FRE-FLOW: OFFLINE MULTIPLE APPROPRIATE FACIAL REACTION GENERATION WITH FREQUENCY-AWARE FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Facial Reaction Generation (FRG) aims to synthesize contextually appropriate listener facial reactions conditioned on speaker behaviors in dyadic interactions. Since the same speaker behavior may elicit multiple appropriate reactions, FRG is inherently a one-to-many generation problem, motivating the Multiple Appropriate Facial Reaction Generation (MAFRG) task. However, existing MAFRG approaches typically employ unified spatio-temporal modeling despite the inherently multi-scale temporal structure of facial reactions. Such frequency-agnostic learning can cause cross-scale entanglement between global reaction trajectories and local variations, leading to unstable global dynamics and over-smoothed local details, while diffusion-based MAFRG approaches additionally incur costly iterative denoising for long-sequence generation. To address these challenges, we propose Fre-Flow, a frequency-aware Conditional Flow Matching (CFM) training strategy for offline MAFRG. Fre-Flow introduces two key components: (i) Frequency Flow Matching (FFM), which constructs frequency-aware spectral states and target velocities with adaptive noise scales, and (ii) Frequency Loss Optimization (FLO), which supervises velocity learning through frequency-weighted spectral matching and high-frequency energy regularization. During inference, Fre-Flow generates diverse listener facial reactions through efficient few-step ODE integration. Experiments on the MARS dataset show strong generation quality and inference efficiency, with a favorable balance among appropriateness, diversity, and structural fidelity.

1 INTRODUCTION

Facial reactions play a crucial role in dyadic communication, as they convey understanding, empathy, and interaction intent Sagliano et al. (2022). Given the same speaker behavior, different listeners may exhibit distinct yet appropriate facial reactions (AFRs), making facial reaction generation inherently a one-to-many generation problem. To address this challenge, Multiple Appropriate Facial Reaction Generation (MAFRG) aims to synthesize diverse and contextually appropriate listener facial reactions conditioned on multimodal speaker behaviors Song et al. (2023b;a; 2024; 2025). In the offline setting, complete speaker behavior sequences are available as conditions, enabling models to capture global interaction context and long-range dependencies for temporally coherent reaction generation. Modeling such one-to-many mappings requires capturing the multimodal distribution of facial reactions rather than predicting a single deterministic response, motivating probabilistic generative modeling for MAFRG.

Existing offline MAFRG methods can be broadly characterized by a progression from deterministic prediction to increasingly expressive probabilistic generative modeling. Earlier regression-based approaches directly map multimodal speaker behaviors to a single facial reaction sequence, limiting their ability to capture reaction uncertainty and multiple plausible responses. To capture reaction uncertainty, subsequent methods introduced probabilistic formulations, including variational autoencoders (VAEs), reversible architectures, and mixture-based models Ng et al. (2022); Morency et al. (2008); Kingma & Welling (2013); Xu et al. (2023); Nguyen et al. (2024). More recently, diffusion-based methods have emerged as a prominent paradigm for MAFRG, showing strong capability in generating multiple AFRs from different noise samples under the same speaker condition.

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.

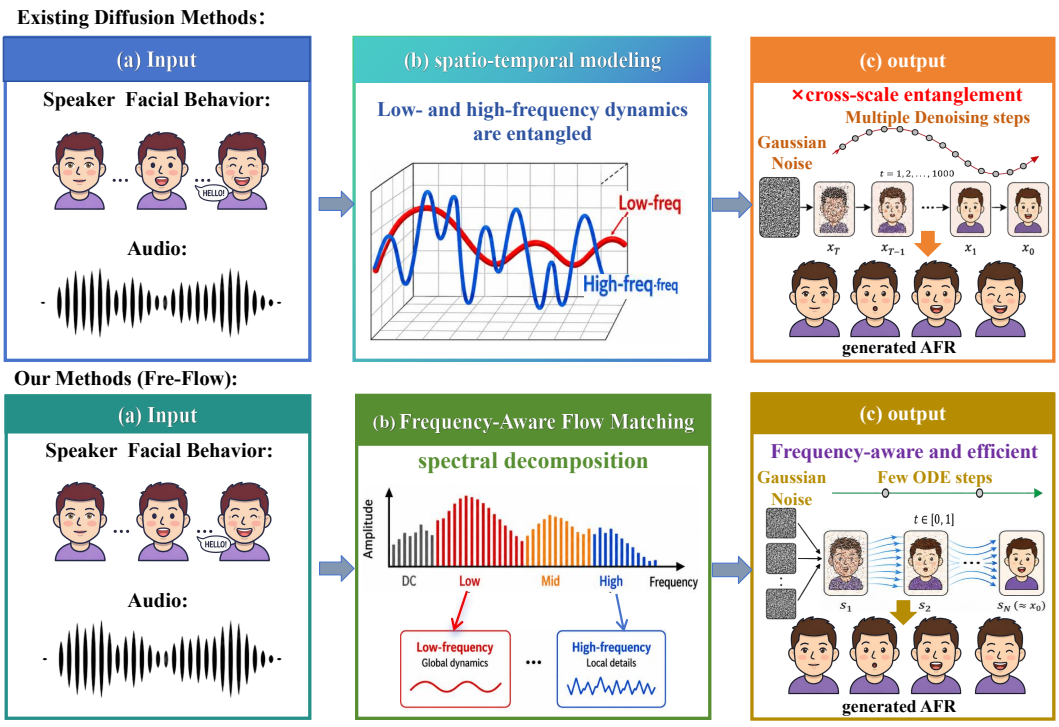


Figure 1: Comparison between diffusion-based MAFRG methods and **Fre-Flow**. Diffusion-based methods jointly model multi-frequency dynamics and rely on iterative denoising, causing cross-scale entanglement and high inference cost. Fre-Flow explicitly differentiates global dynamics and local details through frequency-aware Flow Matching, enabling efficient few-step generation.

Building on this stochastic generation process, existing diffusion-based approaches further explore personalized reaction modeling Zhu et al. (2024), latent-space generation Li et al. (2025); Luo et al. (2025), and structure-aware conditional generation Mao et al. (2025), progressively improving the appropriateness and diversity of generated facial reactions.

Despite these advances, two key problems remain in existing MAFRG approaches. First, most methods model the entire facial reaction sequence through unified spatio-temporal modeling Xu et al. (2023); Nguyen et al. (2024); Zhu et al. (2024); Li et al. (2025); Luo et al. (2025); Mao et al. (2025), without explicitly accounting for its inherent multi-scale temporal structure Rahaman et al. (2019); Durall et al. (2020); Xu et al. (2025). Facial reactions contain dynamics at different temporal frequencies: low-frequency components characterize slowly varying global reaction trajectories, whereas high-frequency components capture rapidly changing local facial details. Modeling these heterogeneous temporal scales within a shared representation can cause **cross-scale entanglement**, where global structures and local variations interfere with each other during learning, resulting in unstable global trajectories, over-smoothed local details, and temporal drift Mao et al. (2019); Rahaman et al. (2019) (**Problem 1**). Second, although diffusion-based approaches effectively model multimodal reaction distributions, they rely on iterative denoising during inference, progressively transforming Gaussian noise into facial reaction sequences over multiple sampling steps Ho et al. (2020); Song et al. (2022); Nichol & Dhariwal (2021). This iterative sampling process introduces substantial computational overhead, particularly for long-sequence generation (**Problem 2**).

