TIMESTEP-ADAPTIVE CONDITION MODULATION FOR FLOW MATCHING TEXT-TO-SPEECH

Anonymous authors
Paper under double-blind review

ABSTRACT

In the field of text-to-speech based on flow matching (FM-based TTS), a vector field (VF) predictor is learned to transport the latent state from noise to a target acoustic representation across multiple timesteps under text and reference speech conditions. Current FM-based TTS methods often concatenate the conditions with the latent state at each timestep and feed it into shared fully connected (FC) layers to obtain a mixed result, which is used to learn the VF predictor for estimating the transport path. For those methods, if the mixing process using FC layers is regarded as the guidance of conditions on the latent state, then this guidance is identical at any timestep. However, the latent state changes dynamically with timesteps. Therefore, unlike previous methods, we suppose that the guidance of conditions on the latent state also changes dynamically with timesteps. To this end, we consider text, reference speech and latent states as conditional distributions with respect to timesteps to optimize the VF predictor for explicitly analyzing whether the guidance changes with timesteps. After deduction, the dependency of the guidance on the timestep is established. Based on this fact, we propose a Timestep-Adaptive Condition Modulation (TACM) block, which uses the timestep embedding to assign two modulation factors for the text and reference speech conditions at each timestep. In this way, the TACM block enables the conditions to adaptively provide appropriate information at different timesteps for guiding the generation of the target acoustic representation. Experimental results show the superiority of our TACM block and validate our hypothesis.

1 INTRODUCTION

Text-to-speech (TTS) aims to synthesize speech from text and is widely used in smart devices and voice assistants (Ren et al., 2021; Ji et al., 2024; Zhang et al., 2024). Technically, the pipeline of TTS needs to generate an acoustic representation (e.g., Mel-spectrograms or speech tokens) according to the text and then convert that representation into speech (Ren et al., 2021; Siuzdak, 2024). In the stage of acoustic representation, there are mainly two kinds of approaches: autoregressive (AR) generation and non-autoregressive (NAR) generation. AR approaches (Wang et al., 2017) generate the acoustic representation sequentially, whereas NAR approaches (Ren et al., 2021) use parallel generation and achieve higher inference efficiency than AR approaches, contributing to the broader adoption of NAR in TTS. Recently, with advances in generative frameworks such as diffusion (Ho et al., 2020) and flow matching (FM) (Lipman et al., 2023), some researchers (Ju et al., 2024; Mehta et al., 2024) apply these frameworks to NAR-TTS. Compared with diffusion, FM requires fewer iterative steps to transport the latent state from noise to a target acoustic representation, thereby attracting increasing attention in NAR-TTS (Eskimez et al., 2024; Mehta et al., 2024).

Generally speaking, FM-based NAR-TTS methods (Zhu et al., 2025; He et al., 2025; Feng et al., 2026; Li et al., 2026b) need to learn a vector field (VF) predictor for estimating the transport path of the latent state from noise toward a target acoustic representation across multiple timesteps under conditions including text and reference speech (Lipman et al., 2023; Chen et al., 2025). For those methods, the conditions are concatenated with the latent state at each timestep, and the concatenated feature is fed into shared fully connected (FC) layers to obtain a mixed result. Then, the timestep embedding is added to the mixed result, which is used to learn the VF predictor (Chen et al., 2025; Zhu et al., 2025; He et al., 2025). Based on the above modeling manner, we regard the mixing

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.

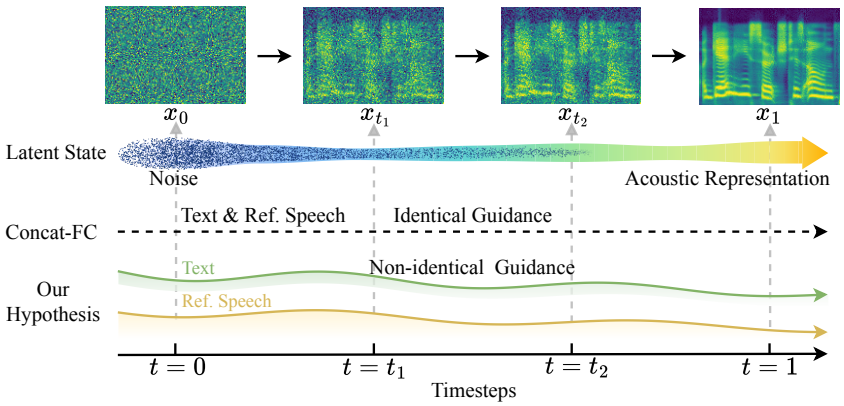


Figure 1: Guidance of text and reference speech (Ref. Speech) conditions on the latent state across timesteps. Concat-FC denotes the concatenation of conditions with the latent state followed by FC-layer processing.

process using FC layers as the guidance of conditions on the latent state. Further, through formula derivation (see Sec. 3.1), we find that such guidance is identical at any timestep.

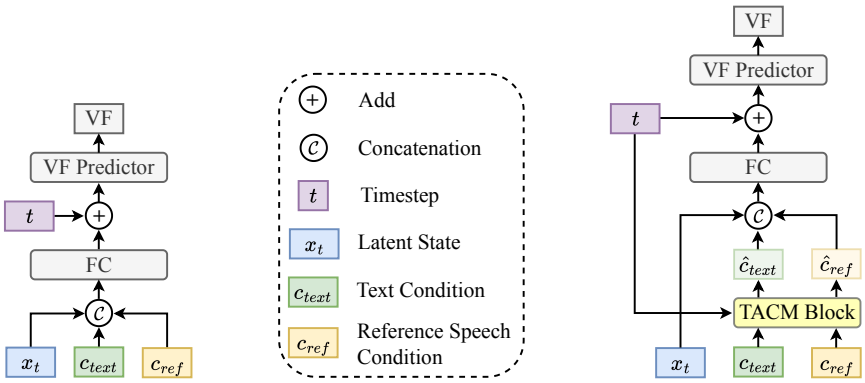
However, as shown in Fig. 1, the latent state evolves across multiple timesteps with the acoustic structure being established in the early timesteps and the acoustic details being refined in the later timesteps. And linguistic content of text and speaker characteristics of reference speech affect the acoustic structure formation and acoustic detail refinement (Chen et al., 2025; Li et al., 2026b). Therefore, unlike previous methods, it is reasonable to hypothesize that *the guidance of text and reference speech conditions on the latent state at each timestep should not be identical*. To verify this hypothesis, we consider text, reference speech, and latent states as conditional distributions with respect to timesteps and accordingly derive the optimal VF (see Sec. 3.2) to determine whether the guidance of conditions on the latent state changes with timesteps. Through derivation, we find that the guidance of conditions on the latent state does depend on the timestep. In other words, the guidance provided by text and reference speech conditions varies across timesteps as the latent state evolves from noise to a target acoustic representation. Based on this fact, we propose a Timestep-Adaptive Condition Modulation (TACM) block. The TACM block uses the timestep embedding to assign two modulation factors to the text and reference speech conditions at each timestep. In this way, the TACM block enables the conditions to adaptively provide appropriate guidance on the latent state at different timesteps for improving the quality of the generated target acoustic representation.

The main contributions of this paper are summarized as follows:

- In this paper, we consider text, reference speech, and latent states as conditional distributions with respect to timesteps to optimize the VF predictor and deduce the dependency of the guidance on the timestep. This result shows that the guidance of conditions on the latent state is timestep-dependent.
- Based on the dependency of the guidance on the timestep, we propose the TACM block. It utilizes timestep embeddings to assign two modulation factors to text and reference speech conditions, enabling these conditions to provide adaptive guidance for the latent state at different timesteps, thereby enhancing synthesized speech quality.
- Our proposed TACM block is pumped into F5-TTS and ZipVoice. The experimental results show consistent improvement on objective and subjective metrics and further validate our hypothesis that the guidance of conditions on the latent state changes with timesteps.

2 Related Work

FM-based NAR-TTS methods (Le et al., 2023; Eskimez et al., 2024; Chen et al., 2025; Zhu et al., 2025; Nguyen et al., 2026) generated target acoustic representations by transporting the latent state. This transport of the latent state was conditioned on text and reference speech, which provided linguistic content and speaker characteristics (Chen et al., 2025; Li et al., 2026b). Among those methods, Voicebox, E2-TTS, F5-TTS, and ZipVoice (Le et al., 2023; Eskimez et al., 2024; Chen et al., 2025; Zhu et al., 2025) used different VF predictors, which nevertheless followed a similar



(a) Concat-FC Architecture (b) TACM Architecture

Figure 2: (a) Concat-FC concatenates the conditions with the latent state and mixes them using FC layers at any timestep. (b) TACM uses the timestep embedding to adaptively modulate the conditions, enabling them to provide appropriate information to the latent state at different timesteps.

mixing process of conditions. In this process, the text and reference speech conditions were concatenated with the latent state, and then the concatenated feature was mixed using FC layers whose weights were reused at each timestep. The mixed result was combined with the timestep embedding and fed into the predictor trained under the FM objective to estimate the transport path.

