MAM-REACTOR: ANCHOR-PRESERVING SET-LEVEL ALIGNMENT FOR MULTIPLE APPROPRIATE FACIAL REACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Offline Multiple Appropriate Facial Reaction Generation (MAFRG) requires more than producing multiple plausible appropriate facial reactions (AFRs): the generated reactions should be both individually appropriate and collectively diverse. Existing offline MAFRG approaches mainly address one-to-many reaction generation through distribution-level modeling or instance-wise supervision. The former models each reaction sequence holistically, capturing global distributional trends but providing weaker fine-grained frame-level alignment, whereas the latter directly supervises individual reactions but does not coordinate multiple predictions as a set. Consequently, neither formulation explicitly organizes the finite output set to complementarily cover multiple valid reactions for the same interaction. We therefore formulate offline MAFRG as a set-level alignment problem. To this end, we propose **Mam-Reactor**, which combines trajectory-level anchoring with set-level alignment. A deterministic paired-reaction anchor provides a reliable fine-grained appropriateness reference, while expression-identified residual queries and diversity regulation expand it into an expressive candidate set in one forward pass. Balanced prediction–target set alignment then jointly optimizes the correspondence between the complete prediction set and multiple valid same-session reactions, encouraging different candidates to cover complementary valid reaction modes rather than redundantly favoring the same target. A lightweight isolated VA refinement further improves continuous affective trajectories without altering the learned set structure. Experiments on the offline MARS benchmark show that our Mam-Reactor achieved an FRC of 1.0888 and an FRD of 134.81 at an FRDiv of 0.1521. It further generates all ten reactions in a single forward pass, providing approximately $857\times$ faster inference than the 50-step diffusion baseline.

1 INTRODUCTION

Multiple Appropriate Facial Reaction Generation (MAFRG) aims to generate several appropriate facial reactions (AFRs) for a listener conditioned on the observed speaker behavior, and provides an important foundation for natural digital humans, virtual characters, and socially interactive agents Song et al. (2023b; 2024; 2025; 2026). A fundamental challenge is the inherent one-to-many of social interaction: the same speaker behavior may elicit multiple distinct yet valid AFRs. Therefore, offline MAFRG requires not only that each facial reaction remains appropriate to the current interaction, but also that the finite prediction set should complementarily cover multiple valid response alternatives rather than repeatedly explaining the same reaction mode.

Existing offline MAFRG approaches mainly organize one-to-many supervision through two paradigms: distribution-level modeling and instance-wise supervision. Distribution-level methods typically represent multiple valid AFRs through a learned conditional reaction distribution, using latent-variable models, mixture distributions, discrete latent representations, or stochastic generative processes to capture the global variability of listener responses Xu et al. (2026); Nguyen et al. (2024c;a); Liu et al. (2024). Under this formulation, AFRs are jointly modeled in a shared response space, and the facial reaction distribution rather than a specific reaction trajectory becomes the primary learning object. In contrast, instance-wise methods usually establish direct supervi-

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.

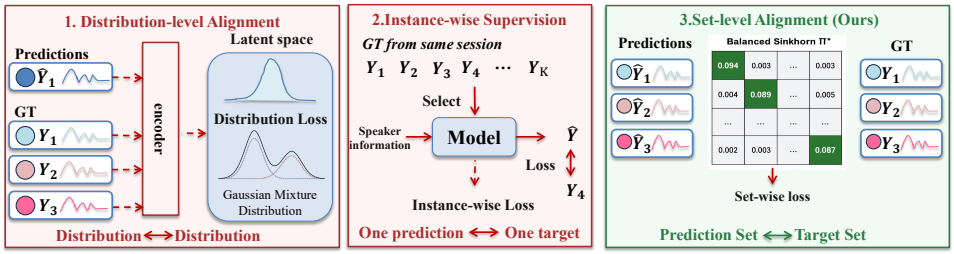


Figure 1: Comparison of supervision paradigms for offline MAFRG. Distribution-level methods optimize global reaction distributions, whereas instance-wise methods supervise individual prediction–target relations. Mam-Reactor shifts the supervision unit to the complete prediction and target AFR sets, explicitly coordinating a finite set of predictions to cover complementary valid reactions.

sion between an individual prediction and a target AFR or a paired speaker-listener example Hoque et al. (2023); Zhu et al. (2024). Such objectives provide explicit reaction-level learning signals by constraining each prediction toward an observed appropriate response. Despite their different implementations, existing approaches therefore organize MAFRG supervision mainly either at the level of a *reaction distribution* or at the level of an *individual prediction-target relation*.

However, neither of these supervision paradigm fully matches the finite-set nature, i.e., the model is required to generate a fixed number of AFR candidates that jointly cover multiple valid responses to the same speaker behavior. Specifically, for distribution-level modeling, global response variability is learned in a shared latent or stochastic space, but the strictly paired AFR is not explicitly preserved as a common fine-grained reference for every generated candidate. Consequently, the generated AFR candidates are not directly anchored to the reliable interaction-specific trajectory provided by paired supervision, namely the strictly paired AFR that is temporally aligned with the current speaker sequence. For instance-wise supervision, each prediction can be explicitly constrained toward an AFR, but different prediction–target relations are still largely optimized independently. It therefore does not explicitly determine how multiple predictions should be jointly organized to cover different valid AFRs. As a result, several candidates may redundantly account for the same valid reaction while other valid alternatives remain insufficiently represented. What is missing is a supervision mechanism that simultaneously preserves the temporally aligned paired-AFR reference and coordinates the complete prediction set toward complementary valid targets. This motivates our central question: how can offline MAFRG maintain fine-grained appropriateness while maximizing the coverage of valid AFRs with a finite prediction set?

To address these offline MAFRG supervision gaps, we propose Mam-Reactor, an anchor-preserving AFR set-level supervision framework for offline MAFRG. Its key idea lies in preserving a reliable facial reactions reference while explicitly organizing multiple predictions to cover different valid AFRs. First, we learn a deterministic anchor from the strictly paired facial reactions and freeze it as a fine-grained reference, ensuring that all generated candidates remain grounded in reliable interaction dynamics. We then expand this anchor into an expressive candidate set using expression-identified residual queries and diversity regulation. Crucially, instead of optimizing these candidates independently, we jointly supervise the predicted and target AFR sets through balanced set-level supervision, encouraging different predictions to account for different valid reactions rather than redundantly concentrating on the same one. In this way, Mam-Reactor expands the coverage of valid reactions while keeping every candidate grounded in a reliable interaction trajectory. Finally, a lightweight isolated VA refinement further improves continuous affective trajectories after the reaction-set structure has been established. The main contributions of this paper are twofold:

- We formulate offline MAFRG as a finite AFR-set supervision problem and propose an anchor-preserving one-step framework that combines a reliable paired-AFR reference with balanced set-level supervision to coordinate multiple predictions toward complementary valid AFRs, without requiring additional cognitive annotations.
- Experiments on MARS demonstrate a strong appropriateness–diversity trade-off, with Mam-Reactor achieving an FRC of 1.0888 and an FRD of 134.81 at an FRDiv of 0.1521.

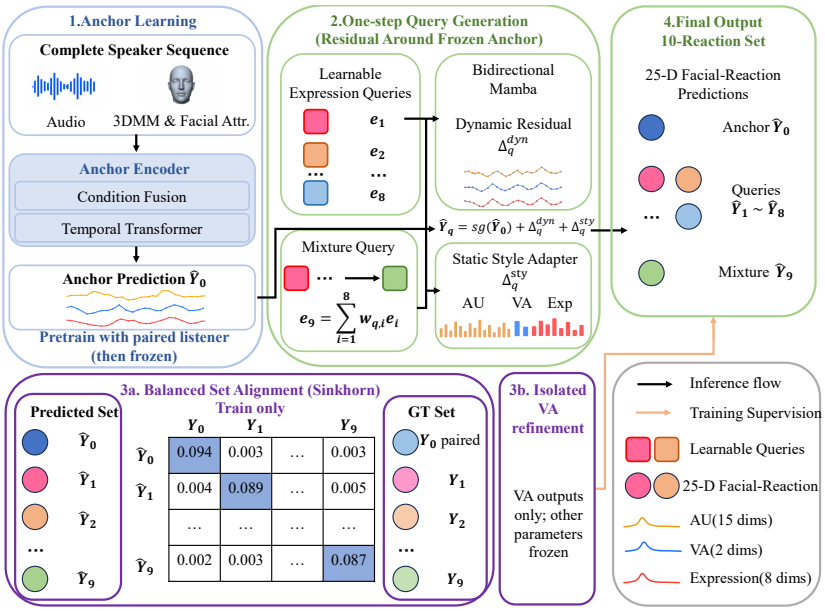


Figure 2: Overview of the proposed Mam-Reactor. A deterministic reaction anchor is pretrained from paired facial reactions and frozen. Expression-identified queries construct an anchor-centered facial reaction set in one forward pass, while diversity regulation expands the candidate space. Balanced Sinkhorn transport then jointly aligns the complete facial reaction set with AFRs during training, followed by isolated VA refinement.

