SEMANTIC-REACT: TEMPORALLY GROUNDED LINGUISTIC CONDITIONING FOR MULTIPLE APPROPRIATE FACIAL REACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In dyadic human-human interactions, linguistic semantics help human listeners interpret the speaker's verbal behaviour and determine when to express appropriate facial reactions (AFRs). Existing automatic Multiple Appropriate Facial Reaction Generation (MAFRG) methods typically generate AFRs conditioned on speaker audio-visual behaviour, where encoded speech representations primarily capture acoustic cues, without explicitly modelling word-level semantics or when those semantics are relevant to the listener's response. Consequently, the generated AFRs are usually appear plausible in motion while remaining semantically mismatched or temporally misaligned with the interaction context. To address this limitation, we propose Semantic-REACT, the first MAFRG solution that specifically encodes speaker's verbal transcripts and aligns the resulting linguistic features with the generated listener facial motion timeline. A gated adapter then injects these aligned linguistic features into the diffusion denoiser. This linguistic conditioning provides temporally aligned guidance while preserving the original speaker audio-visual conditions and the diversity of the generated listener AFRs. Experiments on the official REACT 2025 MARS benchmark show that Semantic-REACT improved the appropriateness of the generated AFRs compared to Trans-VAE, PerFRDiff, and REGNN while preserving decent diversity. Matched control experiments confirmed the importance of temporal alignment for linguistic conditioning. Qualitative analysis and human evaluations further indicate that the generated AFRss exhibit temporal coherence, semantic appropriateness, and favourable perceptual quality.

1 INTRODUCTION

In natural dyadic human–human interactions, linguistic semantics help listeners to interpretthe speaker's verbal behaviour and decide when to produce appropriate facial reactions Bavelas et al. (2000). As the interaction unfolds, listeners respond to what the speaker communicates, and linguistic cues shape not only when listeners' facial reactions should change but also how those reactions evolve over time Paggio & Navarretta (2013). Since the same speaker behaviour may prompt listeners to express multiple appropriate facial reactions (AFRs) that differ in facial movement patterns and response timing Scherer et al. (2001); Barrett et al. (2019), the Multiple Appropriate Facial Reaction Generation (MAFRG) task Song et al. (2023b;a) aims to develop artificial agents that can generate multiple AFRs for each perceived human speaker behaviour, thereby providing responsive facial feedback during human-agent interactions.

Existing MAFRG methods frequently build on dyadic interaction datasets and standardised benchmark protocols Song et al. (2024; 2025). Within these benchmark settings, existing methods leverage various generative AI methods, including variational autoencoders (VAEs) Hoque et al. (2023), latent diffusion models Yu et al. (2023), and vector-quantised diffusion models Nguyen et al. (2024), to generate AFRs conditioned on speaker behaviour. In addition to speaker audio-visual conditioning, linguistic information has been incorporated as an additional conditioning signal Ng et al. (2023); Lai et al. (2025). Beyond direct conditioning on transcripts, recent methods use free-form text and dialogue-level semantics to provide higher-level control over generated listener behaviour Liu et al. (2024); Li et al. (2026). These studies demonstrate the utility of linguistic information for listener

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.

behaviour generation. However, within audio-visual diffusion-based MAFRG, the role of temporal correspondence remains unclear, particularly whether the benefit of linguistic conditioning depends on aligning word-level features with the listener facial motion timeline.

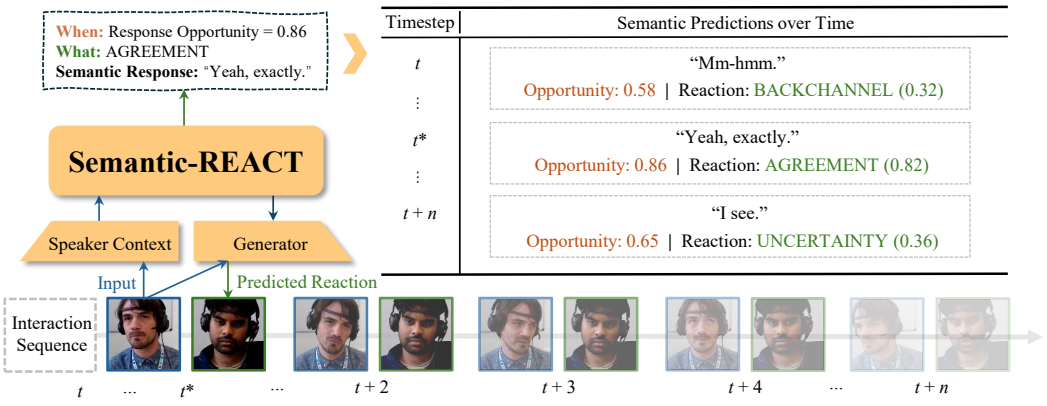

Figure 1: Overview of the Semantic-REACT framework. Given speaker audio-visual behaviour and time-stamped transcripts, the frame-aligned linguistic conditioning pathway provides temporally localised linguistic conditions for stochastic AFRs generation. A separate semantic reaction branch predicts time-indexed response opportunities and high-level reaction semantics to support interpretable analysis. The example illustrates the temporal correspondence among the speaker context, semantic predictions, and generated AFRs.

