalness, and inference efficiency under the non-autoregressive generation paradigm. The baseline systems are summarized as follows:

- **F5-TTS.** F5-TTS is a non-autoregressive TTS system based on Diffusion Transformer and flow matching. It supports zero-shot speech synthesis, expressive speech generation, and speaking-rate control. Trained on a large-scale multilingual speech corpus, it enables expressive zero-shot synthesis and code-switching generation. In our experiments, we use the officially released pretrained checkpoint for inference.
- **Vevo2.** Vevo2 achieves architectural specialization through two-stage modeling itself, injecting specific information at corresponding stages: an AR content-style stage generates content-style tokens prompted by a style reference, followed by a flow-matching acoustic stage that reconstructs acoustics prompted by a timbre reference. Style must cross stages via lossy VQ discretization, while timbre enters only at the second stage; disentanglement is achieved precisely through this pipeline decomposition. This design is architecturally incompatible with our single-stage conditioning path that models style as a continuous reference-driven variable without a discrete intermediate bottleneck, making it difficult to integrate Vevo2 as a modular component.
- **IndexTTS2.** IndexTTS2 is a zero-shot speech synthesis system based on a GPT-style architecture. It introduces an index-based semantic alignment mechanism and combines it with improved speaker-conditional modeling and a BigVGAN vocoder to produce high-fidelity audio. The system supports multilingual zero-shot speech generation and shows stable performance in emotion and timbre control.
- **CosyVoice3.** CosyVoice3 combines the cross-dimensional modeling capability of large language models with high-fidelity waveform generation based on flow matching. The system reconstructs waveforms directly from code sequences predicted by the language model, reducing the need for a separate acoustic model. This design simplifies the generation pipeline while maintaining high naturalness and audio quality, and supports multilingual and cross-lingual speech synthesis.
- **StyleTTS2.** StyleTTS2 introduces a style diffusion mechanism to capture subtle and diverse prosodic variations by modeling the stochastic distribution of the latent style space. It further incorporates adversarial training with a large-scale speech language model, where

semantic priors are used to improve the expressiveness of synthesized speech. Its non-autoregressive design enables efficient inference while maintaining high naturalness and supporting robust zero-shot style transfer.

- **Qwen3-TTS.** Qwen3-TTS is a multilingual, controllable, and streaming text-to-speech model in the Qwen series. It supports natural-language-description-based speech control and short-prompt voice cloning, enabling the generation of speech with specified style, emotion, and speaker characteristics. The model maintains high fidelity and strong multilingual and cross-lingual generation capability, and shows good stability and naturalness in long-form speech generation. We use Qwen3-TTS as an important benchmark for voice cloning, emotional expression, and controllable generation performance.

To ensure fair comparison, all baseline systems are evaluated under the same testing conditions whenever possible, and the generated samples are assessed using the same objective and subjective evaluation protocols.

A.4 Model Configuration

All experiments use the same DiT-based acoustic architecture unless otherwise specified. The DiT backbone operates on 100-dimensional mel-spectrogram frames. At each flow time step, the input representation is formed by concatenating the current noisy mel state, the masked acoustic condition, and the text embedding, followed by a linear projection to a 1024-dimensional hidden space. A convolutional positional embedding is then applied, and the branch embedding is added to the hidden representation. Text is tokenized with a large-vocabulary text tokenizer and embedded into a 512-dimensional representation; four ConvNeXt-style convolutional blocks are used on the text side to enhance local contextual modeling. The backbone contains 22 DiT blocks with 16 attention heads, head dimension 64, a feed-forward expansion ratio of 2, rotary positional embeddings for self-attention, and adaptive layer normalization conditioned on the flow time embedding. The output projection maps hidden states back to a 100-dimensional mel/flow prediction.

For acoustic feature extraction, all audio is represented as 100-channel mel-spectrograms at 24 kHz, with an FFT size of 1024, a window length of 1024, a hop length of 256, and a Vocos-compatible mel frontend. The generated mel-spectrogram is decoded into waveform using a pretrained Vocos-style vocoder. This model configuration is shared across the main CFM, MeanFlow, ablation, and post-training experiments, so differences in results primarily reflect the evaluated training objectives or conditioning components rather than changes in the acoustic backbone.