3 METHOD

In FM-based NAR-TTS methods, the latent state x_t evolves toward a target acoustic representation x_1 along a transport path over timesteps $t \in [0, 1]$, conditioned on text c_{text} and reference speech c_{ref}. Here, $x_t, x_1, c_{\text{text}}, c_{\text{ref}} \in \mathbb{R}^{T\times D}$, where T and D are the temporal length and representation dimension. As shown in Fig. 2(a), previous methods adopt Concat-FC, which denotes concatenating x_t, c_{text}, and c_{ref} and then mixing them using shared FC layers at each timestep, as formulated in Eq. (1). In contrast, as shown in Fig. 2(b), our TACM block modulates c_{text} and c_{ref} according to t, yielding $\hat{c}_{\text{text}}, \hat{c}_{\text{ref}} \in \mathbb{R}^{T\times D}$, which are then concatenated with x_t and mixed using FC layers.

$$h_t^{\text{Concat-FC}} = \text{FC}(X) = XW + b, \qquad X = \mathcal{C}(x_t, c_{\text{text}}, c_{\text{ref}}). \tag{1}$$

where $h_t^{\text{Concat-FC}}$ is the mixed result, $\mathcal{C}(\cdot)$ denotes concatenation along dimension D, W is the FC weight matrix, and b is the bias.

3.1 GUIDANCE OF CONDITIONS IN CONCAT-FC

In this paper, we regard the terms related to c_{text} and c_{ref} as the guidance of conditions on the latent state, as follows.

$$h_t^{\text{Concat-FC}} = [x_t \quad c_{\text{text}} \quad c_{\text{ref}}] \begin{bmatrix} W_x \\ W_{\text{text}} \\ W_{\text{ref}} \end{bmatrix} + b = x_t W_x + \underbrace{c_{\text{text}} W_{\text{text}} + c_{\text{ref}} W_{\text{ref}}}_{\text{Guidance}} + b. \tag{2}$$

where $[x_t, c_{\text{text}}, c_{\text{ref}}] = X$, and W_x, W_{text}, and W_{ref} are the corresponding submatrices of W.

Based on Eq. (2), for any two different timesteps t_i and t_j, their mixing formulations are given by Eq. (3). Since the condition term in Eq. (3) is timestep-independent, the guidance of conditions on the latent state is identical at any timestep.

$$h_{t_k}^{\text{Concat-FC}} = x_{t_k} W_x + \underbrace{c_{\text{text}} W_{\text{text}} + c_{\text{ref}} W_{\text{ref}}}_{\text{Guidance}} + b, \quad k \in \{i, j\},\ t_i \neq t_j. \tag{3}$$

As t increases, x_t evolves from Gaussian noise $x_0 \in \mathbb{R}^{T\times D}$ to x_1 along the transport path, as formulated in Eq. (4). Correspondingly, x_{t_k} in Eq. (3) changes with t_k. Based on this observation,

we believe that the guidance of conditions on x_{t_k} should also change with t_k, rather than remain identical at different t.

$$x_t = (1-t)x_0 + tx_1. \tag{4}$$

3.2 Guidance of Conditions Is Timestep-Dependent

To determine the guidance of conditions on the dynamically evolving latent state, we reformulate the FM objective with respect to t. c_{text} and c_{ref} are jointly denoted by c. We sample $c \sim p(c)$, $x_1 \sim p(x_1 \mid c)$, and $x_0 \sim p(x_0)$, where x_0 is independent of (x_1, c), and t is sampled independently of (x_0, x_1, c). The target VF for the transport path in Eq. (4) is defined as follows.

$$u_t = x_1 - x_0. \tag{5}$$

For a given t, (x_t, u_t) follow from (x_0, x_1) as defined in Eqs. (4) and (5), and their joint distribution with c is denoted by $p(x_t, u_t, c \mid t)$. Since the FM objective depends on (x_t, u_t, c, t), the expectation over (x_0, x_1, c) is equivalent to the expectation under $p(x_t, u_t, c \mid t)$. By the chain rule, $p(x_t, u_t, c \mid t) = p(x_t, c \mid t)p(u_t \mid x_t, c, t)$, where $p(x_t, c \mid t)$ and $p(u_t \mid x_t, c, t)$ are the joint distribution of (x_t, c) given t and the conditional distribution of u_t given (x_t, c, t). Therefore, the FM objective is equivalently rewritten as Eq. (6). The full derivation is provided in App. B.1.

$$\begin{aligned}
\mathcal{L}_{\text{FM}} &= \mathbb{E}_{t,x_0,x_1,c}[\|v_t(x_t,t,c) - u_t\|_F^2] \\
&= \mathbb{E}_t \mathbb{E}_{(x_t,u_t,c)\sim p(x_t,u_t,c|t)}[\|v_t(x_t,t,c) - u_t\|_F^2] \\
&= \mathbb{E}_t \mathbb{E}_{(x_t,c)\sim p(x_t,c|t)} \mathbb{E}_{u_t \sim p(u_t|x_t,c,t)}[\|v_t(x_t,t,c) - u_t\|_F^2],
\end{aligned} \tag{6}$$

where $\mathbb{E}$ denotes the expectation, $v_t(x_t, t, c) \in \mathbb{R}^{T\times D}$ is the VF predictor output, and $\|\cdot\|_F$ denotes the Frobenius norm.

Given (x_t, c, t), the inner expectation in Eq. (6) is the conditional mean squared error (MSE). Hence, v_t^* is obtained as the $\arg\min_v$ of the conditional MSE, where v_t^* is the optimal VF and $\arg\min_v$ denotes the value of v that minimizes the MSE. Expanding the conditional MSE and setting its gradient with respect to v equal to zero yields v_t^*. Substituting Eqs. (4) and (5) further gives its equivalent form, as summarized in Eq. (7). See App. B.2 for details.

$$\begin{aligned}
v_t^*(x_t,t,c) &= \arg\min_v \mathbb{E}[\|v - u_t\|_F^2 \mid x_t, c, t] \\
&= \arg\min_v \Big\{ \operatorname{tr}(v^\top v) - 2\operatorname{tr}\big(v^\top \mathbb{E}[u_t \mid x_t, c, t]\big) + \mathbb{E}\big[\operatorname{tr}(u_t^\top u_t) \mid x_t, c, t\big] \Big\} \\
&= \mathbb{E}[u_t \mid x_t, c, t] \\
&= \frac{\mathbb{E}[x_1 \mid x_t, c, t] - x_t}{1-t}, \qquad 0 < t < 1,
\end{aligned} \tag{7}$$

where $(\cdot)^\top$ denotes the transpose operation, and $\operatorname{tr}(\cdot)$ denotes the trace operation.

To explicitly examine how the condition-dependent component in v_t^* varies with t, we derive an analytical form of $\mathbb{E}[x_1 \mid x_t, c, t]$. Following Gaussian settings used in FM (Lipman et al., 2023; Davtyan et al., 2025), we simplify the derivation by using the one-dimensional case and assume $x_1 \mid c \sim \mathcal{N}(\mu_c, \sigma_1^2)$ and $x_0 \sim \mathcal{N}(0, \sigma_0^2)$, where x_0 is independent of (x_1, c). Under the above assumptions and the linear transport path in Eq. (4), x_1 and x_t are jointly Gaussian conditioned on (c, t). Accordingly, $\mathbb{E}[x_1 \mid x_t, c, t]$ is given by Eq. (8), with the derivation provided in App. B.3.

$$\mathbb{E}[x_1 \mid x_t, c, t] = \mathbb{E}[x_1 \mid c] + \frac{\operatorname{Cov}(x_1, x_t \mid c, t)}{\operatorname{Var}(x_t \mid c, t)} \left(x_t - \mathbb{E}[x_t \mid c, t]\right), \tag{8}$$

where $\operatorname{Cov}(\cdot)$ denotes the conditional covariance, and $\operatorname{Var}(\cdot)$ denotes the conditional variance.

We derive the conditional means, covariance, and variance in Eq. (8) to obtain the analytical form of $\mathbb{E}[x_1 \mid x_t, c, t]$. Substituting this form into Eq. (7) gives the v_t^* decomposition in Eq. (9). See App. B.4 for the detailed derivations. Here, $\Delta(t) = (1-t)^2\sigma_0^2 + t^2\sigma_1^2$.

