MNCFR: EMBODIED SIMULATION-INSPIRED REACTION FORMATION AND MEMORY-AUGMENTED REGULATION FOR MULTIPLE APPROPRIATE FACIAL REACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In dyadic human–human interactions, observing a speaker's emotional expressions may elicit emotional contagion and facial mimicry in the listener, with embodied simulation considered a potential underlying mechanism. Existing Multiple Appropriate Facial Reaction Generation (MAFRG) methods typically learn direct mappings from the given speaker behavior to multiple appropriate facial reactions (AFRs), without explicitly modeling the perception-to-response transformation. Inspired by embodied simulation, we propose Multimodal Neural Context Formation and Regulation (MNCFR), a novel diffusion framework comprising two complementary components: the Reaction Formation Module (RFM) and the Memory-Augmented Regulation Module (MAR). Specifically, RFM uses listener-specific information derived from listener reaction history to transform the speaker's currently perceived multimodal behavior into an intermediate listener-side reaction representation and then into a generation condition for the current window. This explicit speaker-to-listener transformation models the perception-to-response pathway proposed in embodied simulation. The same intermediate representation is also fed to an auxiliary EEG prediction head, which predicts segment-aligned listener EEG and provides an additional listener-side training signal. Meanwhile, MAR applies diffusion-step-dependent gating to a bounded-memory readout of recent speaker representations and injects the resulting context into the RFM condition, thereby promoting smoother transitions between the generated listener facial reactions across consecutive time windows. By separating current reaction formation from recent-context regulation, MNCFR organizes complementary information across temporal scales to support coherent reaction generation. Experiments under the official offline and online protocols on MARS show that MNCFR achieves better overall performance than reported learning-based methods. Ablation and temporal analyses demonstrate the effectiveness of RFM and MAR, while EEG prediction analysis further shows that higher prediction quality based on RFM's intermediate listener-side reaction representation is associated with larger generation gains.

1 INTRODUCTION

In dyadic human–human interactions, observing others' emotional expressions may spontaneously elicit facial movements in the observer that resemble the observed expressions and influence the observer's affective state, phenomena commonly described as facial mimicry and emotional contagion (Dimberg et al., 2000; Wild et al., 2001; Hess & Fischer, 2014). Embodied simulation provides one account of such observation-to-response phenomena, suggesting that perceiving others' behaviors and emotions may activate corresponding sensorimotor and affective representations in the observer (Gallese, 2007; Niedenthal, 2007; Bastiaansen et al., 2009). From this perception-to-response perspective, MAFRG requires transforming continuously evolving multimodal speaker behavior into facial reactions appropriate to the ongoing interaction (Maatman et al., 2005; Morency et al., 2008; Zhou et al., 2022). Accordingly, modeling this task involves explicitly forming a listener facial reaction condition from perceived speaker behavior, rather than treating the speaker–listener relation solely as direct conditioning.

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.

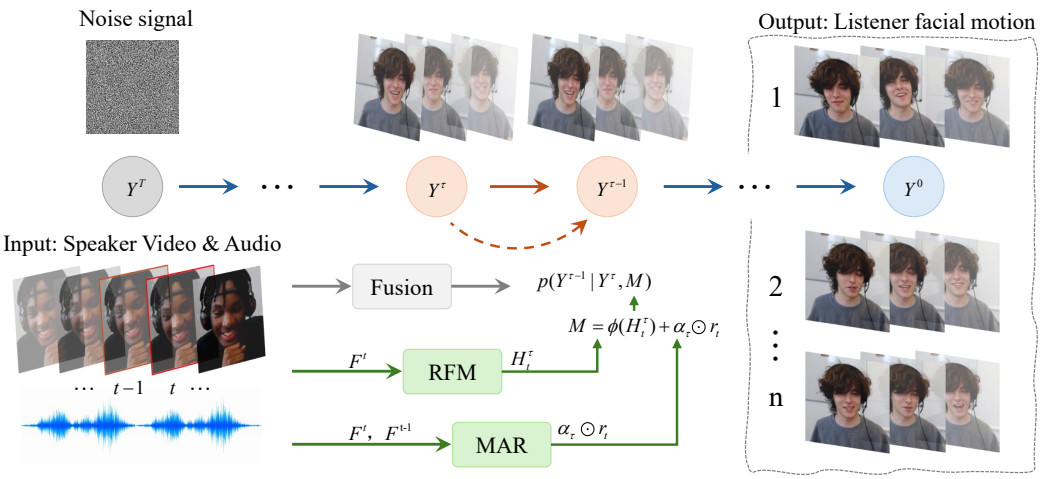


Figure 1: Overview of MNCFR. RFM forms a listener facial reaction condition from current multimodal speaker behavior, MAR regulates it using recent speaker context, and the shared diffusion generator produces listener facial reactions.

Existing MAFRG methods typically use current multimodal speaker behavior, including audio and facial behaviors, as the primary generation condition. Beyond current speaker behavior, these methods further incorporate listener reaction history and interaction context in different ways (Hoque et al., 2023; Zhu et al., 2024; Luo et al., 2025a). Speaker audio and facial behaviors are typically encoded and fused as generation conditions, while listener reaction history is used to provide listener-specific conditioning information (Hoque et al., 2023; Zhu et al., 2024; 2025). Besides, prior interaction information is incorporated to provide temporal context for continuous reaction generation (Luo et al., 2025a; Xie et al., 2025; Song et al., 2024).