A.5 Training and Inference Setup

The model is trained with AdamW using an initial learning rate of 9.5×10^{-5}. We apply a linear warmup schedule for the first 1000 optimization updates, followed by a cosine annealing schedule with warm restarts, where the minimum learning rate is set to 0.1 times the initial learning rate. Training uses sample-based batching with a per-GPU batch size of 32. Gradient accumulation is enabled according to the training configuration, and gradient clipping is applied with a maximum gradient norm of 1.0.

During training, each example is organized as a concatenated mel-spectrogram sequence containing a timbre reference, the target utterance, and a style reference. The target region is masked in the acoustic condition and optimized with the flow/mean-flow reconstruction loss, while the timbre and style reference regions remain visible as conditioning signals. Classifier-free guidance is supported during training by randomly dropping the acoustic condition with probability 0.3 and jointly dropping both acoustic and text conditions with probability 0.2.

During inference, the dual-reference generation interface takes a target text, a timbre reference utterance, and a style reference utterance. The reference waveforms are converted to mono when necessary, RMS-normalized when their energy is below 0.1, resampled to 24 kHz, and converted into 100-dimensional mel-spectrograms with a hop length of 256. The model constructs a three-branch sequence in the order of timbre reference, target generation region, and style reference. The target region is initialized from Gaussian noise and generated by numerical ODE integration conditioned on the visible timbre and style references.

Unless specified, CFM inference uses 32 Euler ODE steps and classifier-free guidance strength 2.0. For the MeanFlow variant, we use the 4-step solver selected according to the step-efficiency analysis. The generated mel-spectrogram is finally decoded into waveform using the pretrained Vocos-style vocoder.

B Evaluation Protocols and Style Definitions

B.1 Definition of Style

In this paper, style refers to the high-level expressive manner of speech beyond textual content and speaker identity. It mainly covers emotional state, speaking attitude, and prosodic organization. Different styles may be reflected through acoustic cues such as intonation contour, rhythm, pause pattern, speaking rate, and energy dynamics. These low-level cues are not treated as separate style categories in this paper; instead, they are regarded as observable manifestations that jointly realize a higher-level speaking style.

The boundaries among style, emotion, prosody, pitch, speaking rate, and speaker habit are often gradual rather than sharp. In natural speech, these factors are frequently coupled: for example, an emotional change may be accompanied by variations in pitch, energy, rhythm, or pause placement. Therefore, we treat style as an open-ended and high-level expressive concept that describes dynamic speech characteristics beyond content and timbre.

B.2 Style Description Lists for Evaluation

To evaluate instruction-based baseline models under complex open-style control scenarios, we convert the style reference utterances used in evaluation into natural-language style descriptions and feed them to CosyVoice3 and Qwen3-TTS for generation. Unlike simple emotion labels, these descriptions cover mixed emotions, intensity variation, prosodic patterns, phonation manner, and character-like expressions. This setup allows us to evaluate whether instruction-based systems can understand complex style descriptions and perform fine-grained controllable generation.

We use two groups of style references. As shown in Table 8, the first group contains 16 basic styles, covering common emotions, tones, and phonation manners; it is used for objective evaluation. As shown in Table 9, the second group is used for subjective evaluation to assess model generalization under complex emotional expression, prosodic variation, and phonation control. For the reference-based model, style is provided by reference audio. For instruction-based baseline models, the corresponding style references are transcribed into natural-language prompts so that different systems receive style targets that are as consistent as possible.

Table 8: Basic style descriptions used in the objective evaluation.

