BEYOND END-TO-END BLACK BOX MAPPING: AN INTENTIONAL AGENT FRAMEWORK FOR COGNITIVE-DRIVEN FACIAL REACTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automatic human-like facial reaction generation (FRG) is essential for building intelligent systems that can authentically and deeply engage in human-computer interaction (HCI). While diverse and context-appropriate facial reactions can reflect latent appraisal and affective processes in human interaction, most existing FRG methods rely on end-to-end architectures that directly map speaker behaviours to listener expressions without an explicit intermediate internal state. In this paper, we reformulate FRG as generation mediated by a structured internal-state process and propose the **Intentional Agent**, which shifts FRG from direct stimulus-response mapping to stimulus-grounded generation through explicit intermediate states. To represent temporal internal-state evolution for FRG, we propose an internal dynamics model that integrates emotional drives with an iterative Inner Thought Flow (ITF) within a structured intermediate state used for subsequent generation. This state can continue to update during conversational silences. Furthermore, to bridge abstract internal states with physiological actions, we explicitly formulate FRG as a downstream affective mapping from this latent thought flow to physical facial expressions. Experiments on the REACT 2025 dataset show an FRDist of 72.39 and an FRDiv of 0.5057; perceptual plausibility is evaluated separately through blinded human ratings. A blinded human evaluation of 96 reactions found no significant difference in mean score between Full and ground truth (5.527 vs. 5.195, $p_{\mathrm{Holm}} = .076$), while Full significantly outperformed Event-Triggered and Heuristic-Only (both $p_{\mathrm{Holm}} < .001$). The Reaction Quality Scorer (RQS) correlated strongly with human judgements (Pearson $r = .855$; Spearman $\rho = .821$, both $p < .05$), supporting its use as a supplementary automatic metric. These results underscore the immense potential of endogenous dynamics in building highly autonomous, human-like agents.

1 INTRODUCTION

Facial Reaction Generation (FRG) is a crucial component for endowing virtual avatars with anthropomorphic social interaction capabilities (Song et al., 2023b; Luo et al., 2024; Song et al., 2023c). In real dyadic human-human interactions, a listener's non-verbal feedback (e.g., subtle changes in facial expressions) plays an irreplaceable role in maintaining conversational rhythm, expressing empathy, and conveying interactive intentions (Stacchio et al., 2025; Ng et al., 2023; Atzil-Slonim et al., 2023). Therefore, how to automatically and adaptively generate appropriate, diverse, and cognitively-driven human-like facial reactions based on the human users' multi-modal behaviours has become a core research challenge in building highly realistic social virtual humans in recent years (Song et al., 2023b; 2025; 2023a).

To address this challenge, extensive research has been dedicated to synthesising cognitively plausible and diverse facial reactions. Existing methods typically treat FRG as a cross-modal sequence translation problem (Song et al., 2023b; 2024; 2025), with the underlying paradigm undergoing a rapid evolution. Transitioning from early deterministic one-to-one regression (Cassell & Thorisson, 1999; Morency et al., 2008), recent Multiple Appropriate Facial Reaction Generation (MAFRG) research has entirely shifted towards stochastic distribution learning to tackle the inherent "one-to-many" mapping challenge. To explicitly model this non-deterministic nature, researchers initially

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.

introduced unified Transformer architectures (Liang et al., 2023), contrastive learning pairs (Hoque et al., 2023), and online synchronous mapping frameworks (Luo et al., 2024). Subsequently, the integration of discrete latent variables (Liu et al., 2024) and Gaussian Mixture Models (GMMs) within VAE frameworks (Nguyen et al., 2024a) allowed for a better capture of the probability space of valid listener reactions. Concurrently, to further push the boundaries of fine-grained diversity and temporal consistency, Latent Diffusion Models have recently dominated MAFRG research (Yu J & Z, 2023; Nguyen et al., 2024b; Nguyen & Kim, 2024). These diffusion-based architectures are continuously being refined to enhance context-awareness through techniques such as multi-scale frequency scattering transforms (Mao et al., 2025), hierarchical 2D-3D multimodal decoupling (Lv et al., 2025), and motion-aware visual dynamic fusions (Huang & Yu, 2025). Most recently, recognizing identical stimuli elicit different responses depending on the listener, the frontier of MAFRG has advanced towards accounting for individual differences. Works such as learning appropriate reaction distributions via Graph Neural Networks (Xu et al., 2026), editing generic network weights to mimic personalized cognitive styles (Zhu et al., 2024), and conditioning diffusion generation on explicit Big Five personality traits (Wang et al., 2025) mark this new trajectory to guide generation process.