Although these methods provide the generator with increasingly rich conditioning information, they commonly combine current speaker behavior, listener reaction history, and interaction context directly within an end-to-end generation process. Consequently, the perception-to-response transformation from current speaker behavior to a listener facial reaction condition remains implicit, making it difficult to explicitly constrain or examine the listener-side information formed during this process (Yu et al., 2023; Nguyen et al., 2024b; Li et al., 2025). Moreover, processing current speaker behavior and recent speaker context through the same conditioning pathway obscures their distinct roles in forming the current reaction and regulating generation across consecutive windows, leaving temporal continuity to be learned implicitly (Luo et al., 2025a; Xie et al., 2025; Luo et al., 2025b). Because existing methods are supervised primarily through final facial reaction outputs, their training objectives provide limited guidance on whether the internal representations capture listener-side response information. These limitations motivate an explicit organization of listener facial reaction conditioning in which current speaker behavior and listener-specific information support reaction formation, while recent speaker context provides complementary temporal regulation.

To address these limitations, we propose Multimodal Neural Context Formation and Regulation (MNCFR), an embodied imulation-inspired diffusion framework that organizes appropriate facial reactions (AFRs) generation and contextual regulation through RFM and MAR. Specifically, RFM uses listener-specific information derived from listener reaction history to transform current multimodal speaker behavior into an intermediate listener-side reaction representation, which is subsequently mapped to an AFRs conditioning representation for the current generation window. This explicit speaker-to-listener transformation models the perception-to-response process motivated by embodied simulation. Meanwhile, MAR applies diffusion-step-dependent gating to a bounded-memory readout of recent speaker representations and injects the resulting context into the listener facial reaction condition formed by RFM, thereby promoting smoother transitions between the generated listener facial reactions across consecutive windows. This separation allows current speaker behavior and recent speaker context to play complementary roles in AFRs generation and temporal regulation, respectively. Furthermore, during training, a lightweight EEG prediction head predicts segment-aligned

listener EEG from RFM's intermediate representation, whose prediction loss provides auxiliary listener-side supervision, while the resulting predictions enable analysis of the relationship between listener EEG prediction quality and generation gains. Under the official offline and online protocols on MARS (Song et al., 2025), MNCFR achieved superior overall performances than reported learning-based methods, while ablation and temporal analyses further demonstrate the effectiveness of RFM and MAR. Figure 1 presents the overall framework and illustrates the complementary roles of RFM and MAR. Our main contributions are threefold:

- We propose RFM to explicitly model the perception-to-response transformation under listener-specific conditioning derived from listener reaction history.
- We propose MAR to regulate the RFM condition using bounded-memory speaker context and diffusion-step-dependent gating, separating current reaction formation from recent-context regulation.
- We propose an auxiliary EEG prediction objective to provide listener-side supervision for RFM's intermediate representation and to support analysis of the relationship between listener EEG prediction quality and generation gains. Under the official offline and online protocols on MARS (Song et al., 2025), MNCFR achieves better overall performance than reported learning-based methods, while ablation and temporal analyses further demonstrate the effectiveness of RFM and MAR.

2 Related Work

Multiple Appropriate Facial Reaction Generation: Research in social psychology has shown that facial and behavioral responses depend on social context, interaction roles, and the interpretation of others' emotional signals; consequently, similar observed behaviors do not necessarily elicit a single fixed response (Hess & Fischer, 2014; Bavelas et al., 2000; Paggio & Navarretta, 2013; Blomsma et al., 2022). MAFRG formalizes this response uncertainty as a one-to-many generation problem and models multiple appropriate reactions using probabilistic generation, discrete latent variables, and diffusion models (Song et al., 2023b; Liu et al., 2024; Li et al., 2025). Diffusion-based MAFRG methods build on established denoising and sampling formulations (Ho et al., 2020; Song et al., 2021; Ho & Salimans, 2022).

The REACT challenge series provides unified datasets, protocols, and evaluation frameworks for MAFRG (Song et al., 2023a; 2024; 2025). Later editions extend the task from offline generation to online and personalized settings and introduce listener EEG prediction (Song et al., 2025; 2026). Representative methods model reaction uncertainty and diversity through non-deterministic motion generation, vector-quantized representations, and reaction-distribution learning (Ng et al., 2022; Nguyen et al., 2024b; Xu et al., 2026).

Multimodal and Listener-Specific Conditioning: MAFRG methods commonly encode speaker audio, facial behavior, and head motion and combine the resulting features through multimodal fusion, cross-modal attention, or Transformer architectures (Hoque et al., 2023; Nguyen et al., 2024a; Liang et al., 2023). Personalized MAFRG methods additionally derive listener-specific conditioning from listener reaction history or adapt generator parameters to individual listeners (Zhu et al., 2024; 2025). Current speaker behavior therefore provides interaction cues, whereas listener reaction history provides information associated with the listener (Zhu et al., 2024; Luo et al., 2025a; Zhu et al., 2025). Most existing MAFRG methods fuse speaker cues and listener-related information within an end-to-end generation process. As a result, the transformation from current speaker behavior to a listener facial reaction condition remains implicit. MNCFR explicitly models this transformation by forming an intermediate listener-side reaction representation before constructing the generation condition.