ID	Style	Description
1	Neutral	A neutral tone with moderate speaking rate and stable intonation.
2	Happy	A cheerful tone with bright intonation and a lively rhythm.
3	Sad	A sad and low tone with slower speaking rate.
4	Angry	A firm and angry tone with stronger emphasis and slightly faster speaking rate.
5	Surprise	A surprised tone with rising intonation and noticeable rhythm variation.
6	Fearful	A nervous and fearful tone with unstable pacing and more pauses.
7	Disgust	A disgusted or rejecting tone with cold intonation and a clear negative attitude.
8	Cry	A crying tone with slower speaking rate and clear pauses.
9	Whisper	A soft whispering voice with lower volume and slower speaking rate.
10	Deep	A deep and steady tone with stable intonation and regular rhythm.
11	Roar	A roaring or shouting style with high volume and strong vocal force.
12	Calm	A calm and relaxed tone with slightly slow speaking rate and stable intonation.
13	Serious	A serious and formal tone with clear articulation and stable rhythm.
14	Gentle	A gentle and friendly tone with soft intonation and moderate speaking rate.

ID	Style	Description
15	Question	A questioning tone with naturally rising sentence endings and moderate rhythm.
16	Excited	An excited and positive tone with slightly faster speaking rate and clear intonation variation.

Table 9: Challenging open-style descriptions used in the subjective evaluation

ID	Style	Description
1	Hesitant Happy	Happy but with slight hesitation and a mild broadcast tone, with bright intonation and moderate speaking rate.
2	Restrained Sad	Sad but restrained, with low intonation, slow speaking rate, natural pauses, and medium emotional intensity.
3	Gentle Question	A gentle and polite questioning tone, with slightly rising endings and moderately slow speaking rate.
4	Curious Surprise	Surprised with slight curiosity, with naturally rising intonation and medium-high emotional intensity.
5	Suppressed Anger	Angry but deliberately suppressed, with a firm tone, moderately fast speaking rate, and controlled intensity.
6	Tired Sad	Tired with a sense of sadness, with low intonation, slow speaking rate, weak energy, and medium emotional intensity.
7	Warm Comfort	A warm and comforting tone, with soft intonation, slow speaking rate, natural pauses, and low emotional intensity.
8	Broadcast Surprise	A surprised tone with a slight broadcast style, clear articulation, obvious intonation variation, and moderate speaking rate.
9	Whisper Sad	A soft whispering tone with sadness, low volume, slow speaking rate, and medium-low emotional intensity.
10	Whisper Urgent	A whispering tone with urgency, low volume, moderately fast speaking rate, short pauses, and tense emotion.
11	Calm Confidence	Calm but confident, with stable intonation, moderate speaking rate, regular rhythm, and low emotional intensity.
12	Serious Narration	A serious narrative tone, with stable intonation, moderate speaking rate, clear articulation, and restrained emotion.
13	Deep Comfort	Deep but comforting, with low and stable intonation, slow speaking rate, and medium emotional intensity.
14	Playful Teasing	A playful and teasing tone, with natural intonation variation, moderately fast speaking rate, and medium emotional intensity.
15	Cold Indifference	A cold and distant tone, with flat intonation, moderate speaking rate, and low emotional intensity.
16	Fearful Hesitation	Fearful with hesitation, with unstable intonation, slow speaking rate, frequent pauses, and high emotional intensity.
17	Excited Anticipation	Excited with anticipation, with bright intonation, moderately fast speaking rate, and high emotional intensity.
18	Soft Apology	A soft apologetic tone, with low intonation, slow speaking rate, restrained emotion, and a sense of regret.
19	Roaring Anger	Roaring anger, with high volume, strong vocal force, fast speaking rate, and very high emotional intensity.
20	Nervous Explanation	A nervous explanatory tone, with slightly fast speaking rate, uneven pauses, unstable intonation, and medium emotional intensity.
21	Child-like Delight	A child-like delighted tone, with higher pitch, lively rhythm, and high emotional intensity.
22	Warm Encouragement	A warm and encouraging tone, with open intonation, moderate speaking rate, positive emotion, and medium intensity.
23	Calm Sadness	Calm with sadness, with low intonation, slow speaking rate, inward emotion, and medium-low intensity.
24	Tense Urgency	A tense and urgent tone, with fast speaking rate, short pauses, tight intonation, and high emotional intensity.

ID	Style	Description
25	Doubtful Surprise	Surprised with doubt, with slightly rising intonation, moderate speaking rate, and medium-high emotional intensity.
26	Gentle Happiness	Gentle happiness, with bright but not exaggerated intonation, moderate speaking rate, and medium emotional intensity.