In this paper, we propose frequency-aware Conditional Flow Matching (CFM) training strategy (called **Fre-Flow**) for offline MAFRG. The proposed Fre-Flow builds upon CFM to learn a continuous conditional velocity field, providing an efficient generative framework for modeling multiple AFRs through few-step ODE integration. To alleviate the cross-scale entanglement caused by frequency-agnostic modeling, Fre-Flow introduces two complementary components. (i) **Frequency Flow Matching (FFM)** constructs frequency-aware intermediate states and target velocities

in the spectral domain using frequency-adaptive noise scales, differentiating low-frequency global facial reaction trajectories from high-frequency local variations to alleviate cross-scale entanglement (**addressing Problem 1**). (ii) **Frequency Loss Optimization (FLO)** further aligns velocity learning with the frequency-aware training process through frequency-weighted spectral velocity matching and high-frequency energy regularization, improving frequency-domain consistency while preserving expressive local variations Yin et al. (2019); Tancik et al. (2020); Maiya et al. (2021). During inference, Fre-Flow integrates the learned conditional velocity field from Gaussian noise to the data distribution using a few-step Euler solver, avoiding repeated diffusion-based denoising and thereby substantially reducing sampling steps and inference overhead (**addressing Problem 2**). The main contributions and novelties of this paper are summarised as follows:

- We propose **Fre-Flow**, a frequency-aware Flow Matching training strategy for offline MAFRG, providing an efficient alternative to diffusion-based methods by incorporating multi-scale frequency information into Conditional Flow Matching.
- We introduce two frequency-aware training components: **FFM** and **FLO**. **FFM** constructs frequency-aware training states and target velocities in the spectral domain using frequency-adaptive noise scales, while **FLO** provides frequency-aware supervision to improve multi-scale consistency and mitigate over-smoothing of local facial reaction details.
- Experiments on the MARS dataset demonstrate that **Fre-Flow** achieves competitive FRCorr (0.70) and a lower FRDist (137.78) than prior MAFRG baselines, while substantially improving inference efficiency.

2 Related Work

Multiple Appropriate Facial Reaction Generation. MAFRG Song et al. (2023b;a; 2024; 2025) addresses the one-to-many mapping between speaker behaviors and listener facial reactions. Non-diffusion methods mainly include regression-based models, reversible architectures Xu et al. (2023), and mixture-based models Nguyen et al. (2024). Regression-based methods often produce deterministic and over-smoothed reactions, while reversible and mixture-based methods model uncertainty through latent distributions or probabilistic components. However, these methods still rely on holistic sequence modeling, limiting their ability to represent complex one-to-many reaction distributions. Recent diffusion-based MAFRG methods improve diversity by gradually denoising Gaussian noise under multimodal speaker conditions. Representative works include personalized diffusion for listener-specific styles Zhu et al. (2024), latent-space conditional diffusion with multimodal interaction modeling Li et al. (2025), temporal diffusion with facial kinematic and AU-dependency priors Luo et al. (2025), and structure-aware diffusion with scattering conditions Mao et al. (2025). Despite these advances, most diffusion-based methods lack explicit low- and high-frequency decomposition, leading to cross-scale entanglement, over-smoothed local details, temporal drift, and high inference cost from iterative denoising.

Frequency-aware Multi-scale Modeling. Neural networks naturally exhibit spectral bias, prioritizing low-frequency structures over high-frequency details Rahaman et al. (2019); Maiya et al. (2021); Ziyin et al. (2020); Xu et al. (2025), often leading to spectral aliasing in generative models Durall et al. (2020); Wang et al. (2020). To address this, recent works explicitly incorporate frequency-domain decomposition, utilizing Fourier features or spectral losses to improve structural consistency and detail fidelity Tancik et al. (2020); Jiang et al. (2021). While frequency-aware modeling has enhanced dynamic stability in generic video generationYin et al. (2019); Zhou et al. (2022), there remains significant room for exploration in applying frequency disentanglement to model the multi-scale dynamics of structured facial reactions.

Flow Matching for Efficient Generation. Diffusion models Ho et al. (2020); Nichol & Dhariwal (2021); Song et al. (2022) have achieved strong performance in generative modeling, but their iterative denoising process often requires many sampling steps, leading to high inference cost for long sequences Salimans & Ho (2022). This limitation has motivated recent interest in more direct generative trajectories. Flow-based paradigms, such as Flow Matching Lipman et al. (2023), Rectified Flow Liu et al. (2022), and Conditional Flow Matching (CFM) Tong et al. (2024), learn continuous velocity fields that transport simple noise distributions to data distributions. During inference, generation can be performed by integrating the learned probability flow with only a few ODE solver steps, rather than repeatedly denoising samples. Recent large-scale image and video generation models,

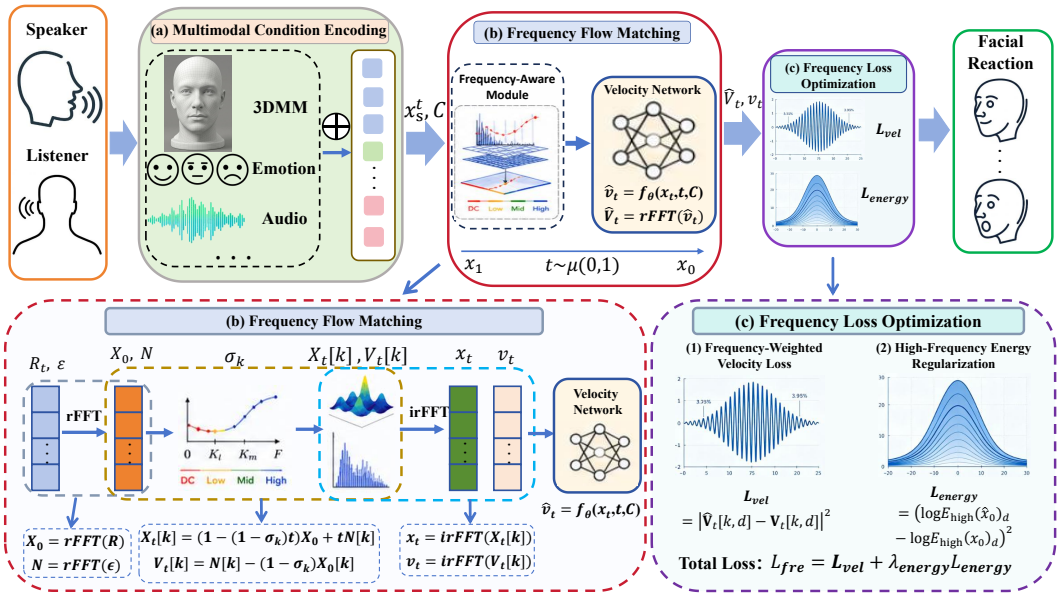


Figure 2: Overview of the proposed **Fre-Flow** training strategy. **(a) Multimodal Condition Encoding (MCE)**: Speaker audio, 3DMM, and facial attribute features are projected and token-concatenated to form the condition C, while the listener facial attribute sequence serves as the target sample $\mathbf{x}_0$. **(b) Frequency Flow Matching (FFM)**: Constructs frequency-aware training states and target velocities in the spectral domain, and learns a speaker-conditioned velocity field for generating listener facial reactions. **Frequency Loss Optimization (FLO)**: The frequency-weighted velocity loss $\mathcal{L}$vel supervises spectral velocity errors, the high-frequency energy regularization $\mathcal{L}_{\text{energy}}$ mitigate over-smoothing of local facial reaction details

such as Stable Diffusion 3 Esser et al. (2024) and Wan2.1 Wan et al. (2025) further demonstrate that Flow Matching can reduce inference cost while maintaining generation quality, making Flow Matching a promising paradigm for efficient long-sequence facial reaction generation.