This unresolved issue raises the question of whether linguistic conditioning benefits audio-visual diffusion-based MAFRG only when word-level features preserve their temporal correspondence with the listener facial motion timeline. Linguistic representations are organised at the word level, whereas listener facial motion evolves continuously at the frame level. This difference in temporal granularity makes it difficult to determine when the speaker's linguistic content should influence the listener's facial reaction. As illustrated in Figure 1, different moments within the same interaction may provide different response opportunities. A linguistic cue may provide relevant guidance when aligned with its corresponding temporal position in the interaction, but may introduce inconsistent guidance when relocated to another part of the interaction. We therefore formulate linguistic conditioning for MAFRG as the problem of temporally grounding the speaker's linguistic content in the listener facial motion timeline.

To address this temporal grounding problem, we propose Semantic-REACT, which represents linguistic information as a temporally localised condition for MAFRG. Semantic-REACT uses Whisper-tiny to obtain time-stamped speaker transcripts Radford et al. (2023). A language encoder maps the transcribed words to linguistic embeddings. Word-level timestamps are then used to align the resulting linguistic embeddings with the listener facial motion timeline, producing a compact frame-aligned linguistic condition. A lightweight, zero-initialised adapter and a learnable gate incorporate the aligned linguistic features into the diffusion denoiser while preserving the denoiser's original speaker audio-visual conditioning. For interpretable analysis, a separate semantic reaction branch predicts time-indexed response opportunities and high-level reaction semantics. The original stochastic sampling process remains unchanged.

In summary, our main contributions are as follows:

- We formulate linguistic conditioning for MAFRG as a temporal grounding problem between word-level linguistic information from speaker transcripts and the listener facial motion timeline.
- We propose Semantic-REACT, which uses word-level timestamps to align linguistic features extracted by a frozen language encoder with the listener facial motion timeline. A lightweight, zero-initialised adapter and a learnable gate inject the aligned features into the diffusion denoiser while preserving the original speaker audio-visual conditioning.

- Matched control experiments across five independent runs show that linguistic conditioning improves FRC when temporally aligned, whereas temporal displacement removes this benefit, demonstrating the importance of temporal correspondence.

2 Related Work

2.1 Multiple Appropriate Facial Reaction Generation

Speech-driven facial animation maps a speaker's speech to the speaker's own facial motion Fan et al. (2022); Xing et al. (2023). Listener AFR generation instead models how another participant responds to the speaker's behaviour. Early listener models generated facial reactions from speaker audio-visual cues, while subsequent stochastic methods introduced partner-conditioned reaction sampling Zhou et al. (2022); Jonell et al. (2020); Ng et al. (2022). MAFRG extends this setting by generating multiple appropriate listener AFRs for the same speaker behaviour Song et al. (2023b). The REACT benchmarks evaluate generated listener AFRs in terms of appropriateness, diversity, realism, and synchrony with the corresponding speaker behaviour Song et al. (2023a; 2024; 2025).

Subsequent MAFRG research has addressed complementary aspects of listener reaction generation. Reaction-distribution and structured-representation methods model plausible AFR spaces, whereas diffusion-based approaches support stochastic reaction generation Xu et al. (2026); Hoque et al. (2023); Yu et al. (2023); Nguyen et al. (2024); Li et al. (2025). Online and sequential methods capture the temporal evolution of listener behaviour Luo et al. (2025); Nguyen et al. (2025). Personalised methods further adapt generated AFRs to individual listener characteristics Zhu et al. (2024; 2025). Despite these advances, most MAFRG methods condition generation primarily on speaker audio-visual behaviour and infer reaction semantics and timing implicitly from these inputs.

2.2 Language- and Semantics-Conditioned Listener Modelling

Linguistic conditioning introduces interaction content beyond speaker audio-visual behaviour. Ng et al. autoregressively generate listener motion from time-stamped speaker transcripts, showing that lexical semantics and word timing can inform listener responses Ng et al. (2023). LLM-based approaches further combine linguistic content with acoustic and emotional cues for multimodal listener generation Lai et al. (2025). Text-guided methods use free-form descriptions, listener attributes, semantic labels, and emotion-intensity controls to produce expressive and controllable listener behaviour Liu et al. (2024); Li et al. (2026).