B.3 Subjective Evaluation Protocol

We conduct subjective evaluation on generated speech using a five-point mean opinion score protocol. The evaluation contains three dimensions: subjective overall speech quality, subjective style similarity, and subjective timbre similarity. We denote them as S-MOS, SS-SIM, and ST-SIM, respectively.

Table **??** shows the instructions used for subjective overall speech quality evaluation. Table 10 shows the instructions used for subjective style similarity evaluation. Table 11 shows the instructions used for subjective timbre similarity evaluation.

We invite 24 bilingual Chinese–English participants to take part in the subjective evaluation. To improve reliability and reduce listening fatigue, the evaluation is conducted in grouped listening sessions. Each session contains 10–12 speech samples, and participants are given 5–10 seconds after each sample to record their scores. After 20–25 minutes of continuous listening, participants take a break of approximately 10 minutes. During the experiment, participants are asked to wear headphones in a quiet environment and adjust the volume to a comfortable level before the evaluation begins. The volume is kept unchanged within each evaluation session. To reduce subjective bias, all speech samples are randomly shuffled, and participants are not informed of the model identity corresponding to each sample. All participants are ordinary native speakers or evaluators who have received a brief task instruction before the evaluation.

B.4 Explanation of Similarity Scores

The absolute values of S-SIM are conservative across all systems, but the metric remains informative for relative comparison. In particular, S-SIM still reveals clear performance gaps across both models and style categories. DualRef-TTS consistently ranks above the other baselines, while the lowest proxy scores are concentrated on styles such as whisper and roar.

This behavior is expected because S-SIM is computed from a pretrained emotion representation model rather than from a universal style evaluator. Therefore, it is only partially aligned with styles expressed mainly through phonation mode, vocal effort, or character-like delivery. For example, the roar style relies heavily on strong vocal effort and special phonation patterns. Even if two samples both contain clear roaring characteristics, their S-SIM score may remain moderate; when the generated speech lacks the roaring characteristic almost entirely, the S-SIM score drops further. Therefore, we treat S-SIM as a conservative but still discriminative proxy for open-set style similarity and use it primarily for relative comparison rather than absolute interpretation. This also explains why the advantage of DualRef-TTS is more pronounced in human evaluation.

C Additional Validation of Style Reference Construction

This section provides three complementary analyses of the style-reference pathway. We first quantify the attributes retained by the processed style references, then evaluate the effect of inference-side reference processing, and finally examine reference length, excerpt position, and temporal structure. Unless otherwise specified, all experiments use the metrics and evaluation protocols defined in Section 4.3. WER is macro-averaged over utterances and includes insertion errors; it can therefore exceed 100% when the evaluated speech contains repeated or mismatched content.

C.1 Signal-Level Attribute-Perturbation Analysis

We randomly sample 1,000 training utterances and evaluate the processed style references themselves, without involving the generation model. Table 12 measures residual speaker identity using

Table 10: Subjective style similarity instruction.

Introduction
Your task is to evaluate how similar two speech recordings sound in terms of speaking style. Style may be reflected by emotion, intonation, rhythm, speaking rate, pauses, and overall expressive manner.

Task Instructions
In this task, you will hear two speech samples. Please focus on whether the two samples express a similar speaking style, rather than whether they have the same speaker identity. You should give a score for Sample B relative to Sample A.

Scoring scale:

- 5: Excellent
- 4: Good
- 3: Fair
- 2: Poor
- 1: Bad

Table 11: Subjective timbre similarity instruction.

Introduction
Your task is to evaluate how similar two speech recordings sound in terms of speaker timbre, including voice color, vocal tone, and perceived speaker identity.

Task Instructions
In this task, you will hear two speech samples. Please focus on whether the two samples sound as if they are spoken by the same speaker, regardless of text content, emotion, or speaking style. You should give a score for Sample B relative to Sample A.