3 Frequency-aware Conditional Flow Matching Training Strategy

The proposed **Fre-Flow** is a frequency-aware training strategy built upon **Conditional Flow Matching (CFM)**, introducing frequency-aware training states, target velocities, and spectral supervision to model the multi-scale temporal structure of the target AFR sequences. To alleviate cross-scale entanglement caused by frequency-agnostic modeling (**Problem 1**), our Fre-Flow differentiates low- and high-frequency learning to stabilize global trajectories while preserving local variations. Meanwhile, few-step ODE integration of the learned conditional velocity field avoids diffusion-based iterative denoising and reduces inference overhead (**Problem 2**).

Overview: Given a multimodal speaker behavior sequence $\mathbf{S} = \{\mathbf{A}_s, \mathbf{F}_s, \mathbf{E}_s\}$, where $\mathbf{A}_s$, $\mathbf{F}_s$, and $\mathbf{E}_s$ denote the speaker audio, facial geometry, and facial attribute streams, respectively, the main goal of our Fre-Flow is to generate multiple (K) AFR sequences $\hat{\mathbf{R}}_{1:K}$ conditioned on the complete speaker behavior as:

$$\hat{\mathbf{R}}_{1:K} = \text{Fre-Flow}(\mathbf{S}), \tag{1}$$

As illustrated in Figure 2, the proposed Fre-Flow first employs **Multimodal Condition Encoding (MCE)** to extract synchronized multimodal speaker behavioral cues and construct the speaker condition for AFR generation. Specifically, MCE extracts the audio representation $\mathbf{A}_s$ from the speaker audio using Wav2Vec 2.0 Baevski et al. (2020), the facial geometry representation $\mathbf{F}_s$ using FaceVerse Wang et al. (2022), and the facial attribute representation $\mathbf{E}_s$ from expression- and affect-related features. These modality-specific representations provide complementary information

about the speaker's vocal behaviors, facial geometry, and affective states that jointly constitute the multimodal speaker condition $\mathbf{C}$ as:

$$\mathbf{C} = \{\mathbf{A}_s, \mathbf{F}_s, \mathbf{E}_s\}. \tag{2}$$

The resulting condition $\mathbf{C}$ is incorporated into the conditional velocity network through cross-attention, enabling the network to predict the speaker-conditioned AFR velocity $\hat{\mathbf{v}}_t$ that is temporally and semantically aligned with the complete synchronized speaker behavior. Then, the **Frequency Flow Matching (FFM)** introduces frequency-aware training into CFM to alleviate cross-scale entanglement between low-frequency global trajectories and high-frequency local variations. This is achieved by constructing intermediate states and target velocities in the spectral domain. Unlike standard CFM, which applies a unified training process across temporal frequencies, FFM employs frequency-adaptive noise scales to differentiate the learning of low- and high-frequency AFR dynamics. Let $\mathbf{X}_0$ and $\mathbf{N}$ denote the rFFT representations of the ground-truth AFR sequence and Gaussian noise, respectively. For a sampled time $t \in [0, 1]$, FFM constructs an intermediate spectral state $\mathbf{X}_t$ and its corresponding target spectral velocity $\mathbf{V}_t$. For the k-th frequency bin, they are defined as:

$$\mathbf{X}_t[k] = \big(1 - (1 - \sigma_k)t\big)\mathbf{X}_0[k] + t\mathbf{N}[k], \qquad \mathbf{V}_t[k] = \mathbf{N}[k] - (1 - \sigma_k)\mathbf{X}_0[k]. \tag{3}$$

Here, σ_k imposes stronger constraints on low-frequency global trajectories while preserving greater flexibility for high-frequency local variations. The spectral state and target velocity are then transformed back to the time domain for network training, enabling multi-scale AFR modeling and alleviating cross-scale entanglement. Finally, **Frequency Loss Optimization (FLO)** further aligns velocity learning with the frequency-aware training process defined by FFM, aiming to improve multi-scale velocity learning and prevent high-frequency over-smoothing. It combines a frequency-weighted velocity loss $\mathcal{L}_{\text{vel}}$ with high-frequency energy regularization $\mathcal{L}_{\text{energy}}$. The former provides frequency-aware supervision for the predicted velocity, while the latter constrains high-frequency energy to alleviate over-smoothing and encourage expressive local AFR variations. The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{vel}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}}, \tag{4}$$

where λ_{energy} controls the contribution of the high-frequency energy regularization. Overall, FLO provides frequency-aware optimization that improves multi-scale consistency while encouraging fine-grained AFR variations.

3.1 Frequency Flow Matching (FFM)

Conditional Flow Matching (CFM) Lipman et al. (2023) enables strong conditional generation and efficient few-step sampling, but does not explicitly model the multi-scale temporal structure of AFR sequences. To address this issue, our **Frequency Flow Matching (FFM)** introduces frequency-aware training by constructing intermediate states and target velocities in the spectral domain, which consists of two steps: (i) frequency-band decomposition with adaptive scaling; and (ii) frequency-aware state and velocity construction. By differentiating low-frequency global trajectories from high-frequency local variations, it alleviates cross-scale entanglement while retaining CFM's generation efficiency.

During training, each ground-truth AFR sequence $\mathbf{R}$ is treated as a clean sample $\mathbf{x}_0 = \mathbf{R} \in \mathbb{R}^{T \times D_p}$, with T and D_p denoting the sequence length and AFR feature dimension, respectively. Based on this, FFM first transforms $\mathbf{x}_0$ into the spectral domain as:

$$\mathbf{X}_0 = \text{rFFT}(\mathbf{x}_0) \in \mathbb{C}^{F \times D_p}, \tag{5}$$

where $F = \lfloor T/2 \rfloor + 1$ denotes the number of non-redundant frequency bins. Lower-frequency components mainly characterize slowly varying global AFR trajectories, whereas higher-frequency components capture faster local variations. Accordingly, FFM divides the spectrum into four bands: DC ($k = 0$), Low ($1 \le k < K_l$), Mid ($K_l \le k < K_m$), and High ($k \ge K_m$), and assigns a frequency-adaptive coefficient σ_k as:

$$\sigma_k = \begin{cases} \sigma_{\text{dc}}, & k = 0 \text{ (DC)}, \\ \sigma_{\text{low}}, & 1 \le k < K_l \text{ (Low)}, \\ \sigma_{\text{mid}}, & K_l \le k < K_m \text{ (Mid)}, \\ \sigma_{\text{high}}, & k \ge K_m \text{ (High)}. \end{cases} \tag{6}$$

where K_l and K_m denote the cutoff boundaries between adjacent frequency bands. To reflect the progression from global, slowly varying dynamics to fast local variations, FFM assigns progressively larger coefficients from DC to high-frequency bands, imposing stronger constraints on low-frequency components while retaining greater flexibility for high-frequency ones:

$$\sigma_{\mathrm{dc}} \ll \sigma_{\mathrm{low}} < \sigma_{\mathrm{mid}} < \sigma_{\mathrm{high}} \ll 1. \tag{7}$$

Here, smaller coefficients impose stronger constraints on low-frequency components to stabilize global AFR trajectories, while relatively larger coefficients retain greater flexibility in high-frequency components to preserve expressive local AFR variations.