2 Related Work

Facial Reaction Generation (FRG). Existing FRG methods mainly formulate reaction generation as a deterministic one-to-one mapping, where each speaker behavior is paired with a single listener reaction. Early GAN-based methods generate listener facial expressions from speaker facial or affective cues Huang & Khan (2017); Nojavanasghari et al. (2018), and later extensions further improve photorealism and temporal dynamics in dyadic or human-agent interactions Huang & Khan (2018). Overall, these methods focus on predicting a single appropriate reaction for each observed speaker behavior.

Multiple appropriate facial reaction generation (MAFRG). Existing MAFRG methods mainly improve supervision and objective design to address the inherent one-to-many mapping problem, where one speaker behavior may correspond to multiple appropriate listener reactions. Early non-diffusion methods introduce latent-variable, distribution-learning, contrastive, or reconstruction-based objectives to avoid deterministic frame-level regression and improve reaction diversity Hoque et al. (2023); Nguyen et al. (2024a); Xu et al. (2026). More recent methods further use discrete latent variables or vector quantization to organize diverse reaction modes into learnable codebooks Liu et al. (2024); Nguyen et al. (2024b). Diffusion-based methods model the one-to-many space through stochastic denoising objectives Li et al. (2025); Nguyen et al. (2024c), and recent variants introduce additional priors or auxiliary constraints Luo et al. (2025); Mao et al. (2025); Zhu et al. (2024). Despite these advances, candidate reactions are still largely modeled implicitly through latent codes or stochastic sampling. In contrast, Mam-Reactor explicitly organizes multiple reactions through a reliable anchor, expression-identified reaction queries, and set-level alignment.

3 Task Definition

Let $\mathcal{X}_i = \{X_i^a, X_i^f, X_i^m\}$ denote the multimodal behavior of the i-th speaker segment, where X_i^a, X_i^f, and X_i^m represent the audio, facial-attribute, and 3DMM sequences, respectively. Unlike traditional facial reaction generation task, in which one speaker behavior is associated with a single target reaction, **offline MAFRG** assumes that the same speaker behavior may admit a set of dif-

ferent AFRs. We denote by $\mathcal{A}(\mathcal{X}_i) \subseteq \mathbb{R}^{T_i \times D}$ the space of facial reaction sequences that constitute appropriate responses to speaker behavior $\mathcal{X}_i$. The goal of offline MAFRG is to learn a mapping

$$F_\theta : \mathcal{X}_i \longmapsto \hat{\mathcal{Y}}_i = \left\{ \hat{Y}_{i,q} \right\}_{q=0}^{Q-1}, \tag{1}$$

where $\hat{\mathcal{Y}}_i$ denotes a set of Q predicted AFRs, with each $\hat{Y}_{i,q} \in \mathbb{R}^{T_i \times D}$. Ideally, the generated reactions should represent distinct plausible responses within the appropriate reaction space, $\hat{Y}_{i,q} \in \mathcal{A}(\mathcal{X}_i)$. Thus, offline MAFRG aims not only to generate multiple facial reactions, but to capture diverse valid facial reactions that remain appropriate to the same speaker context.

4 METHOD

4.1 THE PROPOSED MAM-REACTOR

Overview: As illustrated in Fig. 2, the proposed Mam-Reactor is designed to construct a finite AFR set that is both grounded in reliable interaction dynamics and coordinated to cover multiple valid reactions. To this end, the framework consists of three coupled components: *AFR Anchor Learning*, *Anchor-Preserving One-step Query Generation*, and *Balanced Set Alignment*, which are introduced in Sec. 4.2, Sec. 4.3, and Sec. 4.4, respectively, followed by a lightweight isolated VA refinement.

First, the *AFR Anchor Learning* module learns a deterministic anchor from the strictly paired AFR given a multimodal speaker behavior $\mathcal{X}_i$:

$$H_i = \mathcal{E}_{\text{anc}}(\mathcal{X}_i), \qquad Y_{i,0}^{\text{raw}} = \mathcal{G}_{\text{rel}}(H_i), \qquad \hat{Y}_{i,0} = \Gamma\big(Y_{i,0}^{\text{raw}}\big). \tag{2}$$

The anchor is subsequently frozen and serves as a shared fine-grained reference for all generated AFRs. Second, the *Anchor-Preserving One-step Query Generation* module constructs the remaining candidates as query-conditioned residual variations around the frozen anchor:

$$\Delta_{i,q} = \Delta_{i,q}^{\text{dyn}} + \Delta_{i,q}^{\text{sty}}, \qquad \hat{Y}_{i,q} = \Gamma\big(\text{sg}(Y_{i,0}^{\text{raw}}) + \Delta_{i,q}\big), \quad q = 1, \ldots, Q-1, \tag{3}$$

Together with the frozen anchor, these query-conditioned candidates form the complete AFR prediction set $\hat{\mathcal{Y}}_i = \{\hat{Y}_{i,q}\}_{q=0}^{Q-1}$. Expression-identified queries produce distinct AFR variations, while diversity regulation encourages the generated candidates to cover a broader range of AFRs. Third, the *Balanced Set Alignment* module constructs the prediction–target cost matrix between the predicted AFR set and the valid target AFR set:

$$C_{i,qk} = d\Big(\hat{Y}_{i,q}, Y_{i,k}\Big), \tag{4}$$

and obtain a balanced soft correspondence through Sinkhorn transport:

$$P_i^\star = \text{Sinkhorn}(C_i), \qquad \mathcal{T}(C_i) = \sum_{q,k} P_{i,qk}^\star C_{i,qk}. \tag{5}$$

The resulting balanced correspondence assigns different predictions to different valid AFRs through joint set-level optimization, encouraging complementary coverage rather than redundant predictions. In practice, $\mathcal{L}_{\text{align}}$ combines complementary reaction-style and FRD-oriented objectives instantiated through $\mathcal{T}(\cdot)$; exact definitions are provided in the Appendix. Together, these components achieve the finite-set objective by learning a reliable AFR anchor, generating anchor-preserving candidate reactions through one-step queries, and coordinating the prediction set with valid AFRs through balanced set alignment.

4.2 AFR ANCHOR-LEARNING

For a given $\mathcal{X}_i$, the valid target set $\mathcal{Y}_i$ contains facial reactions, but only the paired AFR Y_i^{pair} is strictly temporally aligned with the current speaker sequence. We therefore first learn a deterministic AFR anchor from this paired supervision, providing a reliable fine-grained reference. In contrast to directly generating multiple AFRs from unconstrained latent variations, subsequent candidates are constructed as controlled deviations from this anchor.

Specifically, $\mathcal{X}_i$ are first projected into a shared feature space and fused into a frame-wise representation. A temporal Transformer then models the complete speaker sequence and produces the contextual representation

$$H_i = \mathcal{E}_{\text{anc}}(\mathcal{X}_i), \qquad H_i \in \mathbb{R}^{T_i \times d_h}, \tag{6}$$

where $\mathcal{E}_{\text{anc}}$ denotes the anchor encoder. To account for the dependencies among different facial attributes, a relation-aware prediction head maps H_i to an unconstrained reaction representation $Y_{i,0}^{\text{raw}}$, followed by channel-specific output transformations:

$$Y_{i,0}^{\text{raw}} = \mathcal{G}_{\text{rel}}(H_i), \qquad \hat{Y}_{i,0} = \Gamma\big(Y_{i,0}^{\text{raw}}\big)\,. \tag{7}$$

Here, $\Gamma(\cdot)$ maps the raw outputs to the corresponding AU, VA, and expression spaces, yielding the deterministic AFR anchor $\hat{Y}_{i,0} \in \mathbb{R}^{T_i \times D}$. The anchor is optimized only with the strictly paired AFR and is frozen after pretraining. Consequently, both its temporal representation H_i and raw reaction prediction $Y_{i,0}^{\text{raw}}$ provide fixed references for the subsequent one-step query generation. The paired-anchor training objective and its regularization terms are described in Sec. 4.4.