These approaches differ in both purpose and temporal scope. Time-stamped transcripts provide local lexical cues, whereas utterance- and description-level controls specify broader behavioural characteristics. Existing studies therefore establish that linguistic information can guide listener behaviour. However, the role of temporal correspondence remains unclear within audio-visual diffusion-based MAFRG. In particular, it is unclear whether linguistic conditioning remains beneficial when word-level features are displaced from their corresponding positions on the listener facial motion timeline. Semantic-REACT examines this question through frame-aligned linguistic conditioning and a matched temporal-displacement control.

2.3 Dialogue Feedback and Temporal Response Modelling

Listener feedback depends on both its interactional function and its temporal placement. Backchannels allow listeners to participate while the speaker retains the conversational floor, and their timing can influence how the interaction develops Yngve (1970); Bavelas et al. (2000); Tolins & Fox Tree (2014). Head movements, facial expressions, and turn transitions occur at locally relevant conversational moments Paggio & Navarretta (2013); Sacks et al. (1974). Their realisation and latency may also vary across languages and cultures Stivers et al. (2009). These findings indicate that the appropriateness of a listener AFR depends not only on its facial motion pattern but also on its position within the local dialogue structure.

Computational dialogue models represent response timing through transition opportunities and the compatibility of prospective responses with those opportunities Ekstedt & Skantze (2020); Jiang

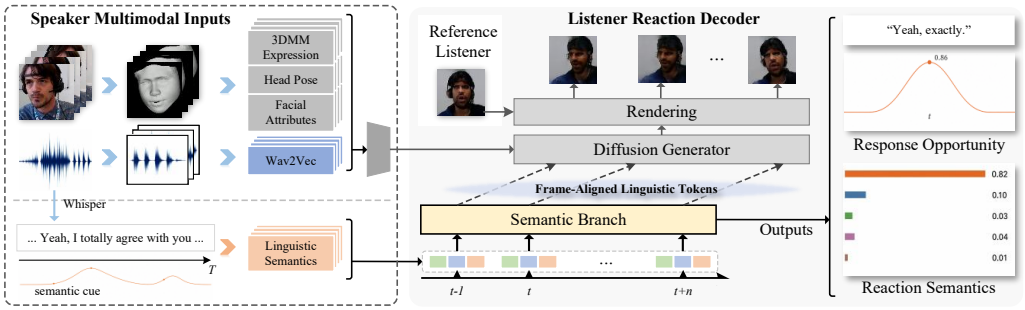

Figure 2: Overview of Semantic-REACT. Time-stamped linguistic features are aligned with the listener facial motion timeline and incorporated into the original speaker audio-visual conditions to guide diffusion-based listener facial reaction generation. The semantic analysis branch identifies response opportunities and high-level reaction semantics, while stochastic sampling produces multiple AFRs candidates.

et al. (2023). Conversational-agent studies further examine feedback as interactive behaviour and its effects on perceived social attributes Axelsson et al. (2022); Blomsma et al. (2022). These formulations primarily describe discrete dialogue events, whereas MAFRG generates continuous listener facial motion. Connecting local response opportunities with the listener facial motion timeline is therefore necessary for analysing the temporal appropriateness of generated AFRs.

Taken together, prior MAFRG research provides audio-visual generative frameworks, language-conditioned methods demonstrate the value of semantic information, and dialogue studies establish the importance of response timing. What remains insufficiently understood is whether linguistic conditioning benefits audio-visual diffusion-based MAFRG because its word-level features preserve their temporal correspondence with listener facial motion.

3 METHOD

Given speaker audio-visual behaviour $X_{1:T} = \{x_t\}_{t=1}^{T}$ and the corresponding time-stamped transcript $\mathcal{D}$, Semantic-REACT generates K AFRs candidates $\{\hat{Y}_{1:T}^{(k)}\}_{k=1}^{K}$. Each candidate is a sequence of listener facial motion features defined over T frames.

From $\mathcal{D}$, Semantic-REACT constructs a frame-aligned linguistic condition $L_{1:T} = \{l_t\}_{t=1}^{T}$, where l_t represents the linguistic information assigned to listener facial motion frame t. Conditioned on the speaker audio-visual behaviour and the frame-aligned linguistic condition, each listener facial reaction candidate is sampled as

$$\hat{Y}_{1:T}^{(k)} \sim p_\theta \left(Y_{1:T} \mid X_{1:T}, L_{1:T} \right), \qquad k = 1, \ldots, K, \tag{1}$$

where p_θ denotes the conditional distribution learned by the listener facial reaction generator and K denotes the number of sampled reaction candidates.