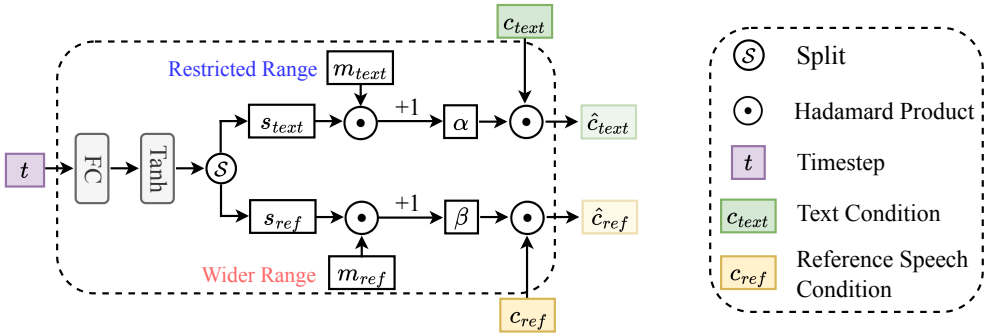


Figure 3: The TACM block uses the current timestep embedding to assign the factors $\alpha(t)$ and $\beta(t)$, which adaptively modulate the text and reference speech conditions to provide the guidance for the latent state.

$$v_t^*(x_t, t, c) = a_c(t)\mu_c + a_x(t)x_t, \qquad a_c(t) = \frac{(1-t)\sigma_0^2}{\Delta(t)}, \qquad a_x(t) = \frac{t\sigma_1^2 - (1-t)\sigma_0^2}{\Delta(t)}. \tag{9}$$

Under the Gaussian assumptions, the dependence of $p(x_1 \mid c)$ on c is represented by $\mu_c = \mathbb{E}[x_1 \mid c]$, so $a_c(t)\mu_c$ is the part of v_t^* related to the condition. As $a_c(t)$ varies with t, the guidance of conditions on the latent state is timestep-dependent.

3.3 Timestep-Adaptive Condition Modulation Block

The above derivation shows that the guidance of conditions on the latent state changes with timestep. In FM-based NAR-TTS, the guidance of text and reference speech is further shaped by task-specific conditioning and training mechanisms, e.g., reference speech is randomly masked during training (Chen et al., 2025; Zhu et al., 2025). Therefore, the analytical $a_c(t)$ derived under the Gaussian setting cannot be directly applied to the TTS task. To retain this timestep dependence while accounting for the TTS setting, we propose a TACM block that assigns two learnable modulation factors to the text and reference speech conditions based on the timestep embedding.

The timestep embedding e_t is mapped through an FC layer followed by a Tanh function, and the output is split into two parts for the conditions, as defined in Eq. (10). With the two outputs bounded to $[-1, 1]$ by the Tanh function, their signs indicate an increase or decrease in the corresponding guidance. Since text affects both prompt and target regions, while reference speech affects only the prompt region, we use a restricted range for text and a wider range for reference speech (see App. C). Then, the two outputs are scaled by their corresponding ranges and combined with the base value 1 in a residual form to obtain the modulation factors $\alpha(t)$ and $\beta(t)$, as formulated in Eq. (11).

$$s_{\text{text}}(t), s_{\text{ref}}(t) = \mathcal{S}(\text{Tanh}(\text{FC}(e_t))), \tag{10}$$

where $\mathcal{S}(\cdot)$ denotes splitting, with $s_{\text{text}}(t)$ and $s_{\text{ref}}(t)$ as the text and reference speech outputs.

$$\alpha(t) = 1 + m_{\text{text}} \odot s_{\text{text}}(t), \qquad \beta(t) = 1 + m_{\text{ref}} \odot s_{\text{ref}}(t), \tag{11}$$

where $\odot$ denotes the Hadamard product, m_{text} and m_{ref} control the modulation range widths, and $\alpha(t)$ and $\beta(t)$ are scalar factors broadcast over both dimensions.

We use $\alpha(t)$ and $\beta(t)$ to modulate the text and reference speech conditions, as expressed in Eq. (12). Afterward, the modulated conditions are concatenated with the latent state and mixed through the FC layers, as shown in Eq. (13). As $\alpha(t)$ and $\beta(t)$ vary with timestep, the modulated conditions provide adaptive guidance to the latent state, aiming to improve synthesized speech quality.

$$\hat{c}_{\text{text}} = \alpha(t) \odot c_{\text{text}}, \qquad \hat{c}_{\text{ref}} = \beta(t) \odot c_{\text{ref}}. \tag{12}$$

$$h_t^{\text{TACM}} = \text{FC}(\mathcal{C}(x_t, \hat{c}_{\text{text}}, \hat{c}_{\text{ref}})), \qquad \hat{c}_{\text{text}}, \hat{c}_{\text{ref}} = \text{TACM}(e_t, c_{\text{text}}, c_{\text{ref}}), \tag{13}$$

where $\hat{c}_{\text{text}}$ and $\hat{c}_{\text{ref}}$ are the modulated text and reference speech conditions.

Table 1: Training and inference configurations on LibriTTS. Each row applies to both the backbone and its TACM variant. NFE and CFG are the number of function evaluations and classifier-free guidance.

Backbone	Epochs	Batch Size	Learning Rate	Solver	Warmup	CFG	Seed
F5-TTS	686	64	7.5×10^{-5}	Euler (32 NFE)	20k steps	2.0	666
ZipVoice	600	Dynamic	1×10^{-2}	Euler (16 NFE)	70k steps	1.0	42

Table 2: Comparison with zero-shot TTS models on the LibriSpeech-PC test-clean benchmark, including paired comparisons of F5-TTS and ZipVoice with the TACM block.

Model	SIM-o ↑	WER ↓	UTMOS ↑	CMOS ↑	SMOS ↑
Ground Truth	0.694	1.774	4.100	0.00	3.65
F5-TTS (Chen et al., 2025)	0.625	1.871	3.877	−0.04	3.55
ZipVoice (Zhu et al., 2025)	0.608	1.503	4.155	+0.15	3.49
VoiceStar (Peng et al., 2026)	0.601	1.953	4.247	+0.09	3.30
DiFlow-TTS (Nguyen et al., 2026)	0.446	2.496	3.951	−0.10	3.12
FlexiCodec-TTS (Li et al., 2026a)	0.622	2.483	3.864	−0.05	3.09
F5-TTS + TACM	**0.673**	1.682	3.887	+0.01	**3.65**
ZipVoice + TACM	0.660	**1.413**	**4.286**	**+0.22**	3.60

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the TACM block under the zero-shot TTS setting. LibriTTS (Zen et al., 2019) is used for training and validation, with all training subsets used for training and dev-clean used for validation. For evaluation, we use the LibriSpeech-PC test-clean subset (Meister et al., 2023), which is commonly used in zero-shot TTS studies (Zhu et al., 2025; Chen et al., 2025). This test set contains 1,127 utterances from unseen speakers, with each utterance lasting 4–10 seconds.

Evaluation Metrics. Generated speech is evaluated using both objective and subjective metrics. Objective metrics include Speaker Similarity to Original (SIM-o), Word Error Rate (WER), and UTMOS for speaker similarity, linguistic content accuracy, and speech quality. Subjective metrics include Comparative Mean Opinion Score (CMOS) and Similarity Mean Opinion Score (SMOS) for speech naturalness and speaker similarity. Details are provided in App. D. For external baselines, we use the officially released checkpoints and evaluate them under the same protocol.

Implementation Details. F5-TTS (Chen et al., 2025) and ZipVoice (Zhu et al., 2025) serve as backbones, with the TACM block pumped into each backbone while all other components remain unchanged. The training and inference configurations are listed in Tab. 1. Both backbones are trained on LibriTTS using NVIDIA RTX 3090 Ti GPUs with Adam and ScaledAdam and gradient clipping. During inference, we use an Euler solver to integrate the predicted VF, and Vocos (Siuzdak, 2024) converts generated Mel-spectrograms into speech.

TACM Block Configuration. We use the same TACM block architecture and modulation hyperparameters for F5-TTS and ZipVoice. The modulation factors are set to $\alpha(t) \in [0.85, 1.15]$ for text and $\beta(t) \in [0.00, 2.00]$ for reference speech. The FC layer uses a $2 \times d_t$ weight matrix, where d_t is the timestep embedding dimension. Its weights follow $\mathcal{N}(0, 0.01^2)$ at initialization, with zero bias.

4.2 Main Results

Tab. 2 reports the zero-shot TTS evaluation results on the LibriSpeech-PC test-clean benchmark. The TACM block provides adaptive guidance of the conditions at different timesteps, making the linguistic content from the text condition and the speaker characteristics from the reference speech condition better suited to the latent state at the current timestep. As a result, TACM reduces WER by 10.1% and 6.0% on F5-TTS and ZipVoice, while improving SIM-o by 7.7% and 8.6%. Among all compared models, F5-TTS + TACM achieves the highest SIM-o and SMOS, while ZipVoice + TACM obtains the lowest WER and the highest UTMOS and CMOS. Overall, the TACM block achieves consistent improvements in both objective and subjective metrics on F5-TTS and ZipVoice, demonstrating its effectiveness as a plug-in module for FM-based NAR-TTS backbones.