4.3 One-step Query Generation

Given the frozen anchor representation H_i and $Y_{i,0}^{\text{raw}}$, we generate multiple AFRs through a set of parallel learnable queries. Specifically, we introduce eight expression-identified query embeddings $\{e_q\}_{q=1}^{8}$. To adapt each query to the current speaker behavior, H_i is first pooled and projected into the query space, $s_i = P_s(\text{Pool}(H_i))\,$, and each query is combined with the speaker context as

$$z_{i,q} = \phi([e_q,\, s_i,\, e_q \odot s_i])\,, \qquad q = 1, \ldots, 8, \tag{8}$$

where $\phi(\cdot)$ denotes the query–context interaction network and $\odot$ denotes element-wise multiplication. Each query condition $z_{i,q}$ is then injected into a bidirectional Mamba together with H_i. We decompose the query-dependent variation into a temporal residual and a clip-level style residual:

$$\Delta_{i,q}^{\text{dyn}} = \rho_d \tanh\big(W_d\big(\,\text{BiMamba}(H_i, z_{i,q})\big)\big)\,, \tag{9}$$

$$\Delta_{i,q}^{\text{sty}} = \rho_s \tanh(W_s(z_{i,q}))\,, \tag{10}$$

where $\Delta_{i,q}^{\text{dyn}}$ captures the reaction dynamics in time-varying, while $\Delta_{i,q}^{\text{sty}}$ introduces the variation at the sequence-level and is transmitted over the temporal dimension. Rather than generating each AFR independently, the two residuals are applied around the frozen anchor in its unconstrained output space:

$$\hat{Y}_{i,q} = \Gamma\Big(\text{sg}(Y_{i,0}^{\text{raw}}) + \Delta_{i,q}^{\text{dyn}} + \Delta_{i,q}^{\text{sty}}\Big)\,, \qquad q = 1, \ldots, 8, \tag{11}$$

where $\text{sg}(\cdot)$ denotes stop-gradient and $\Gamma(\cdot)$ maps the raw prediction to the valid AU, VA, and expression output spaces. This anchor-preserving formulation allows the learned queries to introduce distinct reaction variations while retaining a shared fine-grained AFR reference. In addition to the eight explicit queries, we construct a context-dependent mixture query

$$e_{i,9} = \sum_{q=1}^{8} \omega_{i,q} e_q, \qquad \sum_{q=1}^{8} \omega_{i,q} = 1, \tag{12}$$

where $\boldsymbol{\omega}_i$ is predicted from the pooled speaker context. The mixture query is processed through the same query-generation pathway. Together with the frozen anchor $\hat{Y}_{i,0}$, the model therefore produces the complete AFR prediction set $\hat{\mathcal{Y}}_i = \Big\{\hat{Y}_{i,0}, \hat{Y}_{i,1}, \ldots, \hat{Y}_{i,9}\Big\}$. All nine query-conditioned candidates are generated in parallel.

4.4 Training Strategies

Paired-AFR anchor pretraining. In the first training stage, the deterministic AFR anchor is optimized using only the strictly paired AFR Y_i^{pair}. We use

$$\mathcal{L}_{\text{anchor}} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{ccc}}\mathcal{L}_{\text{CCC}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}, \tag{13}$$

Table 1: Comparison under the official offline MAFRG evaluation protocol. All methods generate Q=10 AFRs. Baseline results are collected from the corresponding original papers and the official benchmark.

Method	Venue	**Appropriateness** FRC↑	**Appropriateness** FRD↓	Diversity FRDiv↑	Diversity FRVar↑	Synchrony FRSyn↓
GT	–	10.0000	0.0000	0.1876	0.0669	48.66
B_Random	–	0.0300	474.6800	0.3342	0.1671	46.64
B_Mime	–	0.5200	206.0200	0.0000	0.0766	43.70
B_MeanFR	–	0.0000	205.6500	0.0000	0.0000	49.00
Trans-VAE	–	0.3000	181.7200	0.0076	0.0083	49.00
REGNN Xu et al. (2026)	TAFFC'26	0.5400	<u>155.4900</u>	0.0012	0.0048	44.81
PerFRDiff Zhu et al. (2024)	ACM MM'24	0.5800	217.3200	<u>0.1596</u>	<u>0.0862</u>	48.66
SC-Diff Mao et al. (2025)	ACM MM'25	0.5883	214.6757	0.0402	0.0811	49.00
Wang Wang et al. (2025)	ACM MM'25	0.7100	173.0100	0.1405	0.0769	47.77
Huang Huang & Yu (2025)	ACM MM'25	0.6700	178.2700	0.1510	0.0801	48.06
MReactor Luo et al. (2026)	CVPR'26	<u>0.7800</u>	165.2600	**0.2931**	**0.1507**	46.89
Mam-Reactor (ours)	Ours	**1.0888**	**134.8100**	0.1521	0.0601	49.00

where

$$\mathcal{L}_{\mathrm{CCC}} = 1 - \frac{1}{D}\sum_{d=1}^{D} \frac{2\,\mathrm{Cov}(\hat{Y}_{i,0}^{d}, Y_{i}^{\mathrm{pair},d})}{\mathrm{Var}(\hat{Y}_{i,0}^{d}) + \mathrm{Var}(Y_{i}^{\mathrm{pair},d}) + \left(\mu_{\hat{Y}^{d}} - \mu_{Y^{d}}\right)^{2} + \epsilon}, \tag{14}$$

and

$$\mathcal{L}_{\mathrm{vel}} = \frac{1}{(T_i - 1)D}\sum_{t=2}^{T_i} \left\| \left(\hat{Y}_{i,0,t} - \hat{Y}_{i,0,t-1}\right) - \left(Y_{i,t}^{\mathrm{pair}} - Y_{i,t-1}^{\mathrm{pair}}\right) \right\|_{2}^{2}. \tag{15}$$

Let T_i denote the number of valid frames in the i-th sequence and $d \in \{1, \ldots, D\}$ index the facial-attribute dimension. Accordingly, $\hat{Y}_{i,0}^{d}$ and $Y_{i}^{\mathrm{pair},d}$ denote the temporal trajectories of the d-th facial attribute in the anchor prediction and paired AFR, respectively. $\mathcal{L}_{\mathrm{MSE}}$ constrains frame-wise facial-attribute values, $\mathcal{L}_{\mathrm{CCC}}$ encourages temporal agreement over each facial dimension, and $\mathcal{L}_{\mathrm{vel}}$ preserves local motion dynamics by matching first-order temporal differences. After pretraining, the anchor is frozen and provides the fixed reference $(H_i, Y_{i,0}^{\mathrm{raw}})$ for the subsequent query generation.

Candidate-set regularization. To obtain an identifiable and expressive AFR set, we further impose query-level regularization $\mathcal{L}_{\mathrm{qry}}$ on the learned queries and regulate the overall prediction-set diversity by

$$\mathcal{L}_{\mathrm{div}} = \left[m - \bar{D}(\hat{\mathcal{Y}}_i)\right]_{+}. \tag{16}$$

A channel-aware term $\mathcal{L}_{\mathrm{ch}}$ additionally regulates how this diversity is distributed across AU, VA, and expression channels, while $\mathcal{L}_{\mathrm{res}}$ limits excessive deviations from the frozen anchor. The set-learning objective is summarized as

$$\mathcal{L}_{\mathrm{main}} = \mathcal{L}_{\mathrm{align}} + \mathcal{L}_{\mathrm{qry}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}} + \lambda_{\mathrm{ch}}\mathcal{L}_{\mathrm{ch}} + \lambda_{\mathrm{res}}\mathcal{L}_{\mathrm{res}}. \tag{17}$$

Exact definitions of the query regularization, reaction costs, and other regularization terms are provided in the Appendix.

Isolated VA refinement. After the AFR-set structure has been established, we perform a lightweight VA-specific refinement at the end of the second stage. Only the VA parameters of the final residual output projection are updated, while the anchor, query embeddings, temporal backbone, and the remaining output parameters are kept fixed. The refinement uses a VA-specific banded Soft-DTW objective while preserving the learned reaction-set diversity and temporal variation. Its detailed objective is provided in the Appendix.

5 Experiments

5.1 Dataset and Evaluation

Dataset: We evaluate our Mam-Reactor on the offline MARS dataset from the REACT 2025 challenge Song et al. (2025), the first large-scale multi-modal benchmark specifically designed for the

Table 2: Component ablation and progressive construction of Mam-Reactor. *Anchor*, *Q-ID*, *Div.*, *Align.*, and *VA* denote the deterministic reaction anchor, expression-query identity supervision, diversity regulation, balanced prediction–target set alignment, and isolated valence–arousal refinement, respectively. The upper block progressively constructs the complete model, while the lower block isolates the roles of diversity regulation and set alignment. The query-based candidate generator is shared by all multi facial reaction variants; Q-ID indicates only the expression-query identity supervision.