Scoring scale:

- 5: Excellent
- 4: Good
- 3: Fair
- 2: Poor
- 1: Bad

T-SIM, retained style using S-SIM, residual linguistic content using WER, and signal quality using DNSMOS. GT denotes the original unprocessed speech. SVC-only denotes speech passed through the SVC resynthesis pipeline without any attribute perturbation, serving as a control for degradation introduced by resynthesis itself. Full denotes our complete style-reference construction pipeline with timbre, content, and pitch perturbations all enabled. The three w/o conditions are leave-one-out variants of Full, with the indicated perturbation disabled.

Table 12: Attribute perturbation.

Perturbation	T-SIM ↓	S-SIM ↑	WER (%) ↑	DNSMOS ↑
GT	–	–	5.2	3.59
SVC-only	0.850	0.910	5.9	3.58
w/o timbre	0.504	0.529	172.5	3.07
w/o content	0.223	0.581	13.0	3.35
w/o pitch	0.224	0.524	115.1	3.16
Full	**0.174**	0.506	114.0	3.02

GT and SVC-only have nearly identical WER and DNSMOS, showing that SVC resynthesis itself introduces little degradation. The leave-one-out rows isolate the contribution of each perturbation. Removing timbre perturbation increases T-SIM from 0.174 to 0.504, indicating substantial residual speaker identity. Removing content perturbation reduces WER from 114.0% to 13.0% and increases

S-SIM from 0.506 to 0.581. Content removal therefore accounts for most of the observed decrease in the style-similarity score. Removing pitch perturbation increases S-SIM by only 0.018, but also increases T-SIM from 0.174 to 0.224. Pitch perturbation thus removes additional identity cues while contributing relatively little to the loss of measurable style. The full pipeline yields the lowest T-SIM and a high WER, demonstrating the strongest joint suppression of speaker identity and linguistic content. Its S-SIM remains 0.506 despite the accompanying DNSMOS decrease from 3.58 to 3.02, which may cause the style evaluator to underestimate the retained style information.

C.2 Inference-Side Reference Processing

We next evaluate whether the signal degradation introduced by attribute perturbation affects downstream synthesis. At inference time, the original style reference is replaced with its perturbed and resynthesized counterpart, while the timbre reference, target text, model, and decoding configuration remain unchanged. Table 13 reports content accuracy, output quality, timbre preservation, and style similarity under the two reference conditions.

Table 13: Inference-side reference processing.

Style Reference	WER (%) ↓	DNSMOS ↑	T-SIM ↑	S-SIM ↑
Original audio	**0.984**	**3.946**	0.736	**0.624**
Perturbed and resynthesized	1.021	3.921	**0.745**	0.616

Although the full perturbation pipeline reduces the DNSMOS of the style reference from 3.58 for SVC-only resynthesis to 3.02, this degradation has little effect on the synthesized output. Replacing the original style reference with the perturbed and resynthesized reference decreases output DNSMOS by only 0.025, from 3.946 to 3.921, and S-SIM by only 0.008, from 0.624 to 0.616. WER changes slightly from 0.984% to 1.021%, while T-SIM increases from 0.736 to 0.745. Thus, despite the substantial reference-level degradation introduced by the full perturbation pipeline, the final synthesis quality, content accuracy, timbre similarity, and style similarity remain largely unchanged. This result indicates that the degradation intentionally introduced during style-reference construction does not substantially propagate to the synthesized speech.

C.3 Reference Length, Excerpt Position, and Temporal Structure

We examine the effect of reference duration and excerpt position on style transfer. Style references of at least 6 seconds are truncated to their first 2 seconds, first 5 seconds, or retained at full length. For the position analysis, 3-second windows are taken from the beginning, middle, and end of the same reference. All other generation conditions remain fixed. S-SIM measures transferred style information, whereas T-SIM measures preservation of the designated timbre reference.

Table 14: Reference length and position.

Analysis	Condition	S-SIM ↑	T-SIM ↑
Length	First 2 s	0.560	**0.759**
	First 5 s	0.581	0.746
	Full reference	**0.630**	0.728
Position	Beginning	0.573	0.739
	Middle	**0.605**	**0.742**
	End	0.551	0.734

S-SIM decreases smoothly as the reference is shortened, but a 2-second excerpt still reaches 0.560, only 0.070 below the full-reference score of 0.630. Longer references provide stronger style evidence with a modest T-SIM trade-off, while T-SIM remains at least 0.728 in every condition. Across excerpt positions, S-SIM differs by at most 0.054 and T-SIM by at most 0.008. The model is therefore only weakly sensitive to excerpt position when the reference style is temporally uniform.