Embodied Simulation and Temporal Context Modeling: Embodied simulation suggests that observing others' behavior and emotion may activate corresponding sensorimotor and affective representations in the observer (Gallese, 2007; Niedenthal, 2007; Bastiaansen et al., 2009). Existing MAFRG methods use discrete motion representations, personalized reaction embeddings, and listener reaction history to encode information related to listener reactions (Nguyen et al., 2024b; Zhu et al., 2024; 2025). MNCFR draws on embodied simulation to organize the transformation from current speaker behavior to an intermediate listener-side reaction representation. Online and continuous

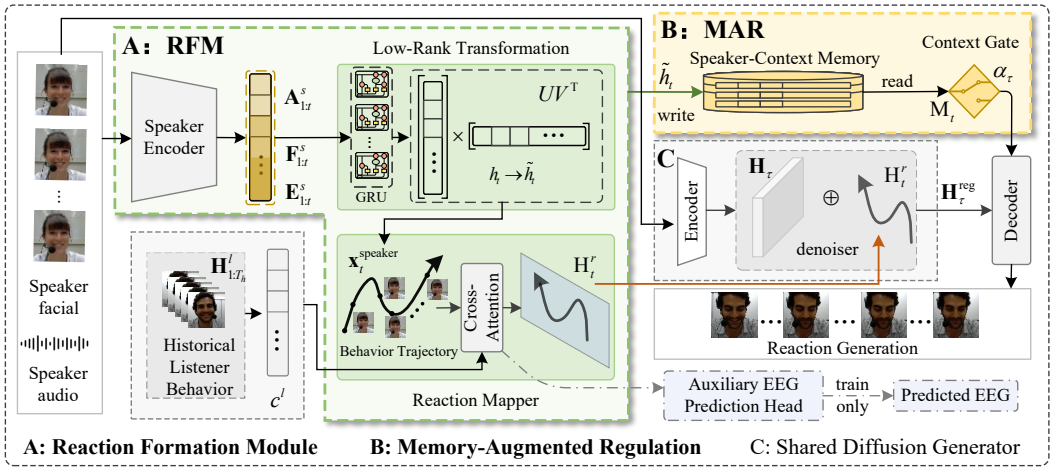


Figure 2: Detailed architecture of RFM and MAR. RFM combines current speaker behavior with listener reaction history to form the listener facial reaction condition, while MAR reads recent speaker-context memory and applies diffusion-step-dependent regulation.

MAFRG methods model temporal dependencies using historical states, previously generated reactions, cross-window feature propagation, temporal priors, or smoothness constraints (Nguyen et al., 2024c; Luo et al., 2025a; Xie et al., 2025). In MNCFR, RFM forms the listener facial reaction condition from current speaker behavior and listener-specific conditioning, whereas MAR uses recent speaker context to regulate generation across consecutive windows. The listener EEG prediction objective provides auxiliary supervision for the intermediate representation and supports analysis of the relationship between listener EEG prediction quality and generation gains (Song et al., 2026).

3 METHODOLOGY

For the current AFRs generation window indexed by t, let $X_t^s = \{A_t^s, F_t^s, E_t^s\}$ denote the current multimodal speaker behavior, where A_t^s, F_t^s, and E_t^s are the temporally aligned audio, facial behavior, and emotion features, respectively, with the superscripts s and l denote the speaker and listener, respectively. The superscripts s and l denote the speaker and listener. Let $H_{1:T_h}^l$ denote the available listener facial reaction history of length T_h. MNCFR conditions on X_t^s and the available listener facial reaction history $H_{1:T_h}^l$ to generate AFRs for the current window while preserving temporal continuity with the AFRs generated in preceding windows, where T_h denotes the history length.

As illustrated in Figure 2, the proposed MNCFR constructs the diffusion condition through two complementary modules: the Reaction Formation Module (RFM) and the Memory-Aware Regulation Module (MAR). Specifically, **RFM** uses listener-specific conditioning derived from the listener facial reaction history to transform the current multimodal speaker behavior into an intermediate listener-side reaction representation and a listener facial reaction condition; during training, the intermediate representation is additionally supervised by segment-aligned listener EEG.

$$(\bar{h}_t, H_t^r) = \mathrm{RFM}\big(X_t^s, H_{1:T_h}^l\big)\,, \tag{1}$$

RFM establishes the current speaker-to-listener response correspondence, whereas MAR complements this window-level reaction modelling with recent speaker context to regulate the generation process across windows. Specifically, MAR retrieves recent speaker representations from a bounded memory and produces a diffusion-step-dependent regulation signal:

$$r_{t,\tau} = \mathrm{MAR}(M_t, \tau). \tag{2}$$

The regulation signal is injected into the AFR conditioning representation produced by RFM:

$$C_{t,\tau} = \phi(H_t^r) + r_{t,\tau}. \tag{3}$$

Here, $\bar{h}_t$ denotes the intermediate listener-side representation, H_t^r is the AFR conditioning representation produced by RFM, $r_{t,\tau}$ is the diffusion-step-dependent regulation signal produced by MAR, and $C_{t,\tau}$ is the final conditioning representation supplied to the diffusion denoiser at diffusion step τ.