Table 3: Ablation of timestep dependence and modulation factor design. Both Constant Modulation (Const. Mod.) and Variable Modulation (Var. Mod.) are timestep-independent. Const. Mod. uses the same factors at all timesteps, while Var. Mod. uses varying factors. T-Dep. indicates timestep-dependent modulation, Sep. Fs. separate text/reference factors, and 1F one shared factor.

Method	T-Dep.	Sep. Fs.	F5-TTS			ZipVoice		
			SIM-o↑	WER↓	UTMOS↑	SIM-o↑	WER↓	UTMOS↑
Concat-FC	–	–	0.625	1.871	3.877	0.608	1.503	4.155
Const. Mod.	–	✓	0.624	1.860	3.871	0.609	1.510	4.156
Var. Mod.	–	✓	0.541	1.903	3.561	0.585	1.869	3.583
TACM w/ 1F	✓	–	0.620	1.834	3.869	0.605	1.468	4.147
TACM	✓	✓	**0.673**	**1.682**	**3.887**	**0.660**	**1.413**	**4.286**

Table 4: Ablation of text and reference speech (Ref. Speech) modulation in the TACM block.

Design	Text	Ref. Speech	F5-TTS			ZipVoice		
			SIM-o↑	WER↓	UTMOS↑	SIM-o↑	WER↓	UTMOS↑
Concat-FC	–	–	0.625	1.871	3.877	0.608	1.503	4.155
TACM	✓	–	0.627	1.701	3.880	0.616	1.425	4.162
TACM	–	✓	0.661	1.854	3.882	0.658	1.486	4.201
TACM	✓	✓	**0.673**	**1.682**	**3.887**	**0.660**	**1.413**	**4.286**

Table 5: Ablation of modulation range designs for the text and reference speech conditions. Restricted and Wide correspond to the ranges $[0.85, 1.15]$ and $[0.00, 2.00]$.

Modulation Range		F5-TTS			ZipVoice		
Text	Ref. Speech	SIM-o↑	WER↓	UTMOS↑	SIM-o↑	WER↓	UTMOS↑
Wide	Wide	0.670	1.975	3.632	**0.662**	1.855	3.830
Restricted	Restricted	0.513	**1.673**	3.384	0.489	**1.405**	3.403
Restricted	Wide	**0.673**	1.682	**3.887**	0.660	1.413	**4.286**
Wide	Restricted	0.482	2.341	3.018	0.477	2.140	3.100

4.3 Ablations

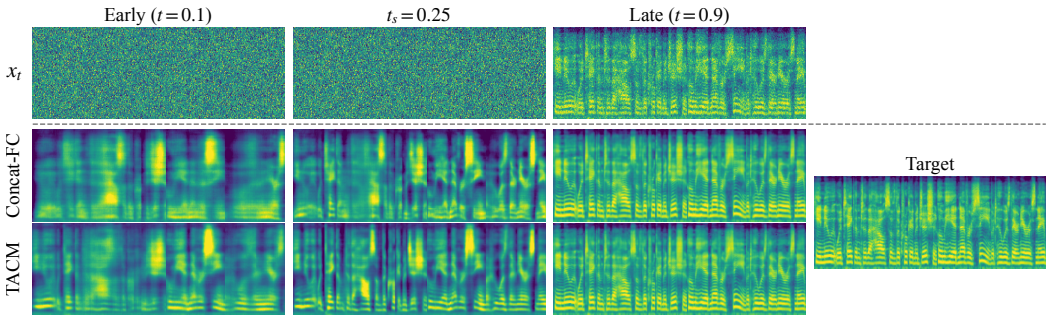
Timestep-Dependent Condition Modulation. In Tab. 3, we compare Concat-FC with timestep-independent Constant Modulation and Variable Modulation, both with learnable factors. Constant Modulation performs similarly to Concat-FC, while Variable Modulation substantially degrades performance, indicating that the modulation factors should vary in a timestep-dependent manner. Then, we compare two timestep-dependent designs using a shared factor and separate factors. The TACM block performs better on both F5-TTS and ZipVoice, supporting separate factors for the text and reference speech conditions. Those results confirm that the guidance of conditions on the latent state should be adapted according to the timestep, with separate modulation for different conditions.

Joint Modulation of Conditions. We investigate the individual effects of condition modulation. As shown in Tab. 4, text-only modulation mainly reduces WER, while reference speech-only modulation provides larger gains in SIM-o. Those results indicate that TACM provides the guidance of conditions based on their different information, allowing the text and reference speech conditions to better play their respective roles. When both conditions are modulated, F5-TTS and ZipVoice achieve the lowest WER and the highest UTMOS while maintaining strong SIM-o performance. Overall, the conditions should be jointly modulated to exploit their complementary effects.

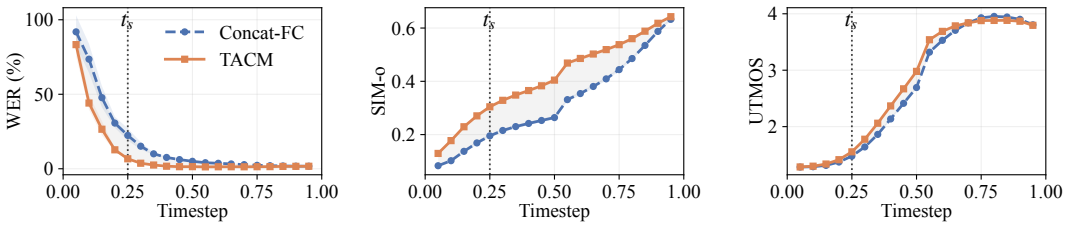
Modulation Range Design. The text and reference speech conditions have different effective regions (prompt + target vs. prompt; see App. C). Therefore, we compare different modulation ranges for the two conditions. Tab. 5 evaluates four combinations of Restricted and Wide ranges. Using the same range for both conditions fails to achieve balanced performance, while using a Wide range for text and a Restricted range for reference speech causes the largest overall degradation. In contrast, using a Restricted range for text and a Wide range for reference speech achieves the best overall balance on both F5-TTS and ZipVoice. Those results support using different modulation ranges for the text and reference speech conditions in the TACM block based on their different effective regions.

Table 6: Hyperparameter study of modulation range widths for text and reference speech conditions.

Range Design	Range	F5-TTS			ZipVoice		
		SIM-o↑	WER↓	UTMOS↑	SIM-o↑	WER↓	UTMOS↑
Text Restricted	$[0.90, 1.10]$	0.650	1.963	3.709	0.609	1.639	4.210
	$[0.85, 1.15]$	**0.673**	**1.682**	3.887	**0.660**	**1.413**	**4.286**
	$[0.80, 1.20]$	0.666	1.859	**3.898**	0.626	1.517	4.103
	$[0.75, 1.25]$	0.647	1.935	3.668	0.618	1.581	4.060
Ref. Speech Wide	$[0.60, 1.40]$	0.616	1.835	3.689	0.623	1.452	4.101
	$[0.40, 1.60]$	0.665	1.761	3.871	0.648	1.431	4.022
	$[0.20, 1.80]$	**0.674**	1.706	3.759	**0.662**	1.429	4.209
	$[0.00, 2.00]$	0.673	**1.682**	**3.887**	0.660	**1.413**	**4.286**



(a) One-step Target Prediction



(b) Objective Metrics across Timesteps

Figure 4: (a) Latent states x_t at different timesteps, together with one-step predictions of the target acoustic representation from Concat-FC and TACM. Both methods use the same x_t at each timestep. (b) WER, SIM-o, and UTMOS across timesteps, where $t_s = 0.25$ is used as an analysis boundary corresponding to the transition from rapid WER reduction to a nearly stable stage.

4.4 Analysis

Modulation Range Width. Based on Tab. 5, we vary one modulation range at a time while keeping the other fixed. As shown in Tab. 6, widening the text range from $[0.90, 1.10]$ to $[0.85, 1.15]$ improves the overall performance, while further widening it to $[0.80, 1.20]$ or $[0.75, 1.25]$ leads to degradation. This indicates that the text condition benefits from a moderately restricted modulation range. For reference speech, wider ranges generally improve the overall performance. The range $[0.20, 1.80]$ achieves the highest SIM-o, while $[0.00, 2.00]$ achieves the best WER and UTMOS on both backbones. Those results show that text and reference speech require different degrees of modulation flexibility, further supporting the range design in Tab. 5.