	Components					Performance			
Variant	Anchor	Q-ID	Div.	Align.	VA	FRC ↑	FRD ↓	FRDiv ↑	FRVar ↑
Progressive construction									
P0: Anchor	✓	–	–	–	–	1.3029	121.23	0.0000	0.0563
P1: + Query generation	✓	✓	–	–	–	1.1595	130.11	0.0420	0.0548
P2: + Diversity	✓	✓	✓	–	–	1.0000	172.26	0.1561	0.0588
P3: + Set alignment	✓	✓	✓	✓	–	1.0879	137.21	0.1530	0.0601
P4: Mam-Reactor	✓	✓	✓	✓	✓	1.0888	134.81	0.1521	0.0601
Isolated controls									
Diversity only	✓	–	✓	–	–	0.9839	170.57	0.1565	0.0693
Set alignment only	✓	–	–	✓	–	1.0215	123.24	0.0736	0.0621

MAFRG task. It comprises 137 natural dyadic interaction clips (23 speakers, 137 listeners) partitioned into 2,856 session pairs across five diverse topics. For frame-level inputs, the dataset provides 768-D audio features extracted via wav2vec 2.0 Baevski et al. (2020), alongside comprehensive facial parameters including 58-D 3DMM coefficients, 15 AUs, 8 expression probabilities, and valence/arousal intensities. Crucially, MARS provides variable-length, multiple ground-truth AFRs per speaker behavior, faithfully reflecting the non-deterministic nature of real-world human interactions. Following the MARS protocol, AFRs recorded under the same session-specific context are treated as valid alternatives for a speaker clip. Only the paired facial reaction is strictly time-aligned with the speaker sequence; the remaining same-session facial reactions provide set-level supervision without treating them as strictly frame-aligned pairs.

Implementation details: We train Mam-Reactor on an NVIDIA A100 GPU. The anchor is first pretrained for 50 epochs and subsequently frozen. We then train the anchor-preserving candidate generator for 30 epochs using a two-layer bidirectional Mamba with a hidden dimension of 128 and eight 64-dimensional expression queries. The resulting diversity-regulated prediction set is further optimized through balanced prediction–target set alignment, followed by isolated VA refinement in the final epoch.

Evaluation metrics: We follow the official REACT evaluation protocol (Song et al., 2023a; 2024; 2025) and report FRCorr and FRDist for appropriateness, FRDiv and FRVar for diversity, and FRSyn for synchrony. For compactness, we denote FRCorr and FRDist as FRC and FRD, respectively.

5.2 Quantitative experiment

Table 1 shows that Mam-Reactor achieves the strongest appropriateness among all compared methods on both FRC and FRD, while maintaining substantial facial-reaction diversity. At nearly identical diversity to Huang et al. (FRDiv 0.1521 vs. 0.1510), Mam-Reactor increases FRC from 0.6700 to 1.0888 and reduces FRD from 178.27 to 134.81. Among prior methods, MReactor reports the highest FRC (0.7800), while REGNN obtains the lowest FRD (155.49); Mam-Reactor improves these two appropriateness measures to 1.0888 and 134.81, respectively. The matched-diversity comparison with Huang et al. further shows that these appropriateness gains are achieved without collapsing the prediction-set diversity.

5.3 Ablation Study

Table 2 shows how the facial-reaction set capability is progressively established. P0 provides a deterministic appropriateness reference but does not introduce inter-reaction diversity (FRDiv = 0). Introducing expression-identified queries in P1 establishes distinct candidate directions, increasing

Table 3: Fixed-checkpoint functional decomposition of Mam-Reactor. The dynamic temporal residual and static style residual are independently disabled at inference time while all learned parameters remain fixed.

Variant	Dynamic	Static	FRC ↑	FRD ↓	FRDiv ↑	FRVar ↑
Anchor only	–	–	1.3029	121.23	0.0000	0.0563
Dynamic only	✓	–	1.1715	125.91	0.1147	0.0617
Static only	–	✓	1.2862	124.78	0.0431	0.0588
Full Mam-Reactor	✓	✓	1.0888	134.81	0.1521	0.0601

Table 4: Effect of deterministic-anchor pretraining under a matched staged-training budget. Both variants follow the same subsequent candidate-set learning and refinement protocol. This controlled staged reference is trained independently from the selected main checkpoint reported in Table 1.

Initialization	FRC ↑	FRD ↓	FRDiv ↑	FRVar ↑
Random joint anchor	0.2406	182.76	0.1535	0.0142
Pretrained P0 anchor	0.9861	157.40	0.1535	0.0568

FRDiv to 0.0420. Diversity regulation in P2 further expands the candidate space to FRDiv = 0.1561. Crucially, balanced set-level alignment in P3 increases FRC from 1.0000 to 1.0879 and reduces FRD from 172.26 to 137.21, while preserving nearly the same diversity (0.1561 $\rightarrow$ 0.1530). The isolated VA refinement then further reduces FRD to 134.81 with essentially unchanged FRC and FRDiv (1.0879 $\rightarrow$ 1.0888 and 0.1530 $\rightarrow$ 0.1521). A complementary fixed-target-set diagnostic further confirms that this gain reflects improved set coverage rather than only prediction-wise appropriateness. Relative to P2, P3 reduces target-to-prediction DTW from 17.4792 to 17.0199 (-2.63%) and the balanced Hungarian set cost from 21.8696 to 20.0736 (-8.21%). These results directly support the complementary roles of the three stages: diversity regulation establishes an expressive candidate space, set-level alignment organizes this diversity toward valid AFRs, and VA refinement improves the resulting affective trajectories. The isolated controls reinforce this interpretation. Diversity regulation alone establishes a broad candidate space (FRDiv $= 0.1565$) while leaving FRD at 170.57. Together, these results show that diversity expansion and set-level coordination address complementary aspects of facial reaction set learning.

In table 3, we further decompose the learned candidate variation by independently disabling the dynamic temporal residual and the static style residual at inference time. The anchor alone collapses to a single facial reaction mode (FRDiv $= 0$). The dynamic branch recovers most of the full-model diversity (FRDiv $= 0.1147$), while the static branch contributes an additional global style shift (FRDiv $= 0.0431$). Combining both yields the complete facial reaction set diversity (FRDiv $= 0.1521$), indicating complementary roles in temporal and style-level candidate variation.

We next examine whether the deterministic anchor can be replaced by equal-budget joint training from random initialization in Table 4. At exactly the same inter-reaction diversity (FRDiv = 0.1535), deterministic P0 pretraining increases FRC from 0.2406 to 0.9861 and reduces FRD from 182.76 to 157.40. It also increases FRVar from 0.0142 to 0.0568. These results show that anchor pretraining provides a meaningful interaction-specific reaction prior rather than merely facilitating candidate separation.

Table 5 compares the computational cost of Mam-Reactor with representative deterministic and iterative generative baselines. Unlike diffusion- or flow-based methods that require repeated denoising or integration steps, Mam-Reactor constructs the complete reaction set in a single forward pass. Consequently, it requires only 41.15 minutes of training and 6.61 GiB of training memory, substantially reducing the cost of iterative generative baselines. At inference time, Mam-Reactor generates the full test set in 25.19 seconds with only 0.56 GiB memory. Although REGNN has slightly lower per-sample latency, it produces one reaction per sampling pass, whereas Mam-Reactor produces the complete multi facial reaction set in one pass.

Table 5: Single-GPU training cost and generation efficiency. Mam-Reactor and REGNN training times are measured from completed runs; PerFRDiff training time is extrapolated from its official 500-epoch protocol. Generation time covers all 1,142 test examples.