However, whether relying on deterministic rules or probabilistic distribution learning, these methods ignore modelling the internal cognitive states of models, establishing only direct "external stimulus to facial reaction" black-box mappings (Marsella & Gratch, 2009; Cassell & Tartaro, 2007). In real-world interactions, high-quality listener facial reactions are not merely passive, low-level reflexes to the speaker's audio-visual cues. Fundamentally, they are the embodied physical manifestations of the listener's latent cognitive processes and social intentionality (Scherer et al., 2001; Scherer, 2005). A natural facial expression, from a human perspective, is a downstream product that emerges after understanding the interactive context, forming an internal stance, and experiencing emotional fluctuations (Scherer et al., 2001; Russell & Mehrabian, 1977; Wagner et al., 2024).

Furthermore, even the latest MAFRG methods attempting to integrate personality traits, personalized cognitive styles, or implicit cognitive processors (Wang et al., 2025; Zhu et al., 2024; Xu et al., 2026) predominantly treat these elements as static conditional priors or fixed network weights. Without modelling the temporal evolution of the listener's inner state, the generation process remains fundamentally stimulus-driven. By treating generation merely as a conditioned passive echo to the speaker's unilateral input, these paradigms do not explicitly represent how an intermediate listener state evolves over time. An appropriate and explanable social facial reaction requires the listener to proactively synthesize the interlocutor's verbal expressions and multimodal cues, construct a coherent train of thought internally, and, based on the evolution of this inner thought, spontaneously project corresponding expressive feedback to the outside world (Premack & Woodruff, 1978; Chanes et al., 2018; Dimberg et al., 2000). Therefore, to achieve human-like FRG, one important modelling challenge is how to represent the intermediate process from context interpretation and internal-state evolution to affective facial generation. This motivates us to formulate FRG with an explicit internal-state process and to use an agent architecture that can maintain and update such intermediate information across time.

In this paper, we propose the novel 'Intentional Agent' MAFRG approach, shifting facial reaction generation from direct reactive mapping to generation mediated by an explicit internal state. We introduce a self-sustaining internal dynamics model and leverage a Large Language Model (LLM) (Park et al., 2023; Xi et al., 2025) to instantiate an Inner Thought Flow (ITF) within a structured intermediate state. At each update, the thought generator uses the current structured state to produce the next thought, after which the action generator produces a possible action state. These outputs subsequently enter the affective and facial generation process. Our main contributions are:

- We reformulate FRG as an internal-state-mediated generation problem, explicitly connecting structured internal-state evolution to physical facial generation rather than direct stimulus–reaction mapping.
- We propose the Intentional Agent, whose structured state maintains memory, affect, thought, and expectation across updates, enabling state evolution and action generation during conversational silence without new external triggers.
- Experiments on REACT 2025 show the lowest FRDist and highest FRDiv and FRVar among the compared learned methods. In blinded human evaluation, Full did not differ significantly

from GT in mean score ($p_{\text{Holm}} = .076$) and significantly outperformed Event-Triggered and Heuristic-Only (both $p_{\text{Holm}} < .001$).

2 Related Work

Facial Reaction Generation. Early FRG methods primarily formulated listener behaviour as deterministic mappings from speaker cues (Cassell & Thorisson, 1999; Morency et al., 2008). Subsequent work drew on general image synthesis and Talking Head generation to improve visual fidelity (Karras et al., 2019; Thies et al., 2016; Vougioukas et al., 2020; Zhang et al., 2023), while Multiple Appropriate Facial Reaction Generation (MAFRG) addressed the one-to-many nature of listener responses (Song et al., 2023c; Xu et al., 2026). Existing MAFRG methods include Transformer and contrastive-learning approaches (Liang et al., 2023; Hoque et al., 2023), latent-variable and VAE-based models (Luo et al., 2024; Liu et al., 2024; Nguyen et al., 2024a), and diffusion-based generation (Yu J & Z, 2023; Nguyen et al., 2024b; Nguyen & Kim, 2024). Recent methods further improve temporal and contextual modelling through multi-scale features and multimodal decoupling (Mao et al., 2025; Lv et al., 2025; Huang & Yu, 2025).

Listener-specific variation has also been modelled through weight editing, reaction-distribution learning, and explicit personality conditioning (Zhu et al., 2024; Xu et al., 2026; Wang et al., 2025). However, these methods remain primarily conditioned on external interaction signals or relatively static listener-specific representations, without explicitly maintaining a temporally evolving intermediate state through which memory, affect, thought, and expectation jointly influence later reactions. Our work instead models this internal-state process explicitly as the mediator between interaction context and facial generation.