Based on the frequency-adaptive coefficients, FFM then constructs frequency-aware training states and target velocities, where Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is sampled in the time domain and transformed as $\mathbf{N} = \mathrm{rFFT}(\boldsymbol{\epsilon})$. For a sampled time $t \in [0, 1]$, the intermediate spectral state $\mathbf{X}_t[k]$ and its target velocity $\mathbf{V}_t[k]$ at the k-th frequency bin are defined as:

$$\mathbf{X}_t[k] = \big(1 - (1 - \sigma_k)t\big)\mathbf{X}_0[k] + t\mathbf{N}[k], \tag{8}$$

The frequency-adaptive coefficient σ_k controls the contribution of the ground-truth spectral component along the training path. The corresponding spectral target velocity $\mathbf{V}_t[k]$ is analytically derived as:

$$\mathbf{V}_t[k] = \frac{d\mathbf{X}_t[k]}{dt} = \mathbf{N}[k] - (1 - \sigma_k)\mathbf{X}_0[k]. \tag{9}$$

Through σ_k, **different temporal frequencies follow differentiated training trajectories rather than sharing a single uniform process, and thus alleviate interference between global and local AFR dynamics during learning.** Since the conditional velocity network operates in the time domain, the spectral state and target velocity are transformed back through inverse rFFT as:

$$\mathbf{x}_t = \mathrm{irFFT}(\mathbf{X}_t), \qquad \mathbf{v}_t = \mathrm{irFFT}(\mathbf{V}_t). \tag{10}$$

The resulting state $\mathbf{x}_t$ is then fed into the conditional velocity network f_θ, together with the sampled time t and multimodal speaker condition $\mathbf{C}$, while $\mathbf{v}_t$ serves as the corresponding training target to predict the velocity as:

$$\hat{\mathbf{v}}_t = f_\theta(\mathbf{x}_t, t, \mathbf{C}), \tag{11}$$

where $\hat{\mathbf{v}}_t$ denotes the predicted velocity. In this way, **FFM** provides frequency-aware training states and target velocities without modifying the velocity network architecture, enabling the trained model to maintain stable global AFR trajectories while preserving expressive local AFR details.

3.2 Frequency Loss Optimization (FLO)

To align the optimization of the conditional velocity network with the frequency-aware training of **FFM**, we propose **Frequency Loss Optimization (FLO)**. While existing time-domain supervision typically aggregates errors across temporal frequencies, which may weaken global trajectory learning and suppress high-frequency details, our FLO addresses this through two complementary objectives: (i) **frequency-weighted velocity supervision**; and (ii) **high-frequency energy regularization**, which jointly improve multi-scale consistency across global and local temporal scales while preserving fine-grained AFR variations.

It starts with imposing frequency-aware supervision on the predicted velocity. Since the conditional velocity network produces $\hat{\mathbf{v}}_t$ in the time domain, we first transform it into the spectral domain as:

$$\hat{\mathbf{V}}_t = \mathrm{rFFT}(\hat{\mathbf{v}}_t). \tag{12}$$

The predicted spectral velocity $\hat{\mathbf{V}}_t$ is supervised by the target $\mathbf{V}_t$ constructed in FFM, allowing direct frequency-wise alignment between the predicted and target velocities. To further account for the unequal contributions of different frequencies and AFR feature dimensions, we introduce $w_k^{(\mathrm{freq})}$ and $w_d^{(\mathrm{dim})}$ for the k-th frequency bin and d-th AFR feature dimension, respectively. The frequency-weighted velocity loss is then computed as:

$$\mathcal{L}_{\mathrm{vel}} = \frac{1}{FD_p} \sum_{k=0}^{F-1} \sum_{d=0}^{D_p-1} w_k^{(\mathrm{freq})} w_d^{(\mathrm{dim})} \left| \hat{\mathbf{V}}_t[k, d] - \mathbf{V}_t[k, d] \right|^2. \tag{13}$$

where $w_k^{(\text{freq})}$ assigns greater importance to low-frequency components to reinforce global AFR trajectories, while $w_d^{(\text{dim})}$ balances the contributions of different AFR feature dimensions. Compared with uniform time-domain supervision, this frequency-weighted objective reinforces low-frequency trajectory learning while maintaining frequency-resolved supervision for local variations, making the optimization consistent with the frequency-aware training of FFM.

Then, although $\mathcal{L}_{\text{vel}}$ supervises individual frequency components, it does not explicitly constrain the overall energy of high-frequency AFR variations, leading to over-smoothed local details and weakened fine-grained temporal dynamics. To avoid this issue, our FLO introduces high-frequency energy regularization. Based on the intermediate state $\mathbf{X}_t$, the ground-truth AFR spectrum at $t = 0$ is reconstructed from the predicted spectral velocity as:

$$\hat{\mathbf{X}}_0 = \mathbf{X}_t - t\hat{\mathbf{V}}_t. \tag{14}$$

This reconstruction enables a direct comparison between the predicted and ground-truth high-frequency statistics. Accordingly, for each AFR feature dimension d, the aggregated high-frequency energy is defined as:

$$E_{\text{high}}(\mathbf{X})_d = \sum_{k=K_m}^{F-1} |\mathbf{X}[k,d]|^2, \tag{15}$$

where K_m denotes the lower boundary of the high-frequency band defined in FFM. The corresponding energy regularization is:

$$\mathcal{L}_{\text{energy}} = \frac{1}{D_p} \sum_{d=0}^{D_p-1} w_d^{(\text{dim})} \Big(\log\Big(E_{\text{high}}(\hat{\mathbf{X}}_0)_d + \varepsilon\Big) - \log\left(E_{\text{high}}(\mathbf{X}_0)_d + \varepsilon\right) \Big)^2. \tag{16}$$

where $\varepsilon > 0$ ensures numerical stability. By matching high-frequency energy without directly constraining spectral phase, $\mathcal{L}_{\text{energy}}$ partially addresses the over-smoothing issue while retaining flexibility in local AFR variations.