Method	Training Efficiency		Inference Efficiency			
	Train Time ↓	Peak Mem. ↓	Gen. Time ↓	ms/sample ↓	Peak Mem. ↓	FLOPs ↓
REGNN	17.66 h	12.18 GiB	**16.26 s**	**14.24 ms**	0.70 GiB	**28.11 G**
PerFRDiff (∼50-step Diffusion)	158.26 h	36.44 GiB	∼6 h	18.91 s	21.15 GiB	282.49 T
∼5-step Flow Matching	29.14 h	19.58 GiB	6.57 min	345.48 ms	7.73 GiB	14.14 T
Mam-Reactor	**41.15 min**	**6.61 GiB**	25.19 s	22.06 ms	**0.56 GiB**	51.05 G

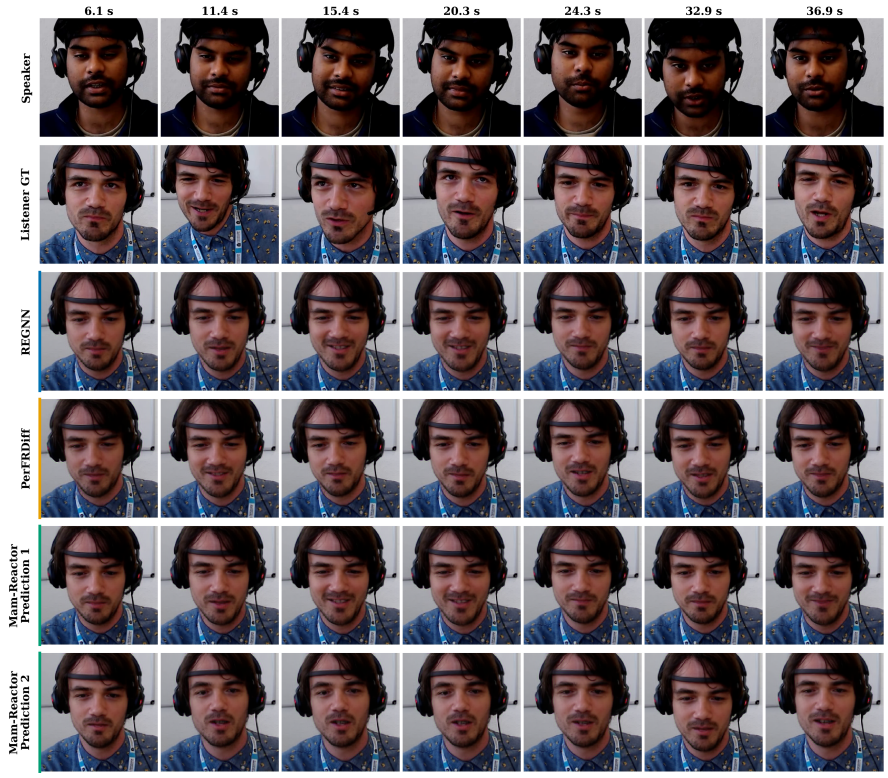


Figure 3: Qualitative comparison on the MARS benchmark.

5.4 Qualitative Experiment

Fig. 3 visualizes listener facial reactions at several timestamps under the same speaker behavior. In addition to the paired ground truth and representative baseline predictions, we show two candidates from the reaction set generated by Mam-Reactor. The two Mam-Reactor predictions follow coherent temporal evolution while exhibiting visibly different facial configurations, particularly in expression intensity and mouth dynamics. This qualitative difference reflects the goal of our set-level formulation: rather than forcing multiple candidates toward a single facial reaction pattern, Mam-Reactor allows different predictions to take complementary roles while remaining plausible under the same interaction context.

6 Conclusion

We present Mam-Reactor, an anchor-preserving one-step framework for offline MAFRG. By combining expression-identified query generation, diversity-regulated balanced set learning, and isolated VA refinement, Mam-Reactor constructs ten appropriate and diverse listener reactions from standard interaction data without additional cognitive annotations. Experiments on MARS demonstrate a strong FRC–FRD operating point together with lightweight training and efficient direct generation.

7 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.

8 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.

9 REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide supplementary code containing the data preprocessing pipeline, the implementation of Mam-Reactor, the forward inference procedure, and the model checkpoints required to reproduce the reported inference results. The main paper describes the training protocol, model configuration, and evaluation metrics, while additional implementation details and exact objective definitions are provided in the supplementary material.

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.

Ximi Hoque, Adamay Mann, Gulshan Sharma, and Abhinav Dhall. Beamer: Behavioral encoder to generate multiple appropriate facial reactions. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 9536–9540, 2023.

Jiajian Huang and Zitong Yu. Multiple appropriate facial reaction generation based on multi-view transformation of speaker video. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13992–13996, 2025.

Yuchi Huang and Saad Khan. A generative approach for dynamically varying photorealistic facial expressions in human-agent interactions. In *Proceedings of the 20th ACM international conference on multimodal interaction*, pp. 437–445, 2018.

Yuchi Huang and Saad M Khan. Dyadgan: Generating facial expressions in dyadic interactions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2259–2266. IEEE, 2017.

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.

Zhenjie Liu, Cong Liang, Jiahe Wang, Haofan Zhang, Yadong Liu, Caichao Zhang, Jialin Gui, and Shangfei Wang. One-to-many appropriate reaction mapping modeling with discrete latent variable. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5. IEEE, 2024.

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.

Jiachen Luo, Jiajun He, Shuai Shen, Lin Wang, Huy Phan, Joshua Reiss, Lin Haijun, Bjoern Schuller, Zeyu Fu, and Siyang Song. Mreactor: Offline multiple appropriate facial reaction generation with hierarchical cognitive disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3354–3363, June 2026.

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.

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, 2024a.

Minh-Duc Nguyen, Hyung-Jeong Yang, Ngoc-Huynh Ho, Soo-Hyung Kim, Seungwon Kim, and Ji-Eun Shin. Vector quantized diffusion models for multiple appropriate reactions generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5. IEEE, 2024b.

Minh-Duc Nguyen, Hyung-Jeong Yang, Soo-Hyung Kim, Ji-Eun Shin, and Seung-Won Kim. Latent behavior diffusion for sequential reaction generation in dyadic setting. In *International Conference on Pattern Recognition*, pp. 233–248. Springer, 2024c.

Behnaz Nojavanasghari, Yuchi Huang, and Saad Khan. Interactive generative adversarial networks for facial expression generation in dyadic interactions. *arXiv preprint arXiv:1801.09092*, 2018.

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, Hengde Zhu, Cheng Luo, Cristina Palmero, German Barquero, Sergio Escalera, Michel Valstar, Mohamed Daoudi, Tobias Baur, Fabien Ringeval, Andrew Howes, Elisabeth Andre, and Hatice Gunes. React 2025: the third multiple appropriate facial reaction generation challenge, 2025. URL https://arxiv.org/abs/2505.17223.

Siyang Song, Micol Spitale, Zijian Wu, Xiangyu Kong, Cheng Luo, Cristina Palmero, German Barquero, Sergio Escalera, Michel Valstar, Mohamed Daoudi, et al. React 2026: The fourth multiple appropriate facial reaction generation challenge: Personalised mafrg and appropriate eeg reaction prediction. *arXiv preprint arXiv:2606.07935*, 2026.

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. *IEEE Transactions on Affective Computing*, 2026.

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.

A Additional Details

A.1 MARS Dataset and Evaluation Metrics

MARS dataset. We evaluate on the Multi-modal Multiple Appropriate Reaction in Social Dyads (MARS) dataset introduced by the REACT 2025 Challenge (Song et al., 2025). MARS contains 137 human–human dyadic interactions recorded from 23 speakers and 137 listeners, resulting in 2,856 multi-modal interaction-session pairs. Each interaction is organized into 23 sessions covering five topics: cultural differences (4 sessions), movie-scene sharing (4), policy changes (4), quizzes and games (8), and scenario-based interviews (3). The official split approximately follows a 60%/20%/20% train/validation/test ratio with a subject-independent protocol. For each session, the conversational context is controlled by predefined prompts. Consequently, for a speaker behavior in a particular session, facial reactions expressed by different listeners under the same session are defined as its appropriate facial reactions (AFRs), i.e., the ground-truth reaction set. MARS provides raw audio-visual recordings together with frame-level descriptors, including 58-D 3DMM coefficients, 768-D wav2vec 2.0 audio features, and 25-D facial attributes consisting of 15 Action Units (AUs), 2 Valence–Arousal (VA) values, and probabilities of 8 categorical facial expressions.

Evaluation protocol. We follow the official REACT evaluator (Song et al., 2025). Let $\hat{\mathbf{Y}}_{i,q} \in \mathbb{R}^{T_i \times D}$ denote the q-th generated reaction for evaluation example i, where $Q = 10$ and $D = 25$, and let $\mathbf{Y}_{i,k}$ denote its k-th valid AFR. For variable-length AFRs, the released evaluator first uses its provided post-processor to convert each target to the temporal length of the corresponding prediction before computing the following metrics.