Proactive and Cognitive Agents. LLM-based agents extend passive instruction following towards proactive interaction through perception, planning, memory, and internal reasoning (Deng et al., 2025; Luo et al., 2025; Lu et al., 2024). Embodied agents further connect language models with environmental perception and action (Wang et al., 2023; Huang et al., 2022), with proactive mechanisms applied to visual analytics (Zhao et al., 2025b) and programming assistance (Zhao et al., 2025a). In social interaction, inner-thought mechanisms can determine intervention timing (Liu et al., 2025), while personalised modelling supports longer-term user adaptation (Qiu et al., 2025; Garbacea & Tan, 2025).

These systems show that internal reasoning can regulate downstream behaviour, but reasoning or action is generally initiated by tasks, observations, or other external triggers. Our Intentional Agent instead maintains endogenous state evolution during continuous non-verbal interaction, allowing retained memory, affect, thought, and expectation-related information to regulate subsequent updates even when no new external event arrives.

3 Methodology

To overcome the limitations of the existing FRG paradigm, we propose an Intentional Agent architecture. Unlike end-to-end black-box models that directly map external inputs to facial pixels, our approach shifts the focus from passive cross-modal alignment to the simulation of an agent's internal cognitive flow and endogenous drives. This methodology is grounded in the concept of Operative Intentionality, where facial expressions are treated as the embodied extension of a self-sustaining internal state rather than mere physical reflections. For a rigorous discussion on the psychological and phenomenological foundations, please refer to Appendix A. For examples of the methods described in this chapter, please refer to Appendix H.

3.1 Internal Dynamics Model

We first propose an internal dynamics model for agents to replace the stimulus-response paradigm found in traditional agent architectures. Traditional agent architectures typically rely on the environment to provide perceptual input and generate actions to feed back to the environment within a single time step. The proactivity constructed under this paradigm remains fundamentally a form of passivity. It still relies on information acquired from the environment at each step to generate behaviour.

In contrast, our internal dynamics model allows the agent to spontaneously generate internal thoughts even in the absence of external input. The agent can then autonomously choose whether to remain silent or output behaviours to the external environment, as illustrated in Figure 1.

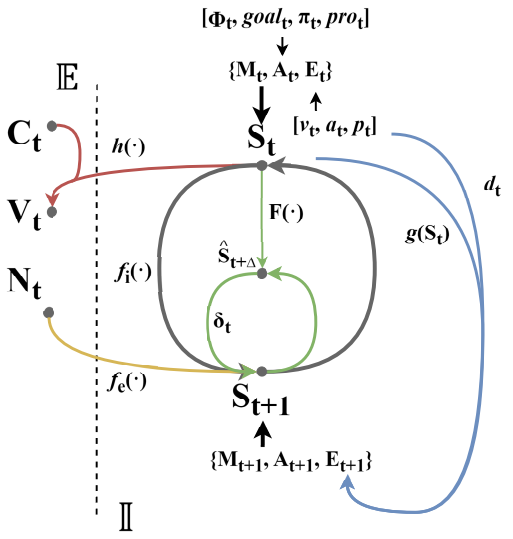


Figure 1: **The internal dynamics model of the intentional agent.**

Specifically, let S_t denote the retained internal state of the agent. Its candidate update over time follows:

$$\tilde{S}_{t+1} = f_i(S_t) + f_e(S_t, N_{t+1}), \qquad (1)$$

where N_{t+1} denotes new external perception, $f_i(\cdot)$ denotes the internal update driven by the retained state, and $f_e(\cdot)$ denotes modulation caused by new external input. The plus sign conceptually separates these two sources rather than indicating element-wise addition over heterogeneous representations. When no new input arrives, $f_e(S_t, \varnothing) = 0$, while $f_i(S_t)$ can continue the state update. Eqs. 1–7 are not derived from a single probabilistic model or objective; rather, they constitute a heuristic formalisation informed by appraisal theory, predictive processing, the Valence–Arousal–Dominance (VAD) model, and embodied intentionality. Their equation-wise theoretical basis, design rationale, and empirical validation are detailed in Appendix B.