Finally, FLO combines the frequency-weighted velocity loss and high-frequency energy regularization as:

$$\mathcal{L} = \mathcal{L}_{\text{vel}} + \lambda_{\text{energy}} \mathcal{L}_{\text{energy}}, \tag{17}$$

where λ_{energy} controls the contribution of the high-frequency energy regularization. The frequency-weighted velocity loss emphasizes reliable multi-scale velocity learning, while the energy regularization constrains high-frequency energy to mitigate over-smoothing. This way, the proposed frequency-weighted velocity supervision and high-frequency energy regularization jointly align the optimization objective with the frequency-aware training of FFM, promoting stable global AFR trajectories and expressive local details.

4 Experiment

4.1 Experimental Settings

Dataset. Our Fre-Flow is evaluated on the MARS dataset Song et al. (2025), a publicly available multimodal dyadic interaction dataset designed for the MAFRG task. The dataset contains 137 dyadic interaction clips involving 23 speakers and 137 listeners, which are divided into 2,856 multimodal session pairs. Each speaker behavior is associated with multiple appropriate ground-truth AFRs. Under the offline setting, the complete speaker behavior sequence is provided as the condition for generating multiple AFR sequences.

Implementation Details. We follow the preprocessing protocol of the REACT2025 benchmark Song et al. (2025). Fre-Flow is implemented in PyTorch and trained on NVIDIA A100 GPUs. The model generates 25-dimensional AFR sequences conditioned on complete multimodal speaker behavior sequences. Training is performed using the AdamW optimizer with a batch size of 4 and a cosine annealing learning rate schedule. During evaluation, the batch size is set to 1, and a fixed-step Euler ODE solver with 10 integration steps is employed for sampling.

Metrics. Following previous MAFRG benchmarks Song et al. (2023a; 2024; 2025), we evaluate the generated facial reaction attributes across four key dimensions: (i) appropriateness, measured

Table 1: Performance comparison with existing offline MAFRG methods.

Method	Appropriateness⋆		Diversity		Synchrony
	FRCorr (↑)	FRDist (↓)	FRDiv (↑)	FRVar (↑)	FRSyn (↓)
Offline Results					
Trans-VAE	0.30	181.72	0.0076	0.0083	49.00
PerfDiffZhu et al. (2024)	0.58	217.32	0.1596	0.0862	48.66
REGENNXu et al. (2023)	0.54	<u>155.49</u>	0.0012	0.0048	44.81
GAFRGWang et al. (2025)	**0.71**	173.01	0.1405	0.0769	47.77
SC-DiffMao et al. (2025)	0.59	214.68	0.0402	0.0811	49.00
Fre-Flow(ours)	<u>0.70</u>	**137.78**	0.1324	0.0738	48.94

by **FRCorr** and **FRDist**; (ii) diversity, measured by **FRDiv** and **FRVar**; (iii) synchrony, measured by **FRSyn**; and (iv) generation efficiency, measured by the number of sampling **Steps** and inference **Latency**.

4.2 Quantitative Results

Table 2: Efficiency comparison with diffusion-based MAFRG methods.

Setting	Method	Efficiency	
		Step	Latency(s)
Offline	PerfDiff	50	17.46
	Fre-Flow(ours)	10	3.44

Table 1 presents a quantitative comparison between Fre-Flow and existing offline MAFRG methods. In terms of appropriateness, Fre-Flow achieves a strong **FRCorr** score of 0.70, demonstrating reliable alignment between the generated AFRs and speaker behaviors. More importantly,Fre-Flow achieves a lower FRDist (137.78) than all prior MAFRG baselines, reducing the trajectory distance by 11.4% compared with the previous best result of REGENN Xu et al. (2023). In terms of diversity, Fre-Flow maintains competitive performance, demonstrating its ability to generate varied AFRs under the same speaker condition. Compared with diffusion-based methods such as PerFRDiff Zhu et al. (2024) and SC-Diff Mao et al. (2025), Fre-Flow demonstrates a more favorable balance between reaction appropriateness, trajectory fidelity, and generation diversity. Overall, these results demonstrate that Fre-Flow effectively balances semantic alignment, structural fidelity, temporal consistency, and one-to-many generation.

As shown in Table 2, Fre-Flow requires only 10 Euler steps and 3.44s to generate an AFR sequence, whereas PerFRDiff requires 50 denoising steps and 17.46s. Fre-Flow therefore reduces both the sampling steps and inference latency by approximately 80%, demonstrating the efficiency of few-step ODE sampling over iterative diffusion denoising.

4.3 Qualitative Results

Figure 3 visualizes the AFR sequences generated by **Fre-Flow**. Given the same multimodal speaker behavior, Fre-Flow produces multiple distinct yet contextually appropriate AFR sequences that align with the speaker's affective state and interaction context while maintaining natural temporal transitions. The generated predictions exhibit different expression intensities and temporal patterns, demonstrating the ability of Fre-Flow to capture the one-to-many nature of MAFRG. Meanwhile, the generated sequences remain temporally coherent across frames, suggesting that the proposed frequency-aware modeling stabilizes global AFR trajectories while preserving expressive local details. These results further validate Fre-Flow for diverse and appropriate AFR generation.

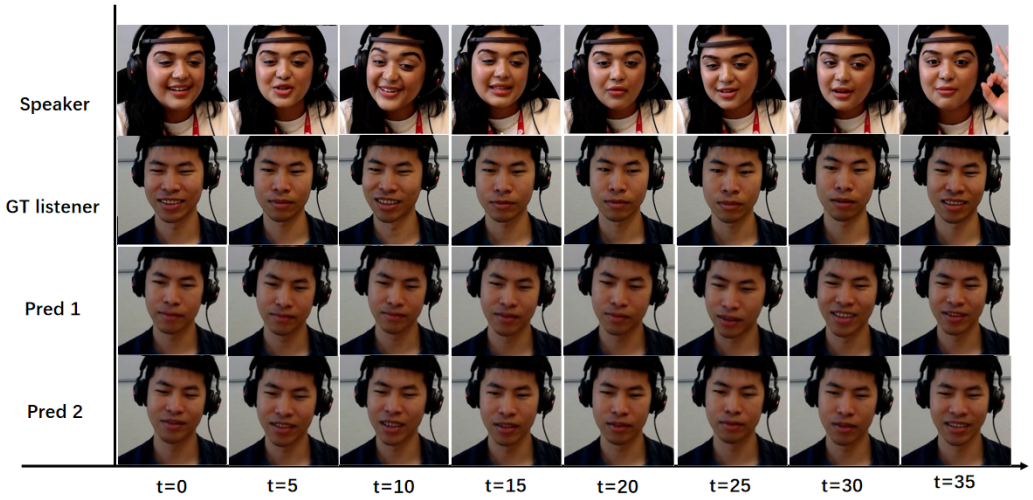


Figure 3: Qualitative results on the MARS dataset: given the speaker's multimodal behavior (top), Fre-Flow synthesizes multiple appropriate and diverse reactions (Pred 1-2) that maintain tight semantic alignment while capturing the inherent one-to-many mapping of human interactions.

Table 3: Ablation study of Fre-Flow.