As illustrated in Figure 2, Semantic-REACT augments the underlying audio-visual diffusion generator with two components: frame-aligned linguistic condition construction and gated linguistic conditioning. The first component uses word-level timestamps to assign linguistic features to their corresponding positions on the listener facial motion timeline, producing a frame-aligned linguistic condition. The second component uses a zero-initialised adapter and a learnable gate to transform the aligned linguistic features into linguistic conditioning tokens. These tokens are appended to the original speaker audio-visual conditioning sequence and jointly serve as the cross-attention memory of the diffusion denoiser. Conditioned on the original speaker audio-visual tokens and the additional linguistic tokens, the diffusion denoiser generates multiple AFRs candidates from independently sampled initial noise.

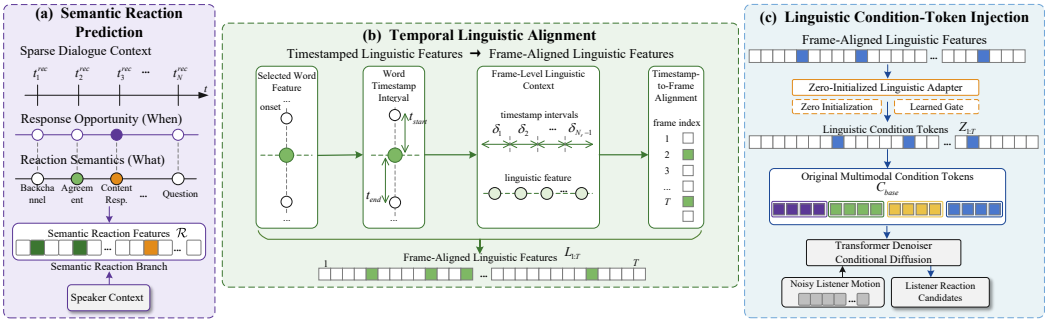

Figure 3: Detailed components of Semantic-REACT. (a) Semantic reaction prediction identifies time-indexed response opportunities and reaction semantics. (b) Temporal linguistic alignment maps time-stamped word features to the listener facial motion timeline. (c) The adapted and gated linguistic tokens are incorporated into the shared cross-attention memory of the diffusion denoiser.

3.1 Frame-Aligned Linguistic Condition Construction

The first component constructs a linguistic condition whose temporal resolution matches that of the listener facial motion sequence. This construction first extracts word-level linguistic features and their temporal intervals from the speaker audio and then assigns each feature to its corresponding listener facial motion frames.

We use `openai/whisper-tiny` to obtain transcribed words and their temporal intervals Radford et al. (2023). A frozen language encoder maps each word to a 768-dimensional embedding, which a fixed random projection reduces to a 64-dimensional word-level linguistic feature. Each projected feature is assigned to listener facial motion frames whose timestamps fall within the corresponding word interval; frames outside all word intervals receive a zero feature:

$$l_t = \begin{cases} \tilde{e}_{j(t)}, & j(t) \neq \varnothing, \\ \mathbf{0}, & j(t) = \varnothing, \end{cases} \qquad l_t \in \mathbb{R}^{64}. \tag{2}$$

Collecting these features gives $L_{1:T} = \{l_t\}_{t=1}^{T} \in \mathbb{R}^{T\times 64}$, preserving each word-level feature at its original position on the listener facial motion timeline. The transcription model, language encoder, and projection remain fixed during training. Full feature-construction details are provided in the supplementary material.

3.2 Gated Linguistic Conditioning for MAFRG

The frame-aligned linguistic condition $L_{1:T} = \{l_t\}_{t=1}^{T}$ must be adapted to the hidden representation space of the pretrained diffusion denoiser before being used for conditioning. We therefore introduce a lightweight linguistic adapter followed by a learnable gate to transform the aligned linguistic features and regulate their contribution to the denoiser.

As shown in Figure 3(c), the linguistic adapter A_ω maps each frame-level condition l_t to an adapted representation u_t in the denoiser's hidden space:

$$u_t = A_\omega(l_t). \tag{3}$$

The adapter sequentially applies non-affine layer normalisation, a linear projection, a second layer-normalisation layer, a GELU activation, and a final linear projection. Its final projection is zero-initialised, so training begins from the original audio-visual generation behaviour before gradually learning a compatible linguistic representation.

To regulate the contribution of the adapted linguistic representation at each listener facial motion frame, we apply a learnable element-wise gate:

$$z_t = \lambda \left[\sigma\left(G_\gamma(u_t)\right) \odot u_t\right], \tag{4}$$

where G_γ denotes the gating network, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. The sigmoid gate regulates each feature dimension, and the fixed coeffi-

cient $\lambda = 0.10$ globally scales the resulting frame-level token z_t. Collecting these tokens gives $Z_{1:T} = \{z_t\}_{t=1}^{T}$.