Here, A_t represents a multi-level action preparation state consisting of four components:

$$A_t = \{\pi_t, \text{goal}_t, \phi_t, \text{pro}_t\} \qquad (2)$$

Where π_t represents the agent's current action tendency over selectable actions. goal_t represents the purpose of the action, ϕ_t represents its feasibility under the current environmental constraints, and pro_t represents its expected result. This factorisation is informed by ITCMA (Zhang et al., 2025) and makes the reason for an action and its expected consequence explicitly inspectable. The expected future state used by the subsequent predictive cycle is represented as:

$$\hat{S}_{t+\Delta} = F(S_t) \qquad (3)$$

Where $\hat{S}_{t+\Delta}$ denotes the expected part of the future state and F is the internal prediction model. In the implementation, the base LLM produces a semantic expectation for comparison with a later observation. The prediction error is then formulated as:

$$\delta_t = C_{\text{LLM}}(\hat{S}_{t+\Delta}, N_{t+\Delta}), \qquad (4)$$

where $C_{\text{LLM}}(\cdot)$ denotes the implemented comparison in which the base LLM receives the previous expectation and later observation and returns a scalar prediction error. A stronger conflict produces a higher error. The error is then used to correct the candidate state:

$$S_{t+1} \leftarrow J(\tilde{S}_{t+1}, \delta_t), \qquad (5)$$

where $\tilde{S}_{t+1}$ denotes the candidate state produced by Eq. 1, while S_{t+1} denotes the final corrected internal state obtained after applying the prediction error through $J(\cdot)$.

$E_t \leftarrow [v_t, a_t, d_t]$ serves as the emotion acting as the driving force. However, it is no longer an emotional representation denoted by basic emotions but a dynamical variable. Here, $v_t \in [-1, 1]$ is the valence dimension, representing the degree of attraction and repulsion of the current state relative to the environment or goal. $a_t \in [0, 1]$ is arousal, representing the overall activity level of the system. High activation accelerates action, while low activation inhibits action. d_t is dominance, representing the quantification of controllability over the external world. The evolution of E_t follows this process:

$$\begin{aligned} E_{t+1} &= \text{clip}\left(\alpha(\delta_t)E_t + [1 - \alpha(\delta_t)]g(S_t, N_{t+1}, \delta_t), -1, 1\right), \\ \alpha(\delta_t) &= 0.7(1 - 0.5\delta_t). \end{aligned} \qquad (6)$$

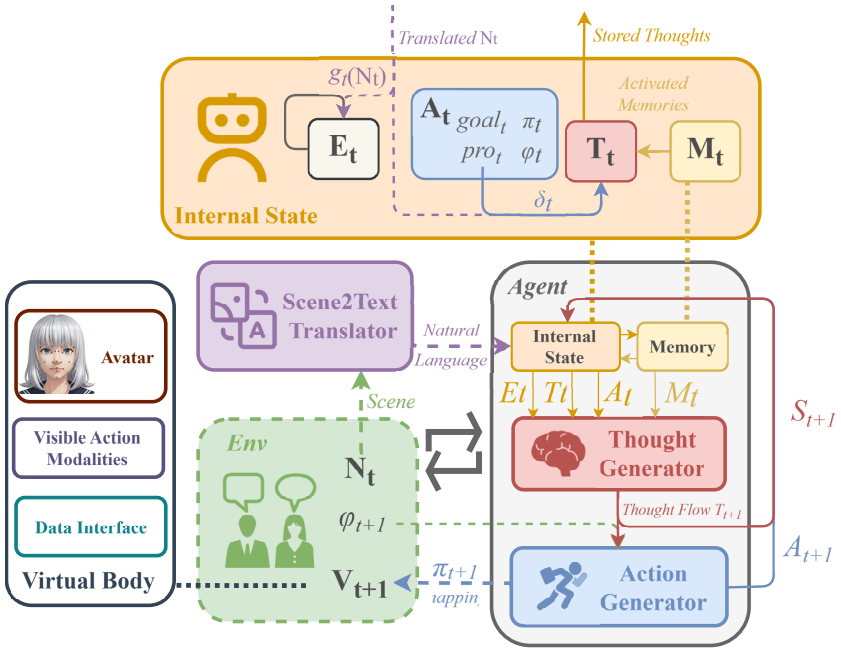


Figure 2: **The overall architecture of the intentional agent.**

Here, the rule-based operator $g(\cdot)$ obtains a new VAD appraisal from current perception, memory, and thought. A small prediction error retains more of the previous emotion, while a large error gives unexpected information more influence. The coefficients 0.7 and 0.5 are empirical choices. E_t continues to influence action selection and directly alters the preferences of the action distribution π_t.