Method	**Appropriateness***		**Diversity**		**Synchrony**
	FRCorr↑	FRDist↓	FRDiv↑	FRVar↑	FRSyn↓
CFM	0.60	136.69	0.1255	0.0669	48.96
CFM+FFM	0.67	141.94	0.1404	0.0701	48.84
CFM+FFM+FLO	0.70	137.78	0.1324	0.0738	48.94

4.4 Ablation Studies

We conduct ablation studies to assess the contributions of **Frequency Flow Matching (FFM)** and **Frequency Loss Optimization (FLO)** to the CFM Tong et al. (2024) baseline. As shown in Table 3, introducing FFM improves **FRCorr** from 0.60 to 0.67 and increases **FRDiv** from 0.1255 to 0.1404, demonstrating improved correlation-based appropriateness and generation diversity. However, **FRDist** increases from 136.69 to 141.94, revealing a trade-off between flexible frequency-aware transport and trajectory fidelity. Incorporating FLO further improves **FRCorr** to 0.70 and reduces **FRDist** to 137.78, largely compensating for the trajectory degradation introduced by FFM. These results show that FFM and FLO serve complementary roles: FFM enhances frequency-aware reaction modeling, while FLO improves frequency-domain alignment and largely recovers the trajectory fidelity degraded by FFM.

5 Conclusion

We propose **Fre-Flow**, a frequency-aware Conditional Flow Matching training strategy that differentiates multi-scale facial dynamics in the spectral domain. Through frequency-aware training, Fre-Flow achieves a favorable balance among appropriateness, diversity, and structural fidelity, while substantially reducing inference cost via few-step sampling. These results suggest frequency-aware Flow Matching as a promising direction for MAFRG. A limitation is the use of predefined frequency decomposition and noise scales, which may limit adaptation to sequence-specific temporal dynamics, which will be addressed by our future work.

6 AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing and presentation, feedback on the interpretation and presentation of experimental results, and software implementation and code refinement. All AI-assisted code changes were manually reviewed, tested, and validated by the authors. All reported quantitative results were obtained from actual experimental runs using the verified code. The authors take full responsibility for the final content of this work.

7 ETHICS STATEMENT

This work uses the publicly available MARS benchmark from the REACT 2025 challenge and does not involve new data collection or direct interaction with human participants. We use the released data solely for its intended research purpose and follow the benchmark protocol and applicable data-use terms. We do not perform identity recognition, collect additional personal information, or release any new participant data.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=FPnUhsQJ5B.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13919–13929, October 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jiaming Li, Sheng Wang, Xin Wang, Yitao Zhu, Honglin Xiong, Zixu Zhuang, and Qian Wang. Reactdiff: Latent diffusion for facial reaction generation. *Neural Networks*, 189:107596, 2025.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. URL https://arxiv.org/abs/2209.03003.

Cheng Luo, Siyang Song, Siyuan Yan, Zhen Yu, and Zongyuan Ge. Reactdiff: Fundamental multiple appropriate facial reaction diffusion model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5607–5616, 2025.

Shishira R Maiya, Max Ehrlich, Vatsal Agarwal, Ser-Nam Lim, Tom Goldstein, and Abhinav Shrivastava. A frequency perspective of adversarial robustness, 2021. URL https://arxiv.org/abs/2111.00861.

Qirong Mao, Qiwei Wu, Na Liu, Yakui Ding, and Lijian Gao. Scattering-conditioned diffusion models for multiple appropriate facial reaction generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13985–13991, 2025.

Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In Helmut Prendinger, James Lester, and Mitsuru Ishizuka (eds.), *Intelligent Virtual Agents*, pp. 176–190, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-85483-8.

Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20395–20405, June 2022.

Dang-Khanh Nguyen, Prabesh Paudel, Seung-Won Kim, Ji-Eun Shin, Soo-Hyung Kim, and Hyung-Jeong Yang. Multiple facial reaction generation using gaussian mixture of models and multimodal bottleneck transformer. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5. IEEE, 2024.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

Laura Sagliano, Marta Ponari, Massimiliano Conson, and Luigi Trojano. Editorial: The interpersonal effects of emotions: The influence of facial expressions on social interactions. *Frontiers in Psychology*, Volume 13 - 2022, 2022. ISSN 1664-1078. doi: 10.3389/fpsyg.2022.1074216. URL https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.1074216.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. URL https://arxiv.org/abs/2202.00512.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Siyang Song, Micol Spitale, Cheng Luo, Germán Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth Andre, et al. React2023: The first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9620–9624, 2023a.

Siyang Song, Micol Spitale, Yiming Luo, Batuhan Bal, and Hatice Gunes. Multiple appropriate facial reaction generation in dyadic interaction settings: What, why and how? *arXiv preprint arXiv:2302.06514*, 2023b.

Siyang Song, Micol Spitale, Cheng Luo, Cristina Palmero, German Barquero, Hengde Zhu, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth André, and Hatice Gunes. React 2024: the second multiple appropriate facial reaction generation challenge. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5, 2024. doi: 10.1109/FG59268.2024.10581935.

Siyang Song, Micol Spitale, Xiangyu Kong, Dr Zhu, Hengde, Cheng Luo, Cristina Palmero, German Barquero, Sergio Escalera, Michel Valstar, Mohamed Daoudi, Tobias Baur, Fabien Ringeval, Andrew Howes, Elisabeth Andrè, and Hatice Gunes. React 2025: the third multiple appropriate facial reaction generation challenge. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 13979–13984, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3762244. URL https://doi.org/10.1145/3746027.3762244.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/55053683268957697aa39fba6f231c68-Paper.pdf.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport, 2024. URL https://arxiv.org/abs/2302.00482.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: A fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20333–20342, June 2022.

Peng Wang, Pujun Xue, Xiaofeng Liu, and Tongjuan Ji. Explaining listener reactions: Personality-guided facial response generation with cross-modal attention. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13997–14003, 2025.

Tong Xu, Micol Spitale, Hao Tang, Lu Liu, Hatice Gunes, and Siyang Song. Reversible graph neural network-based reaction distribution learning for multiple appropriate facial reactions generation. *arXiv preprint arXiv:2305.15270*, 2023.

Zhi-Qin John Xu, Yaoyu Zhang, and Tao Luo. Overview frequency principle/spectral bias in deep learning. *Communications on Applied Mathematics and Computation*, 7(3):827–864, 2025.

Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/b05b57f6add810d3b7490866d74c0053-Paper.pdf.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27268–27286. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhou22g.html.

Hengde Zhu, Xiangyu Kong, Weicheng Xie, Xin Huang, Linlin Shen, Lu Liu, Hatice Gunes, and Siyang Song. Perfrdiff: Personalised weight editing for multiple appropriate facial reaction generation. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 9495–9504, 2024.

Liu Ziyin, Tilman Hartwig, and Masahito Ueda. Neural networks fail to learn periodic functions and how to fix it. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1583–1594. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1160453108d3e537255e9f7b931f4e90-Paper.pdf.

A Supplementary Experiment

A.1 Spectral and Temporal Analysis

Table 4: Spectral and temporal analysis of different training configurations. Lower values indicate better performance for all metrics.

Method	PSD-JSD ↓	High-band Error ↓	Quarter Drift ↓
CFM	0.1246	1.3432	0.1357
CFM+FFM	0.0990	**1.2359**	0.1218
Full Model	**0.0951**	1.4206	**0.1203**

To directly evaluate whether frequency-aware training reduces spectral mismatch and long-horizon temporal drift, we compare CFM, CFM+FFM, and the full Fre-Flow model using three complementary metrics. **PSD-JSD** measures the mismatch between normalized power spectral density distributions, **High-band error** measures the discrepancy of high-frequency energy, and **Quarter drift** measures the generated-to-ground-truth temporal change from the first to the last quarter of a sequence. Lower values indicate better spectral or temporal consistency.