Effects of TACM at Different Timesteps. The timestep analyses in Figs. 4 and 5 are conducted on ZipVoice. In Fig. 4(a), at early timesteps ($t < t_s$), one-step prediction with TACM shows clearer acoustic structures than with Concat-FC. Meanwhile, Fig. 4(b) shows that WER drops quickly before t_s, indicating that TACM guides the formation of acoustic structures at early timesteps. After t_s, WER becomes nearly stable, while SIM-o and UTMOS continue to improve, showing that TACM guides acoustic detail refinement at later timesteps. Those results show that TACM provides guidance suited to the latent state at different timesteps: the text condition (linguistic content) guides acoustic structure formation at early timesteps, while the reference speech condition (speaker characteristics) guides acoustic detail refinement at later timesteps.

Table 7: Performance comparison of Concat-FC and TACM under different NFE on F5-TTS and ZipVoice.

Method	F5-TTS				ZipVoice			
	NFE	SIM-o↑	WER↓	UTMOS↑	NFE	SIM-o↑	WER↓	UTMOS↑
Concat-FC	16	0.629	2.583	3.489	8	0.616	1.551	3.499
	20	0.568	2.399	3.833	10	0.606	1.544	3.482
	32	0.625	1.871	3.877	16	0.608	1.503	4.155
TACM	16	0.678	1.867	3.614	8	0.662	1.521	3.759
	20	0.670	1.738	3.875	10	0.651	1.507	3.906
	32	0.673	1.682	3.887	16	0.660	1.413	4.286

Performance under Different NFE Settings. Since TACM modulates conditions by timestep, we evaluate it under different NFE settings with different numbers of evaluated timesteps. As shown in Tab. 7, under the same NFE, TACM consistently improves all metrics on both F5-TTS and ZipVoice. Moreover, on F5-TTS, TACM with 20 NFE achieves higher SIM-o and lower WER than Concat-FC with 32 NFE while maintaining nearly the same UTMOS. Even with 16 NFE, TACM still achieves higher SIM-o and slightly lower WER than Concat-FC with 32 NFE. Those results show that TACM remains effective with fewer timesteps while maintaining strong performance.

Verification of Timestep-Dependent Guidance. We analyze the learned modulation factors in the TACM block. As shown in Fig. 5, Concat-FC uses identical guidance of conditions at any timestep, while $\alpha(t)$ and $\beta(t)$ vary with the timestep, so the guidance of the text and reference speech conditions on the latent state also varies. Thus, the TACM block uses the timestep to modulate the text and reference speech conditions and adapt their guidance to the latent state. The variation of $\alpha(t)$ and $\beta(t)$ with t in the TACM block verifies the theoretical result in Eq. (9) that the guidance of conditions on the latent state is timestep-dependent.

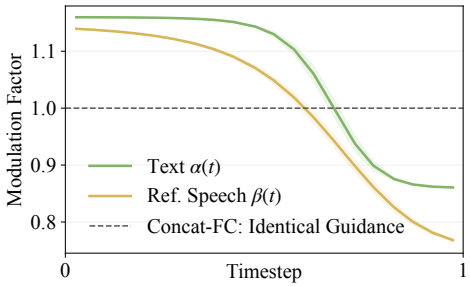


Figure 5: Guidance of text and reference speech conditions at different timesteps in TACM, compared with the identical guidance in Concat-FC.

Inference Efficiency and Computational Overhead. On an NVIDIA RTX 3090 Ti GPU, we report the mean real-time factor (RTF) over 20 test-set runs, together with the number of parameters (Params) and floating-point operations (FLOPs), in Tab. 8. The TACM block introduces less than 0.002% additional Params and 0.001% additional FLOPs, while increasing the RTF by only 0.6% and 1.5% on F5-TTS and ZipVoice. Those results show that the TACM block is a lightweight plug-in module for FM-based NAR-TTS backbones with negligible computational overhead.

Table 8: Inference efficiency and computational overhead of both models, with and without TACM.

Model	RTF	Params	FLOPs
F5-TTS	0.167	158.51M	9.03T
+TACM	+0.001	+1.54K	+0.03G
ZipVoice	0.130	122.61M	3.57T
+TACM	+0.002	+0.3K	+4.7M

5 Conclusion

Our work studies whether the guidance of text and reference speech conditions on the latent state should vary with the timestep in FM-based NAR-TTS. To this end, we consider text, reference speech, and the latent state under conditional distributions with respect to the timestep and accordingly derive the optimal VF. The derivation indicates that the guidance of conditions depends on the timestep. Based on this fact, we propose the TACM block, which uses the timestep embedding to assign two modulation factors to text and reference speech. In this way, the TACM block enables the conditions to provide adaptive guidance to the latent state at different timesteps, improving generated speech quality. Experimental results show consistent improvements with the TACM block and confirm that text and reference speech require separate timestep-dependent modulation. The learned modulation factors verify that the TACM block realizes the derived timestep-dependent guidance.

AI USE STATEMENT

Generative AI tools were not used for any tasks requiring disclosure under the ICLR 2027 AI Policy. Their use was limited to language editing aimed at improving the clarity and readability of the manuscript. Any AI-assisted edits were subsequently reviewed and revised by the authors. Responsibility for the final content of the paper, including text edited with the aid of generative AI, rests entirely with the authors.

REFERENCES

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

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.

Aram Davtyan, Leello Dadi, Volkan Cevher, and Paolo Favaro. Faster inference of flow-based generative models via improved data-noise coupling. In *International Conference on Learning Representations (ICLR)*, volume 2025, pp. 60922–60947, 2025.

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

Fangming Feng, Dongjie Fu, Zequn Xie, Yu Zhang, Yangyang Wu, Zhou Zhao, and Tao Jin. Rectifying the emotional flow: Aligning priors and dynamic guidance for high-arousal text-to-speech. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21874–21888, 2026.

Xiangheng He, Junjie Chen, Zixing Zhang, and Björn Schuller. Prosodyfm: Unsupervised phrasing and intonation control for intelligible speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 24041–24049, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

Shengpeng Ji, Ziyue Jiang, Hanting Wang, Jialong Zuo, and Zhou Zhao. Mobilespeech: A fast and high-fidelity framework for mobile zero-shot text-to-speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13588–13600, 2024.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Eric Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiangyang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. NaturalSpeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. In *International Conference on Machine Learning (ICML)*, 2024.

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 (NeurIPS)*, 36: 14005–14034, 2023.

Jiaqi Li, Yao Qian, Yuxuan Hu, Xiaofei Wang, Heng Lu, Manthan Thakker, Jinyu Li, Zhizheng Wu, et al. Flexicodec: A dynamic neural audio codec for low frame rates. In *International Conference on Learning Representations (ICLR)*, volume 2026, pp. 137256–137283, 2026a.

Yinghao Aaron Li, Xilin Jiang, Fei Tao, Cheng Niu, Kaifeng Xu, Juntong Song, and Nima Mesgarani. Dmospeech 2: Reinforcement learning for duration prediction in metric-optimized speech synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31814–31822, 2026b.

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

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

Aleksandr Meister, Matvei Novikov, Nikolay Karpov, Evelina Bakhturina, Vitaly Lavrukhin, and Boris Ginsburg. Librispeech-pc: Benchmark for evaluation of punctuation and capitalization capabilities of end-to-end asr models. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023.

Ngoc-Son Nguyen, Thanh V. T. Tran, Hieu-Nghia Huynh-Nguyen, Truong-Son Hy, and Van Nguyen. Diflow-tts: Compact and low-latency zero-shot text-to-speech with discrete flow matching. *INTERSPEECH*, 2026.

Puyuan Peng, Zhisheng Zheng, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voicestar: Robust zero-shot autoregressive tts with duration control and extrapolation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 11737–11754, 2026.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pp. 28492–28518, 2023.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations (ICLR)*, 2021.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In *INTERSPEECH*, pp. 4521–4525, 2022.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. In *INTERSPEECH*, pp. 4006–4010, 2017.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *INTERSPEECH*, pp. 1526–1530, 2019.

Leying Zhang, Yao Qian, Long Zhou, Shujie Liu, Dongmei Wang, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Lei He, et al. Covomix: Advancing zero-shot speech generation for human-like multi-talker conversations. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:100291–100317, 2024.

Han Zhu, Wei Kang, Zengwei Yao, Liyong Guo, Fangjun Kuang, Zhaoqing Li, Weiji Zhuang, Long Lin, and Daniel Povey. Zipvoice: Fast and high-quality zero-shot text-to-speech with flow matching. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025.