FRCorr (FRC). For two reactions $\mathbf{A}$ and $\mathbf{B}$, the evaluator computes the Concordance Correlation Coefficient independently over each facial dimension,

$$c(\mathbf{A}, \mathbf{B}) = \frac{1}{D} \sum_{d=1}^{D} \frac{2\rho_d \sigma_{A,d} \sigma_{B,d}}{\sigma_{A,d}^2 + \sigma_{B,d}^2 + (\mu_{A,d} - \mu_{B,d})^2 + \epsilon}, \quad \epsilon = 10^{-8}, \tag{18}$$

and matches every generated reaction to the valid AFR with the highest CCC:

$$\mathrm{FRC} = \frac{1}{N} \sum_{i=1}^{N} \sum_{q=0}^{Q-1} \max_k c(\hat{\mathbf{Y}}_{i,q}, \mathbf{Y}_{i,k}). \tag{19}$$

Higher FRC indicates stronger appropriateness. Since the Q matched scores are summed rather than averaged, the maximum FRC is $Q = 10$.

FRDist (FRD). The official reaction distance combines DTW distances over the three facial attribute groups as

$$d(\mathbf{A}, \mathbf{B}) = \frac{1}{15}\mathrm{DTW}(\mathbf{A}^{\mathrm{AU}}, \mathbf{B}^{\mathrm{AU}}) + \mathrm{DTW}(\mathbf{A}^{\mathrm{VA}}, \mathbf{B}^{\mathrm{VA}}) + \frac{1}{8}\mathrm{DTW}(\mathbf{A}^{\mathrm{Exp}}, \mathbf{B}^{\mathrm{Exp}}), \tag{20}$$

followed by

$$\mathrm{FRD} = \frac{1}{N} \sum_{i=1}^{N} \sum_{q=0}^{Q-1} \min_k d(\hat{\mathbf{Y}}_{i,q}, \mathbf{Y}_{i,k}), \tag{21}$$

where lower values indicate better appropriateness.

FRDiv and FRVar. FRDiv measures pairwise squared distances among the Q reactions generated for the same speaker behavior,

$$\mathrm{FRDiv} = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{q \neq r} \|\hat{\mathbf{Y}}_{i,q} - \hat{\mathbf{Y}}_{i,r}\|_F^2}{Q(Q-1)T_i D}, \tag{22}$$

while FRVar is the mean temporal variance of all generated facial attributes,

$$\mathrm{FRVar} = \frac{1}{NQD} \sum_{i=1}^{N} \sum_{q=0}^{Q-1} \sum_{d=1}^{D} \mathrm{Var}_t \left[\hat{Y}_{i,q,t,d} \right]. \tag{23}$$

Table 6: Main architecture and training hyperparameters of Mam-Reactor.

Hyperparameter	Value	Hyperparameter	Value
Model architecture			
Prediction set size Q	10	Query dimension	64
Mamba hidden dimension	128	Mamba state dimension	16
Mamba layers	2	Mamba convolution width	4
Mamba expansion factor	2	Style-adapter hidden dim.	128
Sequence length	750	AU / VA / Exp. scale	1.8 / 0.4 / 1.6
Training			
Anchor epochs / batch size	50 / 32	Anchor learning rate	1×10^{-4}
Candidate epochs / batch size	30 / 16	Candidate learning rate	2×10^{-4}
Alignment epochs / batch size	3 / 16	Alignment learning rate	2×10^{-6}
Optimizer	AdamW	Weight decay	1×10^{-4}
Training targets K	10	Precision / seed	BF16 / 1
Set regulation and alignment			
Query-identity weight	0.25	Compatibility weight	0.10
Mixture-query weight	0.25	Diversity weight	20
Channel-diversity weight	20	Diversity budget	0.15
Style-transport weight	0.20	Style Sinkhorn temp. / iter.	0.1 / 8
FRD-surrogate weight	2	FRD surrogate stride	12
FRD-transport weight	8	FRD Sinkhorn temp. / iter.	0.1 / 32
FRD risk target R_0	140	Residual regularization	0.005
Isolated VA refinement			
VA learning rate	1×10^{-5}	Activation epoch	3
Trainable VA scalars	258	Preservation ratio	0.99
Mean-risk / proximal weight	0.25 / 0.05	Gradient clipping	0.1

Higher values indicate greater inter-reaction diversity and temporal variation, respectively.

FRSyn. Synchrony is measured by Time-Lagged Cross-Correlation (TLCC). Following the released evaluator, for each example it computes

$$r_i(\ell) = \frac{1}{D} \sum_{d=1}^{D} \mathrm{Corr}\left(\hat{Y}_{i,0,:,d}, X_{i,:,d}; \ell\right), \qquad \ell \in \{-49, \ldots, 49\}, \tag{24}$$

between the first generated reaction and the speaker facial attributes $\mathbf{X}_i$, and reports

$$\mathrm{FRSyn} = \frac{1}{N} \sum_{i=1}^{N} \left| \arg\max_{\ell} r_i(\ell) \right|. \tag{25}$$

FRSyn is expressed in frames, with lower values indicating tighter temporal synchrony.

A.2 Implementation Details

We summarize the main implementation and training hyperparameters in Table 6. Unless otherwise specified, we use BF16 training with random seed 1.

A.3 Training Target-Set Construction

For each speaker sequence X_i, we construct a training target set

$$\mathcal{Y}_i = \{Y_{i,0}, Y_{i,1}, \ldots, Y_{i,K-1}\}, \qquad K = 10. \tag{26}$$

The strictly paired listener reaction is always placed at $Y_{i,0}$. The remaining $K - 1$ targets are selected from listener reactions recorded under the same session-specific context. Thus, $Y_{i,0}$ provides the unique strictly frame-aligned supervision, whereas the remaining reactions provide valid same-session response alternatives without being treated as strictly frame-aligned pairs.

During training, same-session alternatives are ordered using a deterministic epoch-dependent shuffle indexed by the training seed, epoch, and source identity. This makes target-set sampling reproducible while exposing the model to different valid alternatives across epochs. If fewer than $K-1$ alternatives are available, repeated placeholders are used only to preserve a fixed tensor shape. Such placeholders are explicitly marked invalid and receive zero mass in all set-level objectives.

The paired target uses the same temporal crop as the corresponding speaker sequence. In contrast, each non-paired target is independently cropped using a deterministic target-specific offset. Let $m_{i,k,t} \in \{0,1\}$ denote the validity of frame t in target k, and define

$$v_{i,k} = \mathbb{I}\left[\sum_t m_{i,k,t} > 0\right], \qquad K_i = \sum_{k=0}^{K-1} v_{i,k}, \tag{27}$$

where K_i is the number of valid target reactions available for example i.

A.4 Balanced Prediction–Target Set Alignment

Validity-aware balanced transport. Let $C_i \in \mathbb{R}^{Q\times K}$ denote a prediction–target cost matrix for the Q generated reactions and K target reactions. Mam-Reactor assigns uniform mass to all predictions,

$$\mu_q = \frac{1}{Q}, \tag{28}$$

and distributes the target mass uniformly over only the valid targets,

$$\nu_{i,k} = \begin{cases} \frac{1}{K_i}, & v_{i,k} = 1, \\ 0, & v_{i,k} = 0. \end{cases} \tag{29}$$

The feasible transport set is therefore

$$\mathcal{U}_i = \left\{P \geq 0 : P\mathbf{1}_K = \boldsymbol{\mu},\ P^\top \mathbf{1}_Q = \boldsymbol{\nu}_i\right\}. \tag{30}$$

We obtain the balanced correspondence through entropy-regularized optimal transport,

$$P_i^\star = \arg\min_{P\in\mathcal{U}_i} \langle P, C_i\rangle - \tau H(P), \tag{31}$$

where

$$H(P) = -\sum_{q,k} P_{qk}\log P_{qk}. \tag{32}$$

We solve Eq. equation 31 using log-domain Sinkhorn normalization. Unless otherwise specified, we use temperature $\tau = 0.1$ and eight Sinkhorn iterations. Invalid target columns are assigned zero transport mass.

Given a cost matrix C_i, the corresponding balanced set cost is

$$\mathcal{T}(C_i) = \sum_{q=0}^{Q-1}\sum_{k=0}^{K-1} P^\star_{i,qk} C_{i,qk}. \tag{33}$$

Unlike independent nearest-target supervision, the balanced marginals explicitly coordinate the complete prediction set: multiple predictions cannot all concentrate their mass on a single easy target while leaving other valid targets uncovered.

A.4.1 Style-Level Reaction Cost

To compare non-paired reactions without requiring framewise alignment, we first construct a length-invariant reaction descriptor

$$s(Y) = \left[\mu(Y), \sigma(Y), \overline{|\Delta Y|}\right] \in \mathbb{R}^{75}, \tag{34}$$

where the three terms are respectively the temporal mean, temporal standard deviation, and mean absolute first-order velocity of the 25 facial attributes.