Let C_{base} denote the original conditioning-token sequence of the underlying audio-visual diffusion generator. This sequence contains the diffusion-timestep token together with the original speaker audio and visual conditioning tokens. Semantic-REACT appends $Z_{1:T}$ to C_{base}:

$$C_{\text{all}} = [C_{\text{base}}; Z_{1:T}], \tag{5}$$

where the semicolon denotes concatenation along the token dimension.

The combined sequence serves as the Transformer denoiser's shared cross-attention memory, retaining its original tokens while adding temporally ordered linguistic conditioning.

3.3 Training and Sampling Strategies

Semantic-REACT retains the diffusion formulation of the underlying AFRs generator and trains the proposed linguistic conditioning components through its denoising objective Ho et al. (2020). Given a noisy listener facial reaction sequence Y_τ, diffusion step τ, and combined condition C_{all}, the Transformer denoiser predicts the clean sequence $\hat{Y}_0 = D_\theta(Y_\tau, \tau, C_{\text{all}})$.

Following the output representation of the underlying generator, both Y_0 and $\hat{Y}_0$ contain action unit (AU), valence–arousal (VA), and emotion (EM) feature groups. We train the conditioned denoiser using a weighted reconstruction loss over the three feature groups:

$$\mathcal{L}_{\text{train}} = \sum_{q \in \{\text{AU,VA,EM}\}} w_q \left\| \hat{Y}_0^q - Y_0^q \right\|_2^2, \tag{6}$$

where Y_0^q and $\hat{Y}_0^q$ denote the ground-truth and predicted features in group q, respectively, and w_q denotes the corresponding loss weight. This objective provides the optimisation signal for the linguistic adapter, the learnable gate, and the trainable cross-attention modules that incorporate the frame-aligned linguistic condition.

During training, we optimise the linguistic adapter A_ω, the gating network G_γ, and the cross-attention modules in the final two Transformer decoder layers. The transcription model, language encoder, projection matrix, and all remaining parameters of the underlying audio-visual diffusion generator are kept fixed. This selective optimisation adapts the interface between the frame-aligned linguistic condition and the pretrained denoiser while restricting changes to the newly introduced modules and the selected cross-attention layers.

At inference, Semantic-REACT uses DDIM sampling Song et al. (2021). The audio-visual behaviour and frame-aligned linguistic condition are shared across candidates, while K independently sampled initial noise sequences produce varied listener facial reactions under the same conditions. Full diffusion and sampling definitions are provided in the supplementary material.

4 Experiments

We evaluate benchmark performance, the contribution of linguistic conditioning, and the dependence of generation performance on temporal correspondence between linguistic information and listener facial motion. We also assess the semantic and perceptual quality of the generated listener facial reactions.

4.1 Experimental Setup

Dataset and protocol. We evaluate Semantic-REACT on the MARS benchmark under the official REACT 2025 offline protocol Song et al. (2025). We use the official data split, preprocessing pipeline, listener facial reaction representation, and evaluation scripts. All sequences are sampled at 30 fps. For each speaker sequence, we generate $K = 10$ listener facial reaction candidates using a 50-step DDIM sampling process. Published baseline results are taken from studies evaluated under the same official setting.

Evaluation metrics. Following the official protocol Song et al. (2025), we report Facial Reaction Consistency (FRC, ↑) and Facial Reaction Distance (FRD, ↓). We also report Temporal Lagged Cross-Correlation (TLCC, ↓), S-MSE (↓), and Facial Reaction Variation (FRVar, ↑). FRC serves as the primary metric for reaction appropriateness. FRD, TLCC, S-MSE, and FRVar characterise distributional distance, temporal behaviour, reconstruction discrepancy, and reaction variation, respectively.

Implementation details. Semantic-REACT augments the underlying audio-visual diffusion generator with 64-dimensional frame-aligned linguistic features extracted from speaker transcripts. During training, we update the zero-initialised conditioning adapter, the gating module, and the cross-attention modules in the final two Transformer decoder layers. The linguistic injection scale is set to $\lambda = 0.10$.

Controlled evaluation. We report the mean and sample standard deviation across five independently trained runs. We evaluate the contribution of the linguistic conditioning pathway using a seed-matched *w/o Linguistic Conditioning* control. We assess dependence on temporal correspondence using a checkpoint-matched *Shifted Linguistic Condition* control. This control shifts the linguistic sequence at inference while keeping the feature values, model parameters, and evaluation settings fixed. For each control, we report the paired difference, computed as Semantic-REACT minus the control, together with the corresponding 95% confidence interval. This controlled design distinguishes the contribution of the linguistic conditioning pathway from the effect of its temporal placement. Further implementation details and statistical procedures are provided in the supplementary material.

4.2 Comparison with Existing Methods

Table 1: Comparison with existing methods under the official REACT 2025 offline protocol on MARS. Published baseline results are taken from the official benchmark report Song et al. (2025). Results for Semantic-REACT are reported as the mean ± sample standard deviation across five independent runs.