A STATEMENTS

Limitations. Our current study focuses on FM-based NAR-TTS. Whether timestep-adaptive condition modulation generalizes to other TTS paradigms remains unexplored. Future work could extend this study to FM-based AR-TTS and Diffusion-based TTS, which also involve generation over multiple timesteps under the guidance of conditions.

Ethics Statement. Text-to-speech systems can support beneficial applications such as human–computer interaction, accessibility, and content creation, while also posing potential risks related to impersonation, misinformation, and speech generation without a speaker's consent. These risks highlight the importance of responsible development and use of speech generation technologies. The methods and experiments in this work are presented for research purposes. For real-world deployment, we encourage the adoption of appropriate safeguards and consent mechanisms, as well as compliance with applicable data licenses and privacy requirements.

Reproducibility Statement. Information supporting the reproducibility of the proposed method and experimental results is provided throughout the main paper and appendix. The main paper describes the model architecture and configurations, training objective and settings, experimental setup, datasets, evaluation metrics, and comparison methods. Additional theoretical derivations, analyses, and evaluation details are provided in the appendix. These materials are intended to facilitate independent reproduction and verification of the results reported in this work.

B THEORETICAL DERIVATIONS

B.1 FM OBJECTIVE REFORMULATION

We sample $c \sim p(c)$, $x_1 \sim p(x_1 \mid c)$, and $x_0 \sim p(x_0)$, where x_0 is independent of (x_1, c), and t is sampled independently of (x_0, x_1, c). For a fixed t, we consider the joint distribution of (x_0, x_1, c).

$$p(x_0, x_1, c \mid t) = p(c)p(x_1 \mid c)p(x_0). \tag{14}$$

For a given t, (x_t, u_t) are obtained from (x_0, x_1) by Eqs. (4) and (5). Based on the sampling distribution in Eq. (14), the original FM objective is expanded as formulated in Eq. (15).

$$\begin{aligned}
\mathcal{L}_{\mathrm{FM}} &= \mathbb{E}_{t,x_0,x_1,c}[\|v_t(x_t,t,c) - u_t\|_F^2] \\
&= \mathbb{E}_t \mathbb{E}_{(x_0,x_1,c)\sim p(x_0,x_1,c|t)}[\|v_t(x_t,t,c) - u_t\|_F^2] \\
&= \mathbb{E}_t \int p(c)p(x_1 \mid c)p(x_0)\|v_t(x_t,t,c) - u_t\|_F^2 \, dx_0 \, dx_1 \, dc,
\end{aligned} \tag{15}$$

where $\mathbb{E}$ denotes the expectation, and $\|\cdot\|_F$ denotes the Frobenius norm.

The joint distribution of (x_t, u_t, c) conditioned on t is then written using the Dirac delta function $\delta(\cdot)$ as Eq. (16).

$$p(x_t, u_t, c \mid t) = p(c) \int p(x_1 \mid c)p(x_0) \cdot \delta(x_t - (1-t)x_0 - tx_1)\, \delta(u_t - (x_1 - x_0))\, dx_0 \, dx_1. \tag{16}$$

Using Eq. (16), we rewrite the last line of Eq. (15).

$$
\begin{aligned}
\mathcal{L}_{\mathrm{FM}} &= \mathbb{E}_t \int p(c)p(x_1 \mid c)p(x_0)\|v_t(x_t,t,c) - u_t\|_F^2 \, dx_0 \, dx_1 \, dc \\
&= \mathbb{E}_t \int p(c)p(x_1 \mid c)p(x_0) \cdot \delta(x_t - (1-t)x_0 - tx_1)\, \delta(u_t - (x_1 - x_0)) \\
&\qquad \cdot \|v_t(x_t,t,c) - u_t\|_F^2 \, dx_0 \, dx_1 \, dx_t \, du_t \, dc \\
&= \mathbb{E}_t \int p(x_t, u_t, c \mid t)\|v_t(x_t,t,c) - u_t\|_F^2 \, dx_t \, du_t \, dc \\
&= \mathbb{E}_t \mathbb{E}_{(x_t,u_t,c)\sim p(x_t,u_t,c|t)}[\|v_t(x_t,t,c) - u_t\|_F^2].
\end{aligned}
\tag{17}
$$

By integrating Eq. (16) over u_t, we obtain the joint distribution of (x_t, c) conditioned on t.

$$
\begin{aligned}
p(x_t, c \mid t) &= \int p(x_t, u_t, c \mid t) \, du_t \\
&= p(c) \int p(x_1 \mid c)p(x_0)\delta(x_t - (1-t)x_0 - tx_1) \; dx_0 \, dx_1 \\
&= p(x_t \mid c, t)p(c).
\end{aligned}
\tag{18}
$$

Since t is independent of c, $p(c \mid t) = p(c)$. We factorize the joint distribution using the chain rule.

$$
p(x_t, u_t, c \mid t) = p(x_t, c \mid t)p(u_t \mid x_t, c, t), \tag{19}
$$

where $p(x_t, c \mid t)$ is the joint distribution of (x_t, c) given t, and $p(u_t \mid x_t, c, t)$ is the conditional distribution of u_t given (x_t, c, t).

Substituting Eq. (19) into Eq. (17) gives the nested expectation form of the FM objective. Eq. (20) corresponds to Eq. (6) in the main paper.

$$
\begin{aligned}
\mathcal{L}_{\mathrm{FM}} &= \mathbb{E}_t \int p(x_t, u_t, c \mid t)\|v_t(x_t,t,c) - u_t\|_F^2 \, dx_t \, du_t \, dc \\
&= \mathbb{E}_t \int p(x_t, c \mid t)p(u_t \mid x_t, c, t)\|v_t(x_t,t,c) - u_t\|_F^2 \, dx_t \, du_t \, dc \\
&= \mathbb{E}_t \int p(x_t, c \mid t) \left[\int p(u_t \mid x_t, c, t)\|v_t(x_t,t,c) - u_t\|_F^2 \, du_t \right] dx_t \, dc \\
&= \mathbb{E}_t \mathbb{E}_{(x_t,c)\sim p(x_t,c|t)} \mathbb{E}_{u_t \sim p(u_t|x_t,c,t)}[\|v_t(x_t,t,c) - u_t\|_F^2].
\end{aligned}
\tag{20}
$$

B.2 Optimal VF Derivation

Given (x_t, c, t), the inner expectation in Eq. (6) is the conditional mean squared error (MSE) between the predictor output v and the target VF u_t. Therefore, the optimal VF $v_t^*(x_t, t, c)$ is obtained by minimizing this conditional MSE with respect to v.

$$
v_t^*(x_t, t, c) = \arg\min_v \mathbb{E}[\|v - u_t\|_F^2 \mid x_t, c, t], \tag{21}
$$

where $\arg\min_v$ denotes the value of v that minimizes the conditional MSE.

The squared error term in Eq. (21) is expanded as follows.

$$
\begin{aligned}
\|v - u_t\|_F^2 &= \mathrm{tr}((v - u_t)^\top (v - u_t)) \\
&= \mathrm{tr}(v^\top v) - \mathrm{tr}(v^\top u_t) - \mathrm{tr}(u_t^\top v) + \mathrm{tr}(u_t^\top u_t) \\
&= \mathrm{tr}(v^\top v) - 2\,\mathrm{tr}(v^\top u_t) + \mathrm{tr}(u_t^\top u_t),
\end{aligned}
\tag{22}
$$

where $\mathrm{tr}(\cdot)$ denotes the trace operation and $(\cdot)^\top$ denotes the transpose operation.

Substituting Eq. (22) into Eq. (21) gives the expanded conditional MSE.

$$
\begin{aligned}
v_t^*(x_t, t, c) &= \arg\min_v \mathbb{E}[\operatorname{tr}(v^\top v) - 2\operatorname{tr}(v^\top u_t) + \operatorname{tr}(u_t^\top u_t) \mid x_t, c, t] \\
&= \arg\min_v \Big\{ \operatorname{tr}(v^\top v) - 2\operatorname{tr}(v^\top \mathbb{E}[u_t \mid x_t, c, t]) + \mathbb{E}[\operatorname{tr}(u_t^\top u_t) \mid x_t, c, t] \Big\}.
\end{aligned} \tag{23}
$$

We compute the gradient of the conditional MSE in Eq. (23) with respect to v.

$$
\begin{aligned}
\nabla_v \mathbb{E}[\|v - u_t\|_F^2 \mid x_t, c, t] &= \nabla_v \Big\{ \operatorname{tr}(v^\top v) - 2\operatorname{tr}(v^\top \mathbb{E}[u_t \mid x_t, c, t]) + \mathbb{E}[\operatorname{tr}(u_t^\top u_t) \mid x_t, c, t] \Big\} \\
&= 2v - 2\mathbb{E}[u_t \mid x_t, c, t].
\end{aligned} \tag{24}
$$