3.1 Appropriate Facial Reaction Conditioning

RFM explicitly models the transformation from perceived speaker behavior to a listener-side reaction condition. Given the current multimodal speaker behavior X_t^s and the available listener reaction history $H_{1:T_h}^l$, RFM first encodes the speaker behavior as a compact speaker-side representation $\tilde{h}_t$. It then incorporates listener-specific conditioning extracted from the listener reaction history to obtain an intermediate listener-side reaction representation $\bar{h}_t$ and maps this representation to the listener facial reaction condition H_t^r. During training, $\bar{h}_t$ is additionally supervised through an auxiliary EEG prediction task.

To encode the current speaker behavior, RFM projects and fuses the temporally aligned audio, facial behavior, and emotion features at each temporal position i in window t:

$$u_{t,i}^s = \psi\big([W_a A_{t,i}^s;\, W_f F_{t,i}^s;\, W_e E_{t,i}^s]\big)\,, \tag{4}$$

where W_a, W_f, and W_e are modality-specific projection matrices, $[\cdot\,;\cdot]$ denotes feature concatenation, and ψ is the feature-fusion function. The fused sequence is processed by a GRU and a low-rank recurrent accumulation:

$$h_{t,i} = \mathrm{GRU}(u_{t,i}^s, h_{t,i-1}), \qquad z_{t,i} = a \odot z_{t,i-1} + \tanh(h_{t,i} U V^\top + b). \tag{5}$$

The recurrence starts from $h_{t,0} = z_{t,0} = 0$ and proceeds over the L_t temporal positions in the current window. Its final accumulated state, $\tilde{h}_t = z_{t,L_t}$, provides the compact speaker-side representation of the current multimodal speaker behavior. Here, a is a learnable element-wise retention coefficient, while U and V parameterize the low-rank transformation.

To incorporate listener-specific information, the listener reaction history encoder maps the available history to a conditioning representation:

$$c^l = E_l(H_{1:T_h}^l). \tag{6}$$

RFM combines c^l with the current speaker-side representation $\tilde{h}_t$ to obtain the intermediate listener-side reaction representation $\bar{h}_t$ $\bar{h}_t$, which is then projected into the conditioning space of the diffusion denoiser:

$$\bar{h}_t = g(\tilde{h}_t, c^l), \qquad H_t^r = W_p \bar{h}_t. \tag{7}$$

Here, g denotes the transformation conditioned on listener reaction history, and W_p is the projection to the diffusion-conditioning space. Accordingly, $\bar{h}_t$ represents listener-side response information derived jointly from the current speaker behavior and the listener reaction history, whereas H_t^r serves as the condition for generating the listener facial reaction in the current window.

To provide direct listener-side supervision for the intermediate representation, an auxiliary EEG prediction head predicts the listener EEG segment aligned with generation window t:

$$\hat{e}_t^l = H_{\mathrm{eeg}}(\bar{h}_t), \tag{8}$$

where H_{eeg} is a lightweight prediction head and $\hat{e}_t^l$ denotes the predicted listener EEG. This auxiliary task constrains $\bar{h}_t$ with listener-side neural responses during training and enables subsequent analysis of the relationship between EEG prediction quality and facial reaction generation gains. The corresponding EEG prediction loss is introduced together with the overall training objective in the supplementary material.

3.2 Memory-Augmented Regulation Module

While RFM forms the listener facial reaction condition for the current window, continuous generation also requires speaker information accumulated across the current and preceding windows to regulate temporal transitions. MAR provides this complementary information by maintaining a bounded memory of recent speaker-side representations. It reads the stored representations into a speaker-context representation and modulates its contribution according to the current diffusion step before combining it with the RFM condition.

After the speaker encoder produces $\tilde{h}_t$ for window t, MAR appends it to a first-in-first-out memory with capacity C_M:

$$M_t = \mathrm{FIFO}_{C_M}(M_{t-1}, \tilde{h}_t). \tag{9}$$

Thus, M_t contains at most the C_M most recent speaker-side representations, including the representation of the current window. MAR concatenates the representations stored in M_t and maps them to a speaker-context representation:

$$r_t = \mathrm{Read}(\mathrm{Concat}(M_t)). \tag{10}$$

The memory is maintained across consecutive windows within the same interaction and reset at the beginning of each new interaction. Because M_t stores speaker-side representations produced before listener-specific conditioning is incorporated, r_t captures recent speaker context separately from the listener reaction history used by RFM.

MAR then modulates the speaker-context representation at each diffusion step τ. Given the diffusion-step embedding e_τ, a gating network computes

$$\alpha_\tau = \sigma(W_2 \tanh(W_1 e_\tau)), \qquad r_{t,\tau} = \alpha_\tau \odot r_t, \tag{11}$$

where W_1 and W_2 are learnable projection matrices, σ denotes the sigmoid function, and α_τ is an element-wise gate. The resulting $r_{t,\tau}$ represents the contribution of the speaker context at diffusion step τ.