Method	FRC ↑	FRD ↓	TLCC ↓	S-MSE ↓	FRVar ↑
Trans-VAE	0.3000	181.7200	49.0000	0.0079	0.0083
PerFRDiff	0.5800	217.3200	48.6600	0.1386	**0.0862**
REGNN	0.5400	**155.4900**	**44.8100**	**0.0010**	0.0048
Semantic-REACT	**0.7058 ± 0.0477**	173.9510 ± 8.5359	48.2563 ± 0.0484	0.1379 ± 0.0067	0.0775 ± 0.0044

Table 1 shows that Semantic-REACT achieves the highest FRC (0.7058 ± 0.0477); all five runs exceed the strongest published baseline FRC of 0.5800. It also improves FRD and TLCC over Trans-VAE and PerFRDiff, while REGNN remains better on both. The remaining metrics show complementary trade-offs: Semantic-REACT slightly improves S-MSE over PerFRDiff and substantially improves FRVar over Trans-VAE and REGNN, whereas PerFRDiff has the highest FRVar.

4.3 Contribution of Linguistic Conditioning

The seed-matched ablation evaluates the frame-aligned linguistic pathway. Removing it reduces FRC from 0.7058 ± 0.0477 to 0.6705 ± 0.0431, corresponding to a paired improvement of $+0.0353$ with a 95% confidence interval of $[0.0288, 0.0418]$.

Table 2: Matched evaluation of linguistic conditioning and temporal correspondence. Values are reported as the mean ± sample standard deviation across five paired runs. The difference rows report the paired difference between Semantic-REACT and the corresponding control, followed by the 95% confidence interval.

Configuration	FRC ↑	FRD ↓	TLCC ↓	S-MSE ↓	FRVar ↑
Semantic-REACT	0.7058 ± 0.0477	173.9510 ± 8.5359	48.2563 ± 0.0484	0.1379 ± 0.0067	0.0775 ± 0.0044
w/o Linguistic Conditioning	0.6705 ± 0.0431	185.2540 ± 5.6601	48.3820 ± 0.0507	0.1495 ± 0.0031	0.0697 ± 0.0036
Full − w/o Linguistic	$+0.0353\ [0.0288, 0.0418]$	$-11.3030\ [-14.9035, -7.7026]$	$-0.1257\ [-0.1724, -0.0789]$	$-0.0116\ [-0.0172, -0.0060]$	$+0.0078\ [0.0055, 0.0101]$
Shifted Linguistic Condition	0.6594 ± 0.0420	188.0160 ± 5.8360	48.5160 ± 0.0594	0.1564 ± 0.0029	0.0733 ± 0.0032
Full − Shifted Linguistic	$+0.0464\ [0.0384, 0.0544]$	$-14.0650\ [-17.6085, -10.5215]$	$-0.2597\ [-0.3235, -0.1958]$	$-0.0185\ [-0.0244, -0.0127]$	$+0.0043\ [0.0020, 0.0066]$

Relative to *w/o Linguistic Conditioning*, Semantic-REACT reduces FRD, TLCC, and S-MSE while increasing FRVar; all paired 95% confidence intervals exclude zero.

4.4 Dependence on Temporal Correspondence

The checkpoint-matched control shifts the linguistic sequence during inference while holding its values, model parameters, audio-visual conditions, and sampling settings fixed. Alignment improves FRC by 0.0464 over *Shifted Linguistic Condition* (95% CI $[0.0384, 0.0544]$) and consistently improves all five metrics. Mean FRC decreases from Semantic-REACT (0.7058) to *w/o Linguistic Conditioning* (0.6705) and then to the shifted control (0.6594), indicating that temporally misplaced cues can introduce inconsistent guidance. Online response-timing prediction remains outside this offline protocol. Per-run results and further statistical details are provided in the supplementary material.

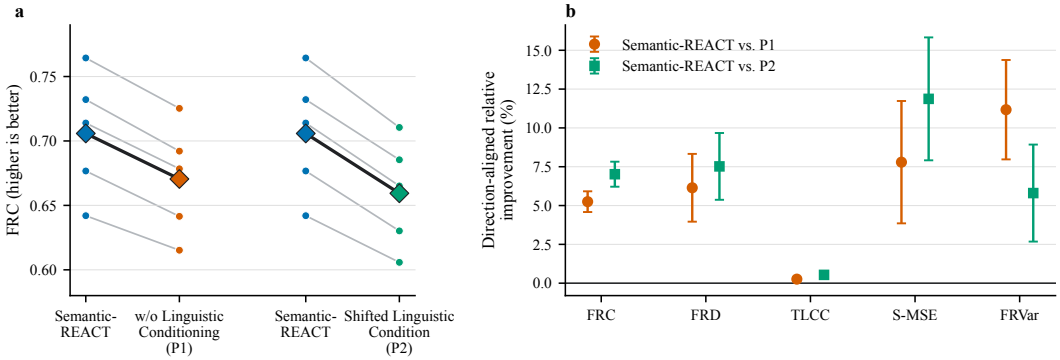

Figure 4: Matched evaluation of linguistic conditioning and temporal correspondence. (a) Paired FRC results across five runs. (b) Direction-aligned paired differences across all metrics, with positive values indicating an advantage for Semantic-REACT.