Let $\bar{s}(Y)$ denote this descriptor after normalization using training-set statistics. The style-level prediction–target cost is

$$C^{\mathrm{sty}}_{i,qk} = \frac{1}{75} \left\| \bar{s}(\hat{Y}_{i,q}) - \bar{s}(Y_{i,k}) \right\|_2^2, \tag{35}$$

and the corresponding balanced set loss is

$$\mathcal{L}_{\mathrm{sty}} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{T}\left(C^{\mathrm{sty}}_i\right). \tag{36}$$

A.4.2 Sequence-Level CCC Diagnostic

For diagnostic purposes, we additionally measure prediction–target trajectory affinity using the temporal Concordance Correlation Coefficient (CCC),

$$C^{\mathrm{ccc}}_{i,qk} = 1 - \mathrm{CCC}\left(\hat{Y}_{i,q}, Y_{i,k}\right). \tag{37}$$

In the reported configuration, this matrix is computed from stop-gradient predictions and is used only to monitor sequence-level affinity. It therefore does not contribute gradients to the optimized set-alignment objective.

A.4.3 FRD-Oriented Training Cost

The official FRD uses DTW over the AU, VA, and expression groups. Computing full differentiable DTW over all prediction–target pairs at every training step is expensive. We therefore use a crop-relative differentiable surrogate that preserves the official channel grouping and prediction-wise target-selection structure while uniformly sampling the independently cropped trajectories. For non-paired reactions, the sampled temporal index denotes relative crop position rather than strict synchronization with the current speaker frames. All reported FRD values are computed using the official exact DTW evaluator.

The official FRD evaluation uses three facial-attribute groups: AU, VA, and expression. We preserve this grouping during optimization with weights

$$w_{\mathrm{AU}} = \frac{1}{15}, \qquad w_{\mathrm{VA}} = 1, \qquad w_{\mathrm{Exp}} = \frac{1}{8}. \tag{38}$$

For efficiency, the surrogate uniformly samples every $s = 12$ frames. Let $\mathcal{S}_{i,k}$ denote the sampled valid positions of target k, $\tilde{n}_{i,k} = |\mathcal{S}_{i,k}|$, and $n_{i,k}$ the corresponding number of full-resolution valid frames. For group g, we define

$$d^{(g)}_{i,qk} = \sqrt{\frac{n_{i,k}}{\tilde{n}_{i,k}} \sum_{t \in \mathcal{S}_{i,k}} \left\| \hat{Y}^{(g)}_{i,q,t} - Y^{(g)}_{i,k,t} \right\|_2^2 + \epsilon}. \tag{39}$$

The first 15 AU channels use straight-through rounding in the forward pass so that the training geometry follows the discrete AU representation used by the official evaluator while retaining gradients.

The pairwise FRD-oriented cost is

$$C^{\mathrm{frd}}_{i,qk} = \alpha \sum_{g \in \{\mathrm{AU,VA,Exp}\}} w_g d^{(g)}_{i,qk}, \tag{40}$$

where α is a fixed scale-calibration factor.

We first define the prediction-wise best-target risk

$$R^{\min}_i = \sum_{q=0}^{Q-1} \min_{k: v_{i,k}=1} C^{\mathrm{frd}}_{i,qk}, \tag{41}$$

and optimize the normalized hinge

$$\mathcal{L}_{\mathrm{frd}} = \frac{1}{B} \sum_i \left[\frac{R^{\min}_i}{R_0} - 1 \right]_+, \qquad R_0 = 140. \tag{42}$$

To explicitly coordinate this cost at the set level, we further compute

$$P_i^{\mathrm{frd}} = \mathrm{Sinkhorn}\left(\frac{C_i^{\mathrm{frd}}}{R_0/Q}\right), \tag{43}$$

using temperature 0.1 and 32 Sinkhorn iterations. The balanced FRD risk is

$$R_i^{\mathrm{OT}} = Q \sum_{q,k} P_{i,qk}^{\mathrm{frd}} C_{i,qk}^{\mathrm{frd}}, \tag{44}$$

with

$$\mathcal{L}_{\text{frd-OT}} = \frac{1}{B} \sum_i \left[\frac{R_i^{\mathrm{OT}}}{R_0} - 1\right]_+ . \tag{45}$$

The balanced set-level component is

$$\boxed{\mathcal{L}_{\mathrm{set}} = \lambda_{\mathrm{sty}} \mathcal{L}_{\mathrm{sty}} + \lambda_{\mathrm{frdOT}} \mathcal{L}_{\text{frd-OT}}}, \tag{46}$$

where

$$\lambda_{\mathrm{sty}} = 0.2, \qquad \lambda_{\mathrm{frdOT}} = 8. \tag{47}$$

We further retain the prediction-wise FRD-oriented risk as an appropriateness regularizer:

$$\boxed{\mathcal{L}_{\mathrm{align}} = \mathcal{L}_{\mathrm{set}} + \lambda_{\mathrm{frd}} \mathcal{L}_{\mathrm{frd}}}, \qquad \lambda_{\mathrm{frd}} = 2. \tag{48}$$

Importantly, all Sinkhorn correspondences and target reactions are required only during training. They introduce no inference-time matching or iterative optimization.

A.5 Candidate-Set Regularization

A.5.1 Expression-Identified Query Supervision

The eight explicit queries $q = 1, \ldots, 8$ are assigned fixed expression identities corresponding to the eight expression channels provided by MARS. For the q-th generated reaction, we compute its clip-level normalized expression profile

$$\pi_{i,q,e} = \frac{\sum_t m_{i,t} \hat{Y}^{\mathrm{Exp}}_{i,q,t,e}}{\sum_{e'} \sum_t m_{i,t} \hat{Y}^{\mathrm{Exp}}_{i,q,t,e'} + \epsilon}, \tag{49}$$

where $e \in \{1, \ldots, 8\}$. Query q is associated with expression identity q, yielding

$$\mathcal{L}_{\mathrm{sem}} = -\frac{1}{8B} \sum_{i=1}^{B} \sum_{q=1}^{8} \log\left(\pi_{i,q,q} + \epsilon\right). \tag{50}$$

This objective makes the reaction queries identifiable while relying only on the standard expression probabilities already available in MARS; no additional cognitive or semantic annotation is introduced.

In addition to the expression-identity objective $\mathcal{L}_{\mathrm{sem}}$, a compatibility head predicts a context-dependent compatibility distribution $\boldsymbol{\omega}_i$ from the pooled speaker context:

$$\boldsymbol{\omega}_i = \mathrm{softmax}(\mathbf{a}_i), \qquad e_{i,9} = \sum_{q=1}^{8} \omega_{i,q} e_q. \tag{51}$$

We use a target-support compatibility term $\mathcal{L}_{\mathrm{comp}}$ and a mixture-query consistency term $\mathcal{L}_{\mathrm{mix}}$, and collect the query-level objectives as

$$\mathcal{L}_{\mathrm{qry}} = \lambda_{\mathrm{sem}} \mathcal{L}_{\mathrm{sem}} + \lambda_{\mathrm{comp}} \mathcal{L}_{\mathrm{comp}} + \lambda_{\mathrm{mix}} \mathcal{L}_{\mathrm{mix}}, \tag{52}$$

where

$$\lambda_{\mathrm{sem}} = 0.25, \qquad \lambda_{\mathrm{comp}} = 0.10, \qquad \lambda_{\mathrm{mix}} = 0.25. \tag{53}$$

Specifically, $\mathcal{L}_{\mathrm{comp}}$ is a binary cross-entropy loss between the predicted compatibility probabilities and the expression support observed in the valid target set, while $\mathcal{L}_{\mathrm{mix}}$ matches the mixture-query expression profile to the normalized aggregate target expression support.

A.5.2 Global Diversity Regulation

We regulate candidate-set diversity directly in the geometry of the official FRDiv metric. Let $\tilde{Y}$ denote the generated reaction after straight-through rounding of the AU channels. For example i, the prediction-set diversity is

$$D_i^{\text{pred}} = \frac{\sum_{q \neq r} \left\| \tilde{Y}_{i,q} - \tilde{Y}_{i,r} \right\|_F^2}{Q(Q-1)T_i D}. \tag{54}$$

The batch-level diversity is

$$\bar{D}^{\text{pred}} = \frac{1}{B} \sum_i D_i^{\text{pred}}, \tag{55}$$

and we use the budget objective

$$\boxed{\mathcal{L}_{\text{div}} = \left[m - \bar{D}^{\text{pred}} \right]_+}, \qquad m = 0.15. \tag{56}$$

The hinge formulation encourages the candidate set to reach the desired diversity range without continuously pushing reactions farther apart after the budget has been satisfied.