Finally, MAR combines the regulated speaker context with the RFM condition:

$$C_{t,\tau} = \phi(H_t^r) + r_{t,\tau}, \tag{12}$$

where ϕ projects H_t^r into the denoiser conditioning space, and $C_{t,\tau}$ is the final condition supplied to the diffusion denoiser. Accordingly, RFM forms the listener facial reaction condition for the current window, whereas MAR regulates this condition with recent speaker context throughout denoising to support coherent transitions across consecutive generation windows.

4 Experiments

4.1 Experimental Setup

We evaluate MNCFR on the MARS benchmark under the official REACT offline and online MAFRG protocols. Each interaction contains temporally aligned speaker audio, speaker facial behavior, and listener facial reactions. Each listener segment is paired with a 14-dimensional processed listener EEG target and a validity mask for unavailable entries. The target supplies auxiliary supervision to the intermediate listener-side reaction representation during training. We report FRCorr and FRDist for listener facial reaction appropriateness, FRDiv for candidate diversity, FRVar for temporal variation, and FRSyn for speaker–listener synchrony. Higher FRCorr, FRDiv, and FRVar are preferred, whereas lower FRDist and FRSyn are preferred. The inherited diffusion setting uses $K = 10$ candidates, AdamW with a learning rate of 10^{-4}, batch size 4, at most 100 epochs, 10 denoising steps, RFM rank 2, and MAR capacity 2. Detailed preprocessing, complete seed results, and parameter sweeps are provided in the supplementary material.

4.2 Overall Performance

In the offline setting, MNCFR achieves the highest FRCorr among the reported learned models and the second-lowest FRDist. PerFRDiff retains higher candidate diversity and temporal variation, while REGNN obtains lower FRDist and FRSyn. The main advantage of MNCFR therefore lies in listener facial reaction appropriateness rather than uniform superiority across all metrics. In the online setting, MNCFR improves both appropriateness measures over the reported Trans-VAE and PerFRDiff results, while PerFRDiff obtains the lower FRSyn. Together, the two protocols establish the overall effectiveness of the complete RFM–MAR framework.

4.3 Analysis of Reaction Formation

The shared-backbone control obtains 0.51 FRCorr and 214.20 FRDist. Enabling the RFM condition raises FRCorr to 0.57 and reduces FRDist to 199.33, while the complete MNCFR reaches 0.80

Table 1: Official MARS comparisons. Published baseline values follow the REACT 2025 report. MNCFR offline results are mean ± standard deviation over three seeds; online results are point estimates. A dash denotes an unreported result.

(a) Offline MAFRG

Method	FRCorr↑	FRDist↓	FRDiv↑	FRVar↑	FRSyn↓
Random	0.03	474.68	0.3342	0.1671	46.64
Mime	0.52	206.02	0.0000	0.0766	43.70
Mean face	0.00	205.65	0.0000	0.0000	49.00
Trans-VAE	0.30	181.72	0.0076	0.0083	49.00
PerFRDiff	0.58	217.32	**0.1596**	**0.0862**	48.66
REGNN	0.54	**155.49**	0.0012	0.0048	**44.81**
MNCFR	$\mathbf{0.802} \pm 0.009$	170.43 ± 1.93	0.1250 ± 0.0010	0.0770 ± 0.0010	47.48 ± 0.16

(b) Online MAFRG

Method	FRCorr↑	FRDist↓	FRDiv↑	FRVar↑	FRSyn↓
Trans-VAE	0.25	212.22	0.1082	0.0564	48.20
PerFRDiff	0.55	222.69	**0.1506**	**0.0841**	**46.56**
MNCFR	**0.715**	**134.40**	0.1304	0.0831	48.50

and 170.43. This comparison supports the usefulness of explicitly transforming current multimodal speaker behavior into a listener facial reaction condition before diffusion generation. The complete model further benefits from the complementary recent-context regulation path.

4.4 Analysis of the Intermediate Listener-Side Reaction Representation

This analysis compares listener EEG prediction from the speaker-side representation and the intermediate listener-side reaction representation, evaluates aligned, shuffled, detached, and no-EEG controls, and relates EEG prediction quality to generation gains. The intended interpretation is computational: segment-aligned listener EEG supplies an additional listener-side learning signal, and EEG prediction provides a representation-analysis dimension. Figure 3 presents a complementary mechanism-oriented analysis of temporal modulation, EEG validity, and framewise reaction-error variation. The coordinated patterns across these signals provide empirical support for the embodied-simulation perspective: listener-side physiological organization offers informative supervision for learning and analyzing reaction representations.

4.5 Analysis of Recent-Context Regulation

Removing MAR increases within-window jerk from 0.0776 to 0.0948 and global jerk from 0.0936 to 0.1155. It also reduces diagnostic candidate separation from 0.1445 to 0.1185 and Var_K from 0.0546 to 0.0415. These diagnostics show that the independent MAR path is associated with smoother motion within and across generation windows while retaining greater separation among generated candidates. The component analysis provides a complementary role analysis: recent speaker context is most useful as a regulation signal applied to a listener facial reaction condition already formed by RFM. The supplementary material reports the full component table, the fixed-setting MAR-removal table, the MAR memory-capacity visualization, the appropriateness–diversity control sweep, and the paired sequence-level MAE analysis.