To distinguish temporally structured transfer from global averaging, we concatenate two known source segments into one style reference. A denotes the first segment and B denotes the second, for

example, 3 seconds of angry speech followed by 5 seconds of sad speech. The first and last 40% of the output are each scored against both A and B. The middle 20% is excluded as a transition guard band.

Table 15: Temporal style transfer.

Output segment	S-SIM ↑	S-SIM ↑
First 40%	**0.513**	0.334
Last 40%	0.290	**0.485**

In Table 15, the first output segment is more similar to A than to B (0.513 versus 0.334), whereas the last segment is more similar to B than to A (0.485 versus 0.290). This diagonal dominance matches the concatenation order and demonstrates temporally ordered style transfer. Together, the length, position, and concatenation analyses show that short references preserve most measurable style, excerpt position has a limited effect, and internal style changes are not reduced to a single global average.

D Robustness Across Gender and Language Conditions

The controlled swapping grid contains 16 style references, 8 timbre references (4 male and 4 female), and 16 target texts (8 Chinese and 8 English), yielding $16 \times 8 \times 16 = 2{,}048$ utterances. Same-gender and Cross-gender compare whether the timbre and style references have matching genders, independently of the target-text language. Matched-language and Cross-language compare whether the language of the style reference matches the target text; each aggregate includes both Chinese-to-English and English-to-Chinese cases.

Table 16: Gender and language splits.

Split	WER (%) ↓	DNSMOS ↑	T-SIM ↑	S-SIM ↑
Same-gender	0.904	3.972	0.749	0.610
Cross-gender	1.064	3.920	0.723	0.638
Matched-language	0.934	3.967	0.739	0.631
Cross-language	1.034	3.925	0.733	0.617
All (Table 1)	**0.984**	**3.946**	**0.736**	**0.624**

Changing from same-gender to cross-gender references reduces T-SIM only from 0.749 to 0.723, while S-SIM increases from 0.610 to 0.638, indicating stable cross-gender style transfer. Under a language mismatch, WER increases moderately from 0.934% to 1.034%, whereas DNSMOS, T-SIM, and S-SIM decrease by only 0.042, 0.006, and 0.014, respectively. The All row reproduces the full-grid MeanFlow result from Table 1; it is an aggregate consistency check rather than an additional condition. Overall, neither gender nor language mismatch causes a sharp degradation.

E MeanFlow Step-Efficiency Analysis

As shown in Table 1, MeanFlow reduces RTF from 0.101 to 0.049 relative to the CFM variant while improving DNSMOS from 3.908 to 3.946 and S-SIM from 0.603 to 0.624, with only small changes in WER and T-SIM. This provides a substantially better quality–efficiency trade-off than the standard 32-step CFM setting.

Figure3 shows that MeanFlow reaches stable quality within the first few denoising steps. WER falls markedly from one to four steps, while MOS, T-SIM, and S-SIM improve and then stabilize; RTF increases monotonically with additional steps. We therefore use four-step MeanFlow in all experiments.

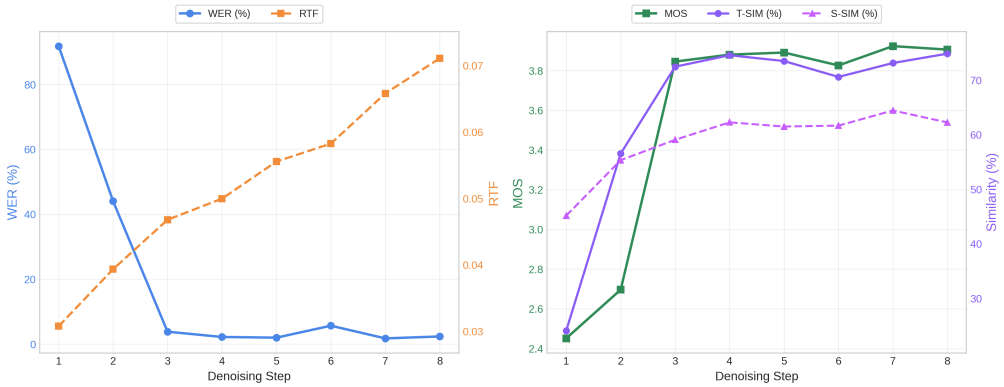

Figure 3: Step-efficiency analysis of MeanFlow.