As shown in Table 4, introducing FFM reduces PSD-JSD from 0.1246 to 0.0990, High-band error from 1.3432 to 1.2359, and Quarter drift from 0.1357 to 0.1218. These results indicate that frequency-aware training improves spectral matching and long-horizon temporal consistency compared with standard CFM. The full model further achieves the lowest PSD-JSD (0.0951) and Quarter drift (0.1203), suggesting improved overall spectral consistency and temporal stability. However, its High-band error increases to 1.4206, indicating that accurate matching of absolute high-frequency energy remains challenging despite the improvement in overall spectral structure and temporal consistency.

B MARS dataset

We conduct experiments on the MARS dataset Song et al. (2025), a large-scale corpus tailored for the MAFRG task. Its main characteristics are summarized as follows:

- **Scale & Demographics:** The dataset contains 137 natural dyadic interaction videos (30 FPS) involving 160 distinct participants (23 speakers, 137 listeners). Participants span diverse ages, genders, and educational backgrounds, supporting robust behavioral modeling.
- **Interaction Sessions:** Raw data is organized into $2,856$ multimodal session pairs. Interactions cover five predefined topics (cultural differences, movie discussions, policy changes, quizzes/games, and scenario-based interviews) to ensure semantic balance.
- **High-Dimensional Modalities:** Frame-level features are provided across distinct streams. Speaker audio is encoded as a 768-D Wav2Vec 2.0 feature. Crucially, rather than early fusion, facial states are decoupled into a 58-D 3DMM vector (52 expression, 3 pose, 3 translation) for geometry, and a 25-D affect vector (15 AUs, 2 Valence-Arousal, 8 categorical emotion probabilities) for affective semantics.
- **One-to-Many :** To capture the stochastic nature of social interaction, each speaker behavior is paired with multiple variable-length ground-truth appropriate facial reactions (AFRs) sourced from different listeners reacting to the same context.

C IMPLEMENTATION DETAILS

In this section, we provide comprehensive details regarding the data preprocessing, model architecture, and training configurations to facilitate reproducibility.

C.1 DATA PREPROCESSING

- **Input Features:** The model ingests three synchronized speaker modalities: 3DMM coefficients ($T \times 58$, comprising 52 expression, 3 rotation, and 3 translation parameters extracted via FaceVerse), audio features ($T \times 768$, pre-extracted using Wav2Vec 2.0), and emotion attributes ($T \times 25$, encompassing 15 AUs, 2 Valence/Arousal dimensions, and 8 Emotion categories).
- **Listener Targets:** The ground-truth listener representation is a 25-dimensional emotion feature vector that concatenates 15 AU intensities, 2 Valence/Arousal scores, and 8 categorical emotion logits, serving as the target signal x_0 for our **Frequency Flow Matching (FFM)** module.
- **Normalization & Image Processing:** Speaker 3DMM coefficients are standardized to a unit-variance space using a predefined per-channel mean and standard deviation precomputed on the training corpus. Listener emotion targets are zero-centered with respect to a pre-computed dataset mean prior to the **spectral decomposition and transport process**. This zero-centering ensures that the **DC component** ($k = 0$) precisely captures the global motion trends without bias. The mean is added back after inference and inverse rFFT. Raw images are resized and center-cropped to 224×224, converted to tensors, and normalized with a mean and standard deviation of $[0.5, 0.5, 0.5]$, scaling outputs to $[-1, 1]$.
- **Sequence Formatting:** Sequences are padded or cropped to a fixed length of 750 frames (representing 25 seconds at 30 FPS). During training, random starting temporal points are sampled to augment temporal diversity; during inference, testing sequences are processed in consecutive 750-frame chunks.

C.2 MODEL ARCHITECTURE

- **Overall Structure:** The architecture follows the proposed Fre-Flow training strategy and conditions the velocity network on three synchronized speaker modalities: audio, facial geometry, and facial attributes. The core generative model is a Transformer-based sequence model trained with Frequency Flow Matching (FFM) and sampled using a 10-step Euler ODE solver at inference.
- **Transformer Sequence Model:** The 9-layer Transformer Decoder employs cross-attention with a hidden dimension of 512, feed-forward dimension of 1024, 8 attention heads, 0.1 dropout, and GELU activation. Learned positional encodings are applied to both the query (noised listener sequence) and the memory (condition tokens). A 64-D sinusoidal timestep encoding is projected to 512-D via a two-layer embedding network. The cross-attention memory is formed by concatenating, along the sequence axis, the timestep embedding, speaker audio tokens (768→512), speaker 3DMM tokens (58→512), and speaker facial attribute tokens (25→512), and the frozen speaker latent embedding (512-D). For the 25-D listener facial reactions target, the signal is mapped between the reaction space and the 512-D latent space via plain linear layers ($25 \leftrightarrow 512$).
- **Frequency Flow Matching (FFM):** The model is trained with continuous flow matching over $t \in [0, 1]$. To explicitly decouple multi-scale dynamics, the transport trajectory is formulated in the spectral domain via rFFT and partitioned into four bands with a frequency-adaptive stochasticity function σ_k: $\sigma_{\mathrm{DC}} = 10^{-6}$ ($k = 0$, ensuring global structural consistency), $\sigma_{\mathrm{low}} = 10^{-4}$ (bins $1 \le k < K_l$, modeling deterministic low-frequency trends), $\sigma_{\mathrm{mid}} = 10^{-3}$ (bins $K_l \le k < K_m$, handling mid-frequency transitions), and $\sigma_{\mathrm{high}} = 10^{-2}$ (high-frequency bins $k \ge K_m$, targeting stochastic micro-expressions), where $K_l \approx 5$ and $K_m \approx 30$ out of 376 rFFT bins ($T = 750$). At inference, sampling is initialized from Gaussian noise and the learned conditional velocity field is integrated from $t = 1$ to $t = 0$ using a fixed-step Euler solver with 10 steps.

- **Classifier-Free Guidance:** Condition dropout is applied independently to audio, 3DMM, and emotion encodings at a rate of 0.2 during training. A guidance scale of 1.0 is used at inference.

C.3 Training Configuration

- **Hardware:** All models are trained on a multi-GPU configuration (NVIDIA A100 80GB GPUs) using PyTorch. To ensure stable data loading and prevent asynchronous transfer bottlenecks in this distributed environment, `pin_memory` is disabled in the dataloader. We utilize 8 workers for training and 4 for validation/testing.
- **Hyperparameters:** The model is optimized using AdamW with a learning rate of 1×10^{-4}, weight decay of 1×10^{-4}, and betas of $[0.9, 0.999]$. A CosineAnnealingLR scheduler is employed to dynamically adjust the learning rate. The training spans 300 epochs with a batch size of 4 per GPU.
- **Loss Functions:** The model is trained end-to-end using a joint objective ($\mathcal{L} = \mathcal{L}_{\text{vel}} + \lambda_{\text{energy}}\mathcal{L}_{\text{energy}}$) to explicitly optimize the frequency-aware transport dynamics.
  - **Conditional Flow Matching Loss ($\mathcal{L}_{\text{CFM}}$):** The standard time-domain objective that learns the conditional velocity field guided by the multimodal speaker behavior.
  - **Frequency Loss Optimization:** A dual-term spectral-domain objective. It comprises the Frequency-Weighted Velocity Loss ($\mathcal{L}_{\text{vel}}$) for band-wise dynamics supervision, and the High-Frequency Energy Regularization ($\mathcal{L}_{\text{energy}}$) to match global spectral statistics and preserve diverse micro-expressions without over-smoothing.