The memory-capacity sweep in Figure 4(a) varies nonmonotonically across the five metrics, indicating that a longer context is not uniformly beneficial; the main setting uses capacity two as a bounded recent-context path. Figure 4(b) provides a complementary sequence-level comparison over 780 test sequences. MNCFR reduces mean MAE from 0.2383 to 0.2189 relative to the matched shared-backbone control, with a paired effect size of $d_z = 0.545$. Figure 4(c) shows that sampling controls expose an appropriateness–diversity trade-off: for example, guidance 0.50 raises FRDiv from 0.126 to 0.145 but lowers FRCorr from 0.812 to 0.650 relative to the default setting in the fixed control sweep.

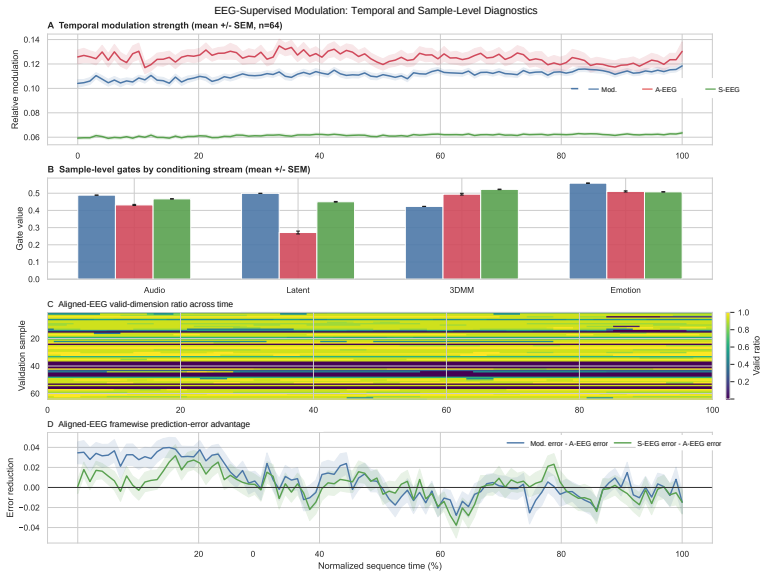
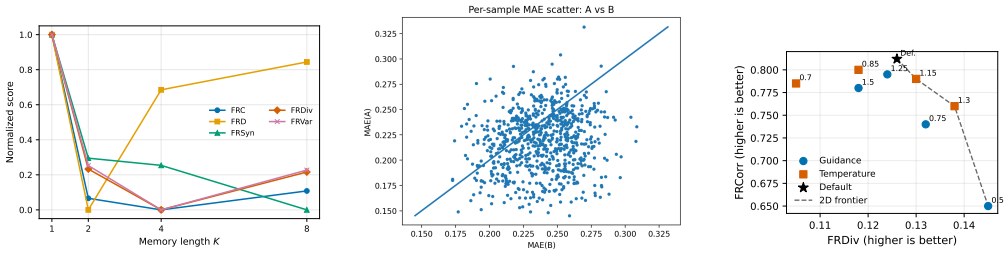


Figure 3: Mechanism-oriented temporal analysis of EEG-supervised reaction representation learning. The four panels show modulation strength, stream-wise gates, EEG validity, and framewise reaction-error differences, jointly demonstrating the value of listener-side physiological organization for embodied reaction modeling.



(a) MAR memory capacity (b) Paired sequence-level MAE (c) Generation-control trade-off

Figure 4: Key diagnostic visualizations. (a) Per-metric normalized values for MAR memory capacities $\{1, 2, 4, 8\}$; normalization is metric-specific. (b) Per-sequence MAE for MNCFR and the matched shared-backbone control, where points below the diagonal favor MNCFR. (c) FRCorr–FRDiv trade-offs under guidance and temperature controls; each point is one controlled single-run configuration.

4.6 Perceptual and Qualitative Evaluation

Sixty evaluators rated 15 generated clips, producing 900 clip-level ratings for each MOS criterion. A separate 2AFC study involved 25 evaluators and 375 paired judgments. The 2AFC result measures perceptual acceptability relative to the recorded listener response rather than reconstruction of a unique ground-truth trajectory.

Figure 5 displays generated listener facial reactions across time. The supplementary material further presents two representative failure patterns. In the first, the generated reaction follows a plausible pattern but exhibits excessive intensity, indicating limited calibration of reaction magnitude. In the second, head pose or facial appearance gradually drifts over time, revealing remaining limitations in long-range visual consistency. Together, the perceptual scores and qualitative examples indicate that MNCFR produces listener facial reactions that are generally natural, contextually appropriate, and temporally coherent, while these failure cases delimit the remaining visual errors.