F Post-Training Details

F.1 DPO Setup

F.1.1 DPO Training Configuration

We use Direct Preference Optimization (DPO) for post-training to further improve naturalness, emotional expression, and style consistency. The DPO fine-tuning process lasts for two epochs. The KL regularization coefficient β, which controls deviation from the reference policy, is set to 0.3. Training is organized in mini-batches, with a batch size of 8. The maximum gradient norm is clipped to 1.0. The learning rate is set to 1×10^{-7}.

F.1.2 DPO Data Construction

To improve controllability in timbre, emotion, and style, we construct preference pairs for DPO post-training. Specifically, we first select high-quality style speech samples as reference or preferred samples. Then, multiple existing TTS systems are used to generate candidate utterances for the same target content. Generating candidates from multiple systems provides diverse error types and quality levels, which helps build more informative negative samples for preference learning.

We then manually filter the generated utterances and select 500 samples with clear quality issues. These issues include content errors, missing words, incorrect pronunciation, unnatural prosody, abnormal speaking rate, weak emotional expression, style mismatch, and poor overall listening quality. For each preference pair, the original high-quality utterance is used as the chosen sample, while the low-quality synthesized utterance is used as the rejected sample.

Through this construction process, the model learns a generation preference aligned with human judgment, namely producing speech that is content-accurate, natural, timbre-stable, and consistent with the target emotional style. This post-training stage further improves fine-grained emotional control, style expression, and overall speech naturalness.

G Limitations and Ethical Considerations

G.1 Limitations and Future Work

The proposed method still has several limitations. Although the training data are carefully filtered and cover Chinese, English, and multiple emotional styles, the overall data scale and training duration have not yet reached those of ultra-large-scale speech generation systems. Therefore, as a label-free paradigm, this study has not fully validated the scaling behavior of the proposed framework under substantially larger datasets, longer training schedules, or larger model capacities. The upper bound of model performance remains to be further explored.

Future work will expand both the scale of the training data and the diversity of covered styles. We will also investigate the model's generalization ability, emotional controllability, and timbre preservation under larger-scale training conditions.

G.2 Ethics Statement

This paper studies style-controllable zero-shot speech synthesis for academic research. The goal is to improve the naturalness, style expressiveness, and timbre preservation of synthesized speech, which may benefit applications such as assistive communication, audiobook production, personalized interfaces, and human–computer interaction. The speech data used in our experiments come from publicly available or properly licensed datasets, and are filtered for content consistency, acoustic quality, and style relevance before training. We cite the original datasets and baseline systems and follow their corresponding terms of use. We do not redistribute restricted training data.

We recognize that voice is a sensitive biometric signal and that style-controllable zero-shot TTS can be misused for speaker impersonation, unauthorized voice replication, deceptive audio generation, or misleading style manipulation. Therefore, the system should only be used with voice references for which proper authorization or licensing has been obtained. Any generated speech should be clearly disclosed as synthetic, and the model should not be used for identity impersonation, fraud, misinformation, harassment, or any scenario that violates privacy, informed consent, or applicable laws and regulations.

For responsible release, the submission provides only qualitative demo samples for inspection and does not release an unrestricted voice-cloning service, model checkpoint, or processed training corpus at submission time. Any future release of checkpoints, demos, or services should include usage restrictions, consent requirements for reference voices, and appropriate access controls. We also encourage the use of provenance or watermarking mechanisms and misuse monitoring when deploying such systems in real-world applications.

The human listening evaluation is conducted with anonymized and randomized speech samples. Participants are informed of the evaluation task before participation, and no personally identifiable information is collected.