A.5.3 Channel-Aware Diversity Allocation

A single global diversity value does not specify how variation should be distributed across AUs, VA, and expression. We therefore additionally match the group-wise diversity composition of the valid target reaction set.

Let

$$D_{i,g}^{\text{pred}}, \qquad D_{i,g}^{\text{tar}}, \qquad g \in \{\text{AU}, \text{VA}, \text{Exp}\}, \tag{57}$$

denote the prediction- and target-set contributions to FRDiv. All groups retain the same $T_i D$ denominator so that

$$\sum_g D_{i,g} = D_i. \tag{58}$$

We first define the target-set diversity proportions

$$r_{i,g} = \frac{D_{i,g}^{\text{tar}}}{\sum_{g'} D_{i,g'}^{\text{tar}} + \epsilon}. \tag{59}$$

The desired total diversity is capped by the global budget,

$$D_i^\star = \min \left(\sum_g D_{i,g}^{\text{tar}}, m \right), \tag{60}$$

and the desired group-wise allocation is

$$D_{i,g}^\star = r_{i,g} D_i^\star. \tag{61}$$

For examples whose target set contains valid non-zero diversity, we optimize

$$\boxed{\mathcal{L}_{\text{ch}} = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} \frac{1}{3} \sum_g \text{SmoothL1} \left(\frac{D_{i,g}^{\text{pred}}}{m}, \frac{D_{i,g}^\star}{m} \right)} \tag{62}$$

where $\mathcal{V}$ denotes the set of valid training examples. Thus, Mam-Reactor regulates not only the amount of reaction diversity but also its allocation across facial-reaction channels.

A.5.4 Residual Regularization

For each learned candidate, let

$$\Delta_{i,q} = \Delta_{i,q}^{\mathrm{dyn}} + \Delta_{i,q}^{\mathrm{sty}} \tag{63}$$

denote the gated residual in the unconstrained anchor output space. We limit unnecessarily large deviations from the frozen anchor through

$$\mathcal{L}_{\mathrm{res}} = \frac{1}{|\Delta|} \|\Delta\|_2^2 . \tag{64}$$

The Stage-2 candidate-set objective is

$$\boxed{\mathcal{L}_{\mathrm{cand}} = \mathcal{L}_{\mathrm{qry}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}} + \lambda_{\mathrm{ch}}\mathcal{L}_{\mathrm{ch}} + \lambda_{\mathrm{res}}\mathcal{L}_{\mathrm{res}}}. \tag{65}$$

During Stage 3, balanced set alignment is activated while the candidate-set regularizers are retained:

$$\boxed{\mathcal{L}_{\mathrm{main}} = \mathcal{L}_{\mathrm{align}} + \mathcal{L}_{\mathrm{qry}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}} + \lambda_{\mathrm{ch}}\mathcal{L}_{\mathrm{ch}} + \lambda_{\mathrm{res}}\mathcal{L}_{\mathrm{res}}}. \tag{66}$$

with

$$\lambda_{\mathrm{div}} = 20, \qquad \lambda_{\mathrm{ch}} = 20, \qquad \lambda_{\mathrm{res}} = 0.005. \tag{67}$$

A.6 Isolated Valence–Arousal Refinement

After the prediction-set structure has been established, we perform a restricted VA refinement. This stage does not introduce a new model component. Only the two VA rows of the existing dynamic residual output projection are updated. With hidden dimension 128, this corresponds to

$$2 \times 128 + 2 = 258 \tag{68}$$

trainable scalars. The anchor, query embeddings, temporal backbone, static style adapter, and all non-VA output rows remain frozen.

For each refinement micro-step, we retain a stop-gradient reference reaction set

$$\hat{\mathcal{Y}}_i^{\mathrm{ref}} = \mathrm{sg}\left(\hat{\mathcal{Y}}_i\right). \tag{69}$$

The VA-specific risk is applied only to the nine learned candidates $q = 1, \ldots, 9$; the deterministic anchor $q = 0$ remains unchanged.

VA risk. Let R_i^{VA} and $R_i^{\mathrm{VA,ref}}$ denote the VA-only reaction risks computed using a banded Soft-DTW surrogate. We use temporal stride 6, band ratio 0.15, and Soft-DTW smoothing parameter $\gamma = 0.1$. We define the relative risk

$$r_i = \frac{R_i^{\mathrm{VA}}}{R_i^{\mathrm{VA,ref}} + \epsilon}. \tag{70}$$

To emphasize examples with the largest VA mismatch, we use the mean of the worst 25% relative risks,

$$\mathcal{L}_{\mathrm{tail}} = \mathrm{Mean}\left(\mathrm{Top}_{25\%}\{r_i\}_{i=1}^{B}\right), \tag{71}$$

together with the average risk

$$\mathcal{L}_{\mathrm{mean}} = \frac{1}{B}\sum_i r_i. \tag{72}$$

Proximal constraint. We constrain the VA update around its pre-refinement solution. Define the reference VA residual energy

$$E_{\mathrm{ref}} = \mathrm{Mean}\left[\left(\hat{Y}_{i,q>0}^{\mathrm{ref,VA}} - \hat{Y}_{i,0}^{\mathrm{ref,VA}}\right)^2\right]. \tag{73}$$

The proximal penalty is

$$\mathcal{L}_{\mathrm{prox}} = \frac{\mathrm{Mean}\left[\left(\hat{Y}_{i,q>0}^{\mathrm{VA}} - \hat{Y}_{i,q>0}^{\mathrm{ref,VA}}\right)^2\right]}{E_{\mathrm{ref}} + \epsilon}. \tag{74}$$

Set-structure preservation. To preserve the reaction-set structure established before VA refinement, we explicitly constrain both inter-reaction diversity and temporal variation:

$$r_{\text{div}} = \frac{\text{FRDiv}(\hat{\mathcal{Y}})}{\text{FRDiv}(\hat{\mathcal{Y}}^{\text{ref}}) + \epsilon}, \tag{75}$$

$$r_{\text{var}} = \frac{\text{FRVar}(\hat{\mathcal{Y}})}{\text{FRVar}(\hat{\mathcal{Y}}^{\text{ref}}) + \epsilon}. \tag{76}$$

The preservation penalties are

$$\mathcal{L}_{\text{div-pres}} = [0.99 - r_{\text{div}}]_+^2 , \tag{77}$$

$$\mathcal{L}_{\text{var-pres}} = [0.99 - r_{\text{var}}]_+^2 . \tag{78}$$

The complete isolated VA objective is

$$\boxed{\mathcal{L}_{\text{VA}} = \mathcal{L}_{\text{tail}} + 0.25\mathcal{L}_{\text{mean}} + 0.05\mathcal{L}_{\text{prox}} + 10\mathcal{L}_{\text{div-pres}} + 10\mathcal{L}_{\text{var-pres}}}. \tag{79}$$

The isolated VA optimizer uses AdamW with learning rate 10^{-5}, zero weight decay, and gradient clipping at 0.1. It is activated only in the final training epoch. Since refinement updates only existing VA output parameters, the inference architecture and forward-pass complexity remain unchanged.

A.7 Staged Optimization and Inference

Mam-Reactor follows a staged optimization procedure aligned with the functional roles of its components.

Stage 1: deterministic anchor learning. The anchor is trained for 50 epochs using the strictly paired listener reaction with frame-level MSE, CCC, and first-order velocity supervision. After pretraining, all anchor parameters are frozen.

Stage 2: candidate-set construction. The expression-identified residual generator is trained for 30 epochs with batch size 16. The eight expression queries and one mixture query are generated in parallel around the frozen anchor. Query-level supervision, global diversity regulation, channel-aware diversity regulation, and residual regularization establish the expressive candidate set. Balanced prediction–target alignment is not activated in this stage.

Stage 3: balanced set alignment. Starting from the Stage-2 solution, the diversity-regulated candidate generator is further optimized for 3 epochs using the prediction–target set objectives in Sec. A.4, while retaining the candidate-set regularizers.

Final isolated VA refinement. In the final epoch, the restricted VA micro-update described in Sec. A.6 is activated.

Inference. All set-level matching and target-set objectives are used only during training. At inference time, Mam-Reactor evaluates the frozen anchor and all nine query-conditioned residual candidates in parallel:

$$\hat{\mathcal{Y}}_i = \left\{ \hat{Y}_{i,0}, \hat{Y}_{i,1}, \ldots, \hat{Y}_{i,9} \right\}. \tag{80}$$

Thus, the complete ten-reaction set is produced in a single forward pass, without Sinkhorn matching, target retrieval, iterative denoising, or test-time refinement.