D Evaluation metrics

To accommodate the one-to-many generative paradigm of dyadic interactions, our model generates $K = 10$ stochastic facial reaction predictions $\hat{Y} = \{\hat{Y}^{(1)}, \hat{Y}^{(2)}, \ldots, \hat{Y}^{(K)}\}$ conditioned on a given speaker input. We comprehensively evaluate these predictions across appropriateness, synchrony, diversity, and computational efficiency using the following metrics:

1. Facial Reaction Correlation (FRCorr): FRCorr evaluates semantic appropriateness by measuring the maximum Concordance Correlation Coefficient (CCC) between the K predictions and the ground-truth set Y. For a predicted sequence $\hat{Y}^{(k)}$, the CCC is defined as:

$$\text{CCC}(\hat{Y}^{(k)}, Y) = \frac{2\rho\sigma_{\hat{Y}^{(k)}}\sigma_Y}{\sigma^2_{\hat{Y}^{(k)}} + \sigma^2_{\hat{Y}} + (\mu_{\hat{Y}^{(k)}} - \mu_Y)^2}$$

where ρ is the Pearson correlation coefficient, while μ and σ^2 denote the mean and variance, respectively.

2. Facial Reaction Distance (FRDist): FRDist quantifies structural and kinematic dissimilarity using Dynamic Time Warping (DTW). Given a prediction $\hat{Y}^{(k)}$ of length T and a ground-truth sequence Y of length L, DTW identifies the optimal alignment path $W = w_1, w_2, \ldots, w_P$ to minimize the cumulative distance:

$$\text{DTW}(\hat{Y}^{(k)}, Y) = \min_W \left[\sum_{p=1}^{P} d(w_p)\right]$$

where $d(w_p)$ is the weighted Euclidean distance computing the alignment cost at step w_p. The weights are empirically set to $1/15$, 1, and $1/8$ for the 15 Action Units (AUs), 2 Valence-Arousal (VA) dimensions, and 8 categorical expression dimensions, respectively.

3. Facial Reaction Synchrony (FRSyn / TLCC): To evaluate cross-modal speaker-listener alignment, Time-Lagged Cross-Correlation (TLCC) identifies the optimal temporal offset τ that maximizes the correlation between the speaker's signal X_s and the predicted reaction $\hat{Y}$:

$$\text{TLCC} = \arg\max_\tau \frac{\sum_t (X_s(t) - \mu_{X_s})(\hat{Y}(t+\tau) - \mu_{\hat{Y}})}{\sigma_{X_s}\sigma_{\hat{Y}}}$$

A smaller absolute optimal lag $|\tau^*|$ indicates tighter, more empathetic temporal alignment.

4. Facial Reaction Diversity (FRDiv / s-MSE): To evaluate inter-sample diversity and penalize mode collapse, we calculate the static Mean Squared Error (s-MSE), which computes the mean pairwise squared Euclidean distance across the K generated candidates:

$$\text{s-MSE} = \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j \neq i}^{K} \left\| \hat{Y}^{(i)} - \hat{Y}^{(j)} \right\|_2^2$$

5. Facial Reaction Variance (FRVar): FRVar assesses within-sequence temporal expressiveness by calculating the average per-dimension variance across the temporal axis T:

$$\text{FRVar} = \frac{1}{K \cdot D} \sum_{k=1}^{K} \sum_{d=1}^{D} \left(\frac{1}{T} \sum_{t=1}^{T} (\hat{Y}_{t,d}^{(k)} - \bar{\hat{Y}}_{d}^{(k)})^2 \right)$$

where D represents the dimensionality of the emotional attributes, and $\bar{\hat{Y}}_{d}^{(k)}$ is the temporal mean of the d-th feature.

6. Generation Efficiency (Steps & Latency): To evaluate the computational overhead, we report the number of ODE sampling Steps required to synthesize a reaction sequence and the corresponding inference Latency, defined as the time (in seconds) required to generate a single facial reaction sequence. Lower values indicate higher practical generation efficiency.

E LIMITATIONS

While our proposed Fre-Flow framework demonstrates significant advancements in generating contextually appropriate and diverse listener facial reactions, several limitations closely related to the domain of human affective interaction remain to be addressed:

- **Cultural and Contextual Generalizability**: Facial reactions in dyadic interactions are inherently nuanced and highly sensitive to cultural norms, interpersonal relationships, and social contexts. Although Fre-Flow effectively captures conversational rhythms within the MARS dataset (covering five predefined scenarios), its zero-shot generalization to cross-cultural interactions or high-stakes environments (e.g., psychological counseling or legal negotiations) remains unexplored. Future work should investigate culture- or personality-aware conditionings to explicitly modulate reaction styles.
- **Long-term Interaction Consistency**: Human interactions are not static; listeners dynamically adapt their reactive behaviors based on rapport and the cumulative history of the conversation. Currently, Fre-Flow processes interactions in fixed 750-frame (25-second) windows. While this successfully models short-term reactive semantics, incorporating long-term stateful memory mechanisms to capture evolving relationship dynamics over multi-minute conversations is a critical challenge for future affective computing research.
- **Robustness to "In-the-Wild" Modality Noise**: Moving towards real-world applications, our framework relies heavily on high-fidelity upstream representations (e.g., Wav2Vec 2.0 audio and FaceVerse 3DMM coefficients). In unconstrained environments where audio is corrupted by ambient noise or visual tracking suffers from occlusions, the spectral decomposition (rFFT) within the Frequency Flow Matching (FFM) module may be sensitive to high-frequency artifacts, potentially degrading the geometric authenticity of the generated reactions.
- **Static vs. Dynamic Frequency Decomposition**: While FFM explicitly decouples multi-scale dynamics, the current spectral boundaries (K_l, K_m) and noise scales (σ_k) are statically predefined. Human facial dynamics, however, are highly non-stationary; the optimal threshold separating global trends from micro-expressions may shift based on conversational intensity or individual styles. Developing a dynamic frequency-aware mechanism that adaptively infers spectral boundaries conditioned on real-time context is a vital direction for enhancing expressive fidelity.
- **Real-time Streaming Bottleneck**: For integration into embodied conversational agents, ultra-low-latency response is paramount. Although Fre-Flow utilizes Conditional Flow

Matching to drastically reduce sampling steps (to 10 steps) and inference latency compared to traditional diffusion models, it currently operates on sequence chunks. Exploring flow distillation or sliding-window ODE solvers to achieve true frame-by-frame, real-time streaming generation remains an ongoing engineering challenge.