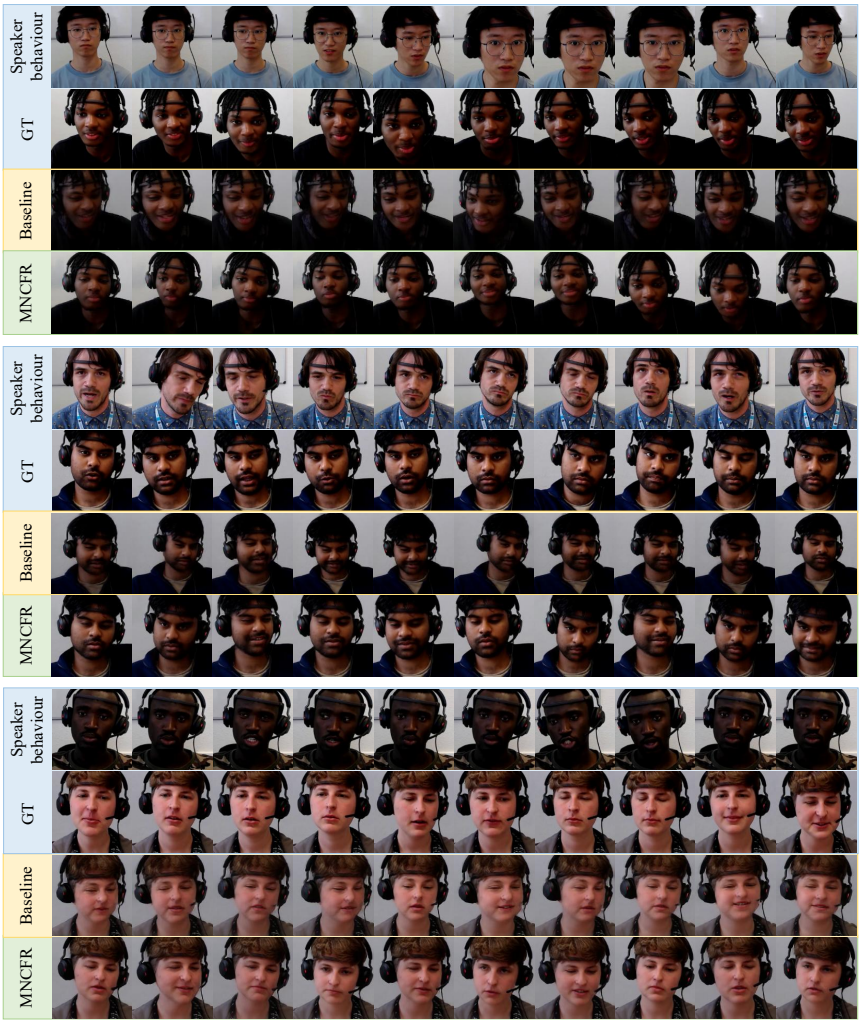


Figure 5: Representative qualitative results showing how MNCFR generates contextually appropriate and temporally coherent listener reactions from evolving speaker behavior, alongside recorded responses and the shared-backbone control.

5 Conclusion

We presented MNCFR, an embodied-simulation-inspired diffusion framework that organizes listener facial reaction generation through explicit reaction formation and recent-context regulation. RFM transforms current multimodal speaker behavior into an intermediate listener-side reaction representation under listener-specific conditioning and constructs the listener facial reaction condition, while MAR uses recent speaker context to regulate its evolution across generation windows. Existing offline and online results show that the principal gains of MNCFR lie in listener facial reaction appropriateness. Component and temporal analyses further associate RFM with improved reaction formation and MAR with smoother within-window and cross-window motion. Perceptual evaluation indicates that the generated listener facial reactions are generally natural, contextually appropriate, and temporally coherent. These findings support the usefulness of separating current-window reaction formation from recent-context regulation within stochastic listener facial reaction generation. Remaining limitations include metric-specific trade-offs, reaction-intensity overshoot, and pose or appearance drift. Future work may improve robustness while preserving the explicit conditioning roles established by MNCFR.

REFERENCES

J. A. C. J. Bastiaansen, M. Thioux, and C. Keysers. Evidence for mirror systems in emotions. *Philosophical Transactions of the Royal Society B*, 364(1528):2391–2404, 2009. doi: 10.1098/rstb.2009.0058.

Janet B. Bavelas, Linda Coates, and Trudy Johnson. Listeners as co-narrators. *Journal of Personality and Social Psychology*, 79(6):941–952, 2000. doi: 10.1037/0022-3514.79.6.941.

Peter Blomsma, Gabriel Skantze, and Marc Swerts. Backchannel behavior influences the perceived personality of human and artificial communication partners. *Frontiers in Artificial Intelligence*, 5: 835298, 2022. doi: 10.3389/frai.2022.835298.

Ulf Dimberg, Monika Thunberg, and Kurt Elmehed. Unconscious facial reactions to emotional facial expressions. *Psychological Science*, 11(1):86–89, 2000. doi: 10.1111/1467-9280.00221.

Vittorio Gallese. Before and below 'theory of mind': Embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B*, 362(1480):659–669, 2007. doi: 10.1098/rstb.2006.2002.

Ursula Hess and Agneta Fischer. Emotional mimicry: Why and when we mimic emotions. *Social and Personality Psychology Compass*, 8(2):45–57, 2014. doi: 10.1111/spc3.12083.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. doi: 10.48550/arXiv.2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Ximi Hoque, Adamay Mann, Gulshan Sharma, and Abhinav Dhall. BEAMER: Behavioral encoder to generate multiple appropriate facial reactions. In *Proceedings of the ACM International Conference on Multimedia*, pp. 9536–9540, 2023. doi: 10.1145/3581783.3612860.