Setting Eq. (24) equal to zero gives the minimizer of the conditional MSE.

$$
\begin{aligned}
2v_t^*(x_t, t, c) - 2\mathbb{E}[u_t \mid x_t, c, t] = 0, \\
v_t^*(x_t, t, c) = \mathbb{E}[u_t \mid x_t, c, t].
\end{aligned} \tag{25}
$$

We next express u_t using the transport path. For $0 < t < 1$, rearranging Eq. (4) gives x_0 as follows.

$$
\begin{aligned}
x_t &= (1-t)x_0 + tx_1, \\
(1-t)x_0 &= x_t - tx_1, \\
x_0 &= \frac{x_t - tx_1}{1-t}.
\end{aligned} \tag{26}
$$

Substituting Eq. (26) into Eq. (5) gives u_t in terms of x_t and x_1.

$$
\begin{aligned}
u_t &= x_1 - x_0 \\
&= x_1 - \frac{x_t - tx_1}{1-t} \\
&= \frac{(1-t)x_1 - x_t + tx_1}{1-t} \\
&= \frac{x_1 - x_t}{1-t}.
\end{aligned} \tag{27}
$$

Substituting Eq. (27) into Eq. (25) gives the equivalent form of v_t^*. Eq. (28) is the form used in Eq. (7) of the main paper.

$$
\begin{aligned}
v_t^*(x_t, t, c) &= \mathbb{E}[\frac{x_1 - x_t}{1-t} \mid x_t, c, t] \\
&= \frac{\mathbb{E}[x_1 \mid x_t, c, t] - \mathbb{E}[x_t \mid x_t, c, t]}{1-t} \\
&= \frac{\mathbb{E}[x_1 \mid x_t, c, t] - x_t}{1-t}, \qquad 0 < t < 1.
\end{aligned} \tag{28}
$$

B.3 Conditional Gaussian Expectation

For a given (c, t), the assumptions in the main paper give $x_1 \mid c \sim \mathcal{N}(\mu_c, \sigma_1^2)$ and $x_0 \sim \mathcal{N}(0, \sigma_0^2)$, where x_0 is independent of (x_1, c). Since x_0 is independent of (x_1, c) and t is independent of (x_0, x_1, c), x_0 and x_1 are conditionally independent given (c, t). Thus, $\operatorname{Cov}(x_0, x_1 \mid c, t) = 0$, giving the zero off-diagonal entries.

$$
\begin{bmatrix} x_1 \\ x_0 \end{bmatrix} \Bigg| \, c, t \sim \mathcal{N}\Big(\begin{bmatrix} \mu_c \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_0^2 \end{bmatrix} \Big). \tag{29}
$$

Based on the linear transport path in Eq. (4), (x_1, x_t) is obtained by a linear transformation of (x_1, x_0), as follows.

$$\begin{bmatrix} x_1 \\ x_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ t & 1-t \end{bmatrix} \begin{bmatrix} x_1 \\ x_0 \end{bmatrix}. \tag{30}$$

A linear transformation of jointly Gaussian variables remains jointly Gaussian. Therefore, conditioned on (c, t), (x_1, x_t) follows the joint Gaussian distribution below.

$$\begin{bmatrix} x_1 \\ x_t \end{bmatrix} \Bigg| \, c, t \sim \mathcal{N}\Big(\begin{bmatrix} \mathbb{E}[x_1 \mid c] \\ \mathbb{E}[x_t \mid c, t] \end{bmatrix}, \begin{bmatrix} \mathrm{Var}(x_1 \mid c) & \mathrm{Cov}(x_1, x_t \mid c, t) \\ \mathrm{Cov}(x_1, x_t \mid c, t) & \mathrm{Var}(x_t \mid c, t) \end{bmatrix} \Big). \tag{31}$$

To derive $\mathbb{E}[x_1 \mid x_t, c, t]$, we define the residual as follows.

$$r = x_1 - \mathbb{E}[x_1 \mid c] - \frac{\mathrm{Cov}(x_1, x_t \mid c, t)}{\mathrm{Var}(x_t \mid c, t)}(x_t - \mathbb{E}[x_t \mid c, t]), \tag{32}$$

where r is the residual after subtracting the linear term based on x_t from x_1, $\mathrm{Cov}(\cdot)$ denotes the conditional covariance, and $\mathrm{Var}(\cdot)$ denotes the conditional variance.

Taking the expectation of Eq. (32) conditioned on (c, t) gives the mean of r as follows. Since t is independent of (x_1, c), $\mathbb{E}[x_1 \mid c, t] = \mathbb{E}[x_1 \mid c]$.

$$\mathbb{E}[r \mid c, t] = \mathbb{E}[x_1 \mid c, t] - \mathbb{E}[x_1 \mid c] - \frac{\mathrm{Cov}(x_1, x_t \mid c, t)}{\mathrm{Var}(x_t \mid c, t)} \Big(\mathbb{E}[x_t \mid c, t] - \mathbb{E}[x_t \mid c, t] \Big) = 0. \tag{33}$$

The conditional covariance between r and x_t is obtained by substituting Eq. (32), as follows.

$$\begin{aligned} \mathrm{Cov}(r, x_t \mid c, t) &= \mathrm{Cov}(x_1, x_t \mid c, t) - \frac{\mathrm{Cov}(x_1, x_t \mid c, t)}{\mathrm{Var}(x_t \mid c, t)} \mathrm{Cov}(x_t, x_t \mid c, t) \\ &= \mathrm{Cov}(x_1, x_t \mid c, t) - \frac{\mathrm{Cov}(x_1, x_t \mid c, t)}{\mathrm{Var}(x_t \mid c, t)} \mathrm{Var}(x_t \mid c, t) \\ &= 0. \end{aligned} \tag{34}$$

Since r and x_t are obtained by linear combinations and constant shifts of the jointly Gaussian variables in Eq. (31), they are jointly Gaussian conditioned on (c, t). Together with Eq. (34), their zero conditional covariance implies that r and x_t are conditionally independent. Therefore, the conditional expectation of r is given as follows.

$$\mathbb{E}[r \mid x_t, c, t] = \mathbb{E}[r \mid c, t] = 0. \tag{35}$$

Rearranging Eq. (32) gives x_1 in terms of x_t and r, as formulated in Eq. (36).

$$x_1 = \mathbb{E}[x_1 \mid c] + \frac{\mathrm{Cov}(x_1, x_t \mid c, t)}{\mathrm{Var}(x_t \mid c, t)}(x_t - \mathbb{E}[x_t \mid c, t]) + r. \tag{36}$$

Taking the expectation of Eq. (36) conditioned on (x_t, c, t) and using Eq. (35) gives the conditional expectation of x_1 as formulated in Eq. (37). Eq. (37) is the form used in Eq. (8) of the main paper.

$$\begin{aligned} \mathbb{E}[x_1 \mid x_t, c, t] &= \mathbb{E}[x_1 \mid c] + \frac{\mathrm{Cov}(x_1, x_t \mid c, t)}{\mathrm{Var}(x_t \mid c, t)}(x_t - \mathbb{E}[x_t \mid c, t]) + \mathbb{E}[r \mid x_t, c, t] \\ &= \mathbb{E}[x_1 \mid c] + \frac{\mathrm{Cov}(x_1, x_t \mid c, t)}{\mathrm{Var}(x_t \mid c, t)}(x_t - \mathbb{E}[x_t \mid c, t]). \end{aligned} \tag{37}$$

B.4 Optimal VF Decomposition

Under the Gaussian assumptions in the main paper, $x_1 \mid c \sim \mathcal{N}(\mu_c, \sigma_1^2)$ and $x_0 \sim \mathcal{N}(0, \sigma_0^2)$. Since x_0 is independent of (x_1, c) and t is sampled independently of (x_0, x_1, c), the corresponding conditional means, variances, and covariance are obtained directly from these assumptions.

$$\begin{aligned} \mathbb{E}[x_1 \mid c, t] = \mathbb{E}[x_1 \mid c] = \mu_c, \qquad \mathbb{E}[x_0 \mid c, t] = \mathbb{E}[x_0] = 0, \qquad \mathrm{Var}(x_1 \mid c, t) = \sigma_1^2, \\ \mathrm{Var}(x_0 \mid c, t) = \sigma_0^2, \qquad \mathrm{Cov}(x_1, x_0 \mid c, t) = 0. \end{aligned} \tag{38}$$

Using Eqs. (4) and (38), we derive the conditional mean of x_t.

$$\begin{aligned} \mathbb{E}[x_t \mid c, t] &= \mathbb{E}[(1-t)x_0 + t x_1 \mid c, t] \\ &= (1-t)\mathbb{E}[x_0 \mid c, t] + t\mathbb{E}[x_1 \mid c, t] \\ &= t\mu_c. \end{aligned} \tag{39}$$