4.5 Semantic and Perceptual Evaluation

Figure 5 presents agreement, backchannel, content-response, question, and uncertainty cases, combining speaker context, generated facial motion, reaction semantics, and response-opportunity curves. The cases illustrate correspondence between time-indexed semantic records and continuous listener facial motion.

We further conduct a two-track user study: Track 1 compares generated and recorded reactions using 375 two-alternative forced-choice trials, while Track 2 asks 60 participants to rate 15 clips on four 10-point perceptual dimensions. Generated reactions are preferred in 42.7% of Track 1 comparisons. As shown in Table 3, Semantic-REACT scores 7.45–7.52 across Track 2 dimensions, with consistently favourable but lower ratings than recorded reactions.

The supplementary material provides representative failure cases from the semantic reaction branch, the complete user study protocols, and the questionnaire definitions.

5 Conclusion

Semantic-REACT demonstrates that the effectiveness of linguistic conditioning in Multiple Appropriate Facial Reaction Generation depends on its temporal correspondence with listener facial motion. The framework aligns time-stamped linguistic features from speaker transcripts with the listener facial motion timeline. A lightweight, zero-initialised gated conditioning pathway incorporates these features into a diffusion-based listener facial reaction generator. The original speaker audio-visual conditions are preserved in the denoiser, while independently sampled diffusion noise yields multiple reaction realisations under the same interaction context.

Under the official REACT 2025 offline protocol on MARS, Semantic-REACT achieves the highest FRC among the published baselines included in the comparison. Matched control experiments

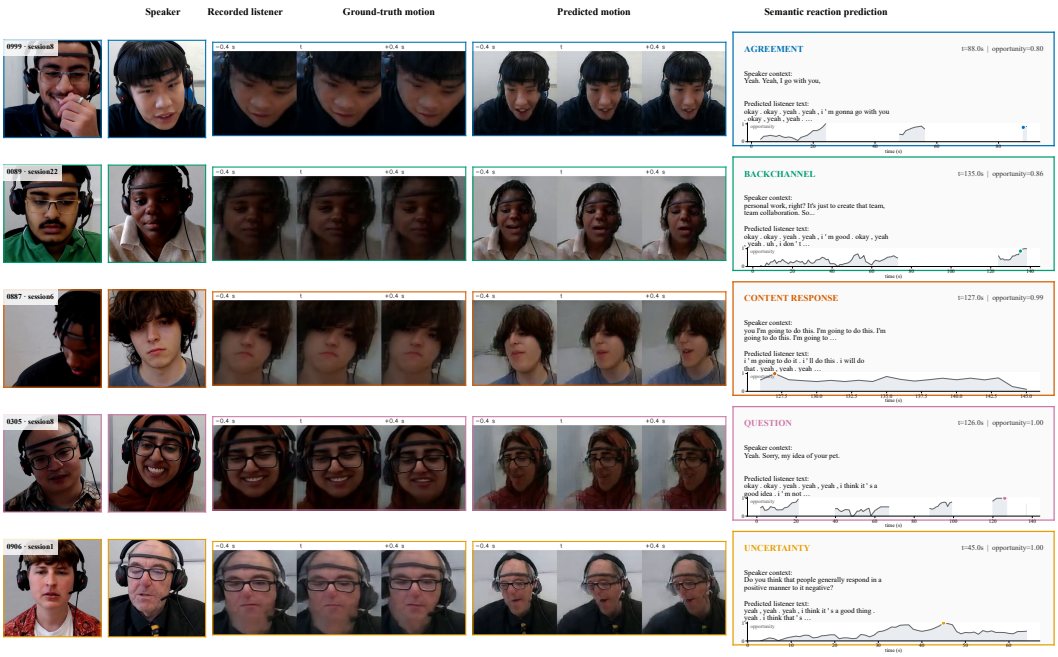

Figure 5: Qualitative analysis of semantic reaction predictions and generated listener facial reactions. Each row presents the speaker context, the recorded listener, and the recorded and generated listener facial motion. The corresponding semantic reaction predictions include the reaction category and response-opportunity curve.

Table 3: Track 2 user study results based on a 10-point rating scale. Values are reported as the mean $\pm$ standard deviation.