Jiaming Li, Sheng Wang, Xin Wang, Yitao Zhu, Honglin Xiong, Zixu Zhuang, and Qian Wang. ReactDiff: Latent diffusion for facial reaction generation. *Neural Networks*, pp. 107596, 2025. doi: 10.1016/j.neunet.2025.107596.

Cong Liang, Jiahe Wang, Haofan Zhang, Bing Tang, Junshan Huang, Shangfei Wang, and Xiaoping Chen. UniFaRN: Unified transformer for facial reaction generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9506–9510, 2023. doi: 10.1145/3581783.3612854.

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 *IEEE FG*, 2024. doi: 10.1109/FG59268.2024.10582013.

Cheng Luo, Siyang Song, Weicheng Xie, Micol Spitale, Zongyuan Ge, Linlin Shen, and Hatice Gunes. ReactFace: Online multiple appropriate facial reaction generation in dyadic interactions. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):6190–6207, 2025a. doi: 10.1109/TVCG.2024.3490613.

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, 2025b. doi: 10.1145/3746027.3755038.

R. M. Maatman, Jonathan Gratch, and Stacy Marsella. Natural behavior of a listening agent. In *Intelligent Virtual Agents*, pp. 25–36, 2005. doi: 10.1007/11550617_3.

Louis-Philippe Morency, Iwan de Kok, and Jonathan Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *Intelligent Virtual Agents*, pp. 176–190, 2008. doi: 10.1007/978-3-540-85483-8_18.

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*, pp. 20395–20405, 2022. doi: 10.1109/CVPR52688.2022.01975.

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 *IEEE FG*, 2024a. doi: 10.1109/FG59268.2024.10581901.

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 *IEEE FG*, 2024b. doi: 10.1109/FG59268.2024.10581978.

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 *Pattern Recognition*, pp. 233–248. Springer Nature Switzerland, 2024c. doi: 10.1007/978-3-031-78389-0_16.

Paula M. Niedenthal. Embodying emotion. *Science*, 316(5827):1002–1005, 2007. doi: 10.1126/science.1136930.

Patrizia Paggio and Costanza Navarretta. Head movements, facial expressions and feedback in conversations: Empirical evidence from danish multimodal data. *Journal on Multimodal User Interfaces*, 7(1):29–37, 2013. doi: 10.1007/s12193-012-0105-9.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Siyang Song, Micol Spitale, Cheng Luo, German Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth Andre, et al. REACT 2023: The first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9620–9624, 2023a. doi: 10.1145/3581783.3612832.

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. doi: 10.48550/arXiv.2302.06514.

Siyang Song, Micol Spitale, Cheng Luo, Cristina Palmero, German Barquero, Hengde Zhu, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, et al. REACT 2024: The second multiple appropriate facial reaction generation challenge. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition*, pp. 1–5. IEEE, 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, et al. REACT 2025: The third multiple appropriate facial reaction generation challenge. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13979–13984, 2025. doi: 10.1145/3746027.3762244.

Siyang Song, Micol Spitale, Zijian Wu, Xiangyu Kong, Cheng Luo, Cristina Palmero, German Barquero, Sergio Escalera, Michel Valstar, Mohamed Daoudi, Fabien Ringeval, Andrew Howes, Elisabeth André, and Hatice Gunes. REACT 2026: The fourth multiple appropriate facial reaction generation challenge: Personalised MAFRG and appropriate EEG reaction prediction. *arXiv preprint arXiv:2606.07935*, 2026. doi: 10.48550/arXiv.2606.07935.

Barbara Wild, Michael Erb, and Mathias Bartels. Are emotions contagious? evoked emotions while viewing emotionally expressive faces: Quality, quantity, time course and gender differences. *Psychiatry Research*, 102(2):109–124, 2001. doi: 10.1016/S0165-1781(01)00225-6.

Weicheng Xie, Chunlin Yan, Siyang Song, Zitong Yu, Linlin Shen, and Laizhong Cui. Smooth online multiple appropriate facial reaction generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5804–5813, 2025. doi: 10.1145/3746027.3755716.

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*, 17(3):3563–3578, 2026. doi: 10.1109/TAFFC.2026.3682774.

Jun Yu, Ji Zhao, Guochen Xie, Fengxin Chen, Ye Yu, Liang Peng, Minglei Li, and Zonghong Dai. Leveraging the latent diffusion models for offline facial multiple appropriate reactions generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9561–9565, 2023. doi: 10.1145/3581783.3612865.

Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, Tiejun Zhao, and Tao Mei. Responsive listening head generation: A benchmark dataset and baseline. In *European Conference on Computer Vision*, pp. 124–142, 2022. doi: 10.1007/978-3-031-19839-7_8.

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. doi: 10.1145/3664647.3680752.

Hengde Zhu, Xiangyu Kong, Weicheng Xie, Xin Huang, Xilin He, Lu Liu, Linlin Shen, Wei Zhang, Hatice Gunes, and Siyang Song. PerReactor: Offline personalised multiple appropriate facial reaction generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39(2), pp. 1665–1673, 2025. doi: 10.1609/aaai.v39i2.32159.