We derive the conditional covariance between x_1 and x_t from the same transport path.

$$\begin{aligned} \mathrm{Cov}(x_1, x_t \mid c, t) &= \mathrm{Cov}(x_1, (1-t)x_0 + t x_1 \mid c, t) \\ &= (1-t)\mathrm{Cov}(x_1, x_0 \mid c, t) + t\mathrm{Cov}(x_1, x_1 \mid c, t) \\ &= (1-t)\mathrm{Cov}(x_1, x_0 \mid c, t) + t\mathrm{Var}(x_1 \mid c, t) \\ &= t\sigma_1^2. \end{aligned} \tag{40}$$

Similarly, the conditional variance of x_t is derived as follows.

$$\begin{aligned} \mathrm{Var}(x_t \mid c, t) &= \mathrm{Var}((1-t)x_0 + t x_1 \mid c, t) \\ &= (1-t)^2\mathrm{Var}(x_0 \mid c, t) + t^2\mathrm{Var}(x_1 \mid c, t) + 2t(1-t)\mathrm{Cov}(x_0, x_1 \mid c, t) \\ &= (1-t)^2\sigma_0^2 + t^2\sigma_1^2. \end{aligned} \tag{41}$$

For brevity, we denote the conditional variance in Eq. (41) by $\Delta(t)$.

$$\Delta(t) = (1-t)^2\sigma_0^2 + t^2\sigma_1^2. \tag{42}$$

Substituting Eqs. (38) to (40) and (42) into Eq. (8) gives $\mathbb{E}[x_1 \mid x_t, c, t]$ as follows.

$$\begin{aligned} \mathbb{E}[x_1 \mid x_t, c, t] &= \mu_c + \frac{t\sigma_1^2}{\Delta(t)}(x_t - t\mu_c) \\ &= \mu_c - \frac{t^2\sigma_1^2}{\Delta(t)}\mu_c + \frac{t\sigma_1^2}{\Delta(t)}x_t \\ &= \frac{\Delta(t) - t^2\sigma_1^2}{\Delta(t)}\mu_c + \frac{t\sigma_1^2}{\Delta(t)}x_t \\ &= \frac{(1-t)^2\sigma_0^2}{\Delta(t)}\mu_c + \frac{t\sigma_1^2}{\Delta(t)}x_t. \end{aligned} \tag{43}$$

Substituting Eq. (43) into Eq. (7) gives v_t^* as follows.

$$\begin{aligned} v_t^*(x_t, t, c) &= \frac{1}{1-t}\left[\frac{(1-t)^2\sigma_0^2}{\Delta(t)}\mu_c + \frac{t\sigma_1^2}{\Delta(t)}x_t - x_t\right] \\ &= \frac{(1-t)\sigma_0^2}{\Delta(t)}\mu_c + \frac{t\sigma_1^2 - \Delta(t)}{(1-t)\Delta(t)}x_t, \qquad 0 < t < 1. \end{aligned} \tag{44}$$

We simplify the numerator of the x_t coefficient in Eq. (44).

$$\begin{aligned} t\sigma_1^2 - \Delta(t) &= t\sigma_1^2 - (1-t)^2\sigma_0^2 - t^2\sigma_1^2 \\ &= t(1-t)\sigma_1^2 - (1-t)^2\sigma_0^2 \\ &= (1-t)[t\sigma_1^2 - (1-t)\sigma_0^2]. \end{aligned} \tag{45}$$

Substituting Eq. (45) into Eq. (44) gives the decomposition of v_t^* as follows. The expressions in Eq. (46) are used in Eq. (9) of the main paper.

$$v_t^*(x_t, t, c) = a_c(t)\mu_c + a_x(t)x_t, \qquad a_c(t) = \frac{(1-t)\sigma_0^2}{\Delta(t)}, \qquad a_x(t) = \frac{t\sigma_1^2 - (1-t)\sigma_0^2}{\Delta(t)}. \tag{46}$$

C Analysis of Condition Modulation Ranges

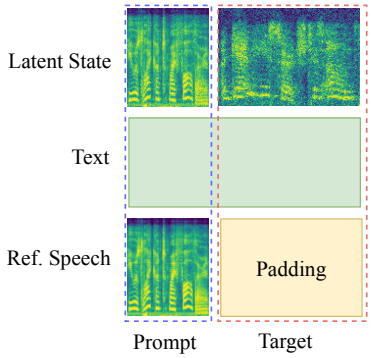


Figure 6: Effective regions of text and reference speech conditions. Text condition covers both prompt and target regions, while reference speech condition is zero-padded in the target region.

The modulation ranges control how much the text and reference speech conditions change. Since the Tanh outputs lie in $[-1, 1]$, $\alpha(t)$ and $\beta(t)$ vary around 1 within their respective ranges. A factor of 1 leaves the condition unchanged, while values below or above 1 reduce or increase its guidance on the latent state. To choose suitable ranges, we examine where the two conditions are effective, as shown in Fig. 6. The text condition covers both the prompt and target regions, whereas the reference speech condition is present only in the prompt region and zero-padded in the target region. Accordingly, the two conditions are represented as follows.

$$c_{\text{text}} = \begin{bmatrix} c_{\text{text}}^{P} \\ c_{\text{text}}^{\text{Tgt}} \end{bmatrix}, \qquad c_{\text{ref}} = \begin{bmatrix} c_{\text{ref}}^{P} \\ 0 \end{bmatrix}, \tag{47}$$

where P and Tgt represent prompt and target regions.

Applying the timestep-dependent factors gives the modulated conditions in Eq. (48). Since $\alpha(t)$ and $\beta(t)$ modulate the entire condition, text is affected in both P and Tgt, whereas reference speech is affected only in P due to zero-padding in Tgt.

$$\hat{c}_{\text{text}} = \begin{bmatrix} \alpha(t) \odot c_{\text{text}}^{P} \\ \alpha(t) \odot c_{\text{text}}^{\text{Tgt}} \end{bmatrix}, \qquad \hat{c}_{\text{ref}} = \begin{bmatrix} \beta(t) \odot c_{\text{ref}}^{P} \\ 0 \end{bmatrix}, \tag{48}$$

where $\hat{c}_{\text{text}}$ and $\hat{c}_{\text{ref}}$ are the modulated text and reference speech conditions.

The modulation changes are shown in Eq. (49).

$$\begin{aligned} \Delta c_{\text{text}} &= \hat{c}_{\text{text}} - c_{\text{text}} = \begin{bmatrix} (\alpha(t) - 1) \odot c_{\text{text}}^{P} \\ (\alpha(t) - 1) \odot c_{\text{text}}^{\text{Tgt}} \end{bmatrix}, \\ \Delta c_{\text{ref}} &= \hat{c}_{\text{ref}} - c_{\text{ref}} = \begin{bmatrix} (\beta(t) - 1) \odot c_{\text{ref}}^{P} \\ 0 \end{bmatrix}, \end{aligned} \tag{49}$$

where Δc_{text} and Δc_{ref} are the corresponding changes.

As shown in Eq. (49), the change in text spans both P and Tgt, while the change in reference speech is limited to P. Under the same modulation strength, text modulation affects a larger region. A large text modulation changes the guidance in both P and Tgt, whereas reference speech modulation remains limited to P. Based on this difference, we use a restricted range for text and a wider range for reference speech. The specific ranges are selected empirically based on Tabs. 5 and 6.

D Evaluation Metrics

For CMOS and SMOS evaluation, 20 volunteers rated the same 10 utterances for each model, with model identities hidden and samples presented in random order. All participants provided informed consent before the listening evaluation.

(1) Speaker Similarity to Original (SIM-o): SIM-o measures speaker similarity between generated and reference speech. We extracted speaker embeddings using WavLM-base-plus-sv (Chen et al., 2022; Snyder et al., 2018) and computed their cosine similarity following the Voicebox evaluation protocol (Le et al., 2023).
(2) Word Error Rate (WER): WER measures the linguistic accuracy of generated speech. We transcribed each generated utterance using Whisper-large-v3 (Radford et al., 2023) and computed the error rate between the transcription and the target text.
(3) UTMOS: UTMOS (Saeki et al., 2022) estimates the perceptual quality of generated speech directly from the waveform. A higher score indicates higher estimated naturalness.
(4) Comparative Mean Opinion Score (CMOS): CMOS compares the naturalness of generated speech with the corresponding ground truth, which is assigned a score of zero. Each pair is rated on a 7-point scale from -3 to $+3$, where positive scores favor the generated speech.
(5) Similarity Mean Opinion Score (SMOS): SMOS measures perceived speaker similarity between generated and reference speech. Each pair is rated on a 5-point scale from 1 (“not similar at all”) to 5 (“extremely similar”).