Method	Naturalness	Context. App.	Temporal Coh.	Semantic App.
Recorded Listener	8.81 ± 1.11	8.81 ± 1.10	8.73 ± 1.14	8.73 ± 1.14
Semantic-REACT	7.45 ± 1.32	7.47 ± 1.32	7.52 ± 1.35	7.45 ± 1.32

across five runs reveal a consistent ordering in mean FRC. Frame-aligned linguistic conditioning performs best, removing the linguistic conditioning pathway reduces FRC, and temporal shifting yields the lowest mean FRC. This ordering indicates that the effectiveness of linguistic conditioning depends on its temporal placement on the listener facial motion timeline. Temporally aligned linguistic information narrows the set of contextually appropriate reactions at each moment, while stochastic diffusion preserves multiple facial motion realisations.

The current evaluation is limited to offline linguistic conditioning and temporal perturbation. Extending the framework to streaming dialogue will require online linguistic alignment, prospective prediction of response opportunities, and dialogue-level semantic modelling for real-time listener facial reaction generation.

REFERENCES

Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. Modeling feedback in interaction with conversational agents—a review. *Frontiers in Computer Science*, 4:744574, 2022. doi: 10.3389/fcomp.2022.744574.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68, 2019. doi: 10.1177/1529100619832930.

Janet B Bavelas, Linda Coates, and Trudy Johnson. Listeners as co-narrators. *Journal of personality and social psychology*, 79(6):941, 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.

Erik Ekstedt and Gabriel Skantze. Turngpt: A transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. doi: 10.18653/v1/2020.findings-emnlp.268.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18770–18780, 2022. doi: 10.1109/CVPR52688.2022.01821.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

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. ACM, 2023. doi: 10.1145/3581783.3612860.

Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. Response-conditioned turn-taking prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12241–12248. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.776. URL https://aclanthology.org/2023.findings-acl.776/.

Patrik Jonell, Taras Kucherenko, Gustav Eje Henter, and Jonas Beskow. Let's face it: Probabilistic multi-modal interlocutor-aware generation of facial gestures in dyadic settings. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pp. 1–8. ACM, 2020. doi: 10.1145/3383652.3423911.

Peiwen Lai, Weizhi Zhong, Yipeng Qin, Xiaohang Ren, Baoyuan Wang, and Guanbin Li. LLM-driven multimodal and multi-identity listening head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10656–10666, 2025. doi: 10.1109/CVPR52734.2025.00996.

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. doi: 10.1016/j.neunet.2025.107596.

Shiying Li, Xingqun Qi, Bingkun Yang, Weile Chen, Zezhao Tian, Muyi Sun, Qifeng Liu, Man Zhang, and Zhenan Sun. Vividlistener: Expressive and controllable listener dynamics modeling for multi-modal responsive interaction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(8):6396–6404, 2026. doi: 10.1609/aaai.v40i8.37567.

Xi Liu, Ying Guo, Cheng Zhen, Tong Li, Yingying Ao, and Pengfei Yan. Customlistener: Text-guided responsive interaction for user-friendly listening head generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2415–2424, 2024. doi: 10.1109/CVPR52733.2024.00234. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_CustomListener_Text-guided_Responsive_Interaction_for_User-friendly_Listening_Head_Generation_CVPR_2024_paper.html.

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, 2025. doi: 10.1109/TVCG.2024.3490613.

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.

Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10083–10093, 2023. doi: 10.1109/ICCV51070.2023.00925.

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, 2024. 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, 2025. doi: 10.1007/978-3-031-78389-0_16.

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.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974. URL https://www.jstor.org/stable/412243.

Klaus R. Scherer, Angela Schorr, and Tom Johnstone (eds.). *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, New York, 2001. ISBN 9780195130072. doi: 10.1093/oso/9780195130072.001.0001.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Siyang Song, Micol Spitale, Cheng Luo, Germán Barquero, Cristina Palmero, Sergio Escalera, Michel Valstar, Tobias Baur, Fabien Ringeval, Elisabeth André, and Hatice Gunes. React2023: The first multiple appropriate facial reaction generation challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9620–9624. ACM, 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. URL https://arxiv.org/abs/2302.06514.

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. 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. ACM, 2025. doi: 10.1145/3746027.3762244.

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009. doi: 10.1073/pnas.0903616106.

Jackson Tolins and Jean E. Fox Tree. Addressee backchannels steer narrative development. *Journal of Pragmatics*, 70:152–164, 2014. doi: 10.1016/j.pragma.2014.06.006.

Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12780–12790, 2023. doi: 10.1109/CVPR52729.2023.01229.

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.

Victor H. Yngve. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society*, pp. 567–578, 1970.

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. ACM, 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. Springer, 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. ACM, 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1665–1673, 2025. doi: 10.1609/aaai.v39i2.32159.