The virtual body V_t serves as the embodied interface between the agent and the external interactive environment. Crucially, within the context of Facial Reaction Generation (FRG), this module is explicitly responsible for mapping the agent's unobservable internal affective states onto the visual avatar. This projection process is formulated as:

$$V_t = h(S_t, C_t) \tag{7}$$

Where V_t denotes the virtual body state (i.e., the avatar's facial expressions and visible action modalities), and C_t represents the external environmental or physical constraints (such as the rigging limitations of the 3D model). The function $h(\cdot)$ acts as the projection mechanism that translates the latent internal state S_t, which now encapsulates the generated multi-dimensional emotional drives, into the physical world. The detailed rules and implementation of affective-to-facial mapping are provided in Appendix D.

3.2 Intentional Agent

Based on the proposed dynamics model, we introduce the intentional agent. This is an agent framework featuring a mechanism similar to overhearing agents (Zhu & Callison-Burch, 2025). It can monitor environmental signals, perceive user activities, and intervene to respond. Furthermore, due to our foundational dynamics model, the intentional agent can continuously generate states and behaviours through the evolution of internal thoughts without external triggers. It can also proactively shape interactions driven by the prediction model and emotional drives. The overall architecture is in Figure 2.

For engineering implementation purposes, the intentional agent maintains an Inner Thought Flow (ITF) T_t as part of its structured internal state. We use ITF as an engineering representation of the intermediate state used for subsequent generation; we do not treat it as evidence that the agent has human consciousness or cognition. ITF is mainly prompt-driven at inference, and the base LLM is not fine-tuned to learn it. During silence, updates are initiated by a timed heartbeat rather than new external events. The thought generator updates from retained information without a new external

event; the prompt keeps expectation separate from observation, prohibits unsupported external events, and requires prediction mismatches to be addressed in the subsequent thought. At each update, the retained state is combined with current perceptual and prediction-error information; the thought generator produces the next thought, after which the action generator produces a possible action state. This mechanism is somewhat similar to the framework by Liu et al. Liu et al. (2025). However, their approach still relies on the passive triggering of inner thoughts by external events as a starting point to address the optimisation of dialogue flows. In contrast, the intentional agent focuses on how the agent initiates intentional actions towards the external world. Under our internal dynamics framework, this implies how the agent acts upon the environment even when it is situated alone. Here, T_t, emotion vector E_t, action vector A_t, and activated memory content M_t together constitute the retained internal state $S_t \leftarrow \{T_t, E_t, A_t, M_t\}$. External perception N_t is supplied separately to the current update and can be empty. Conventional temporal models usually encode interaction history in a learnt hidden vector whose individual dimensions have no predefined meanings. In our framework, information used for subsequent updates is organised into explicit, named components rather than compressed into a single opaque hidden representation. These components can therefore be inspected and changed separately. At each update, the base LLM used as the thought generator receives the current state to generate the next thought, after which the action generator produces the next action state. The updated thought and action then enter the affective and facial generation process. Because T_t largely contains working memory, the memory module primarily stores long-term memory. During the generation of T_{t+1}, T_t and the possible N_t are inputted into the memory module to calculate cosine similarity for extracting the most similar memory elements. Subsequently, a LLM integrates the text of the memory element itself, its corresponding temporal attributes, and potential executed actions into a descriptive text segment M_t. After this, T_t is stored by the memory module.

When there is an external perception input N_t, this contextual information is translated into a text description and supplied to the current update. It intervenes in the update of E_t and calculates the prediction error δ_t together with pro_t to be incorporated into the thought flow. This optimizes the anticipation of the agent regarding the future. During each internal iteration, the internal state first inputs T_t into the memory module to obtain the activated memory M_t. Then, the internal state S_t is inputted into a LLM named the thought generator. This continues the thought flow T_t based on the current state to obtain T_{t+1}. This result is inputted into the action generator to produce the action tendency π_{t+1}, its purpose goal_{t+1}, its feasibility under the environmental constraint ϕ_{t+1}, and its expected result pro_{t+1}. These components form the action vector A_{t+1} at time $t+1$. This vector participates again alongside T_{t+1} to construct the internal state S_{t+1}. If π_{t+1} is not empty, it is inputted into the virtual body V_t to execute actions within the environment. If pro_{t+1} is not empty, it will be compared with the perceived information N_{t+1} upon the next acquisition. The reflection on the expected outcome is then placed into the thought flow T_{t+1}. When there is no external perception input N_t, the emotion vector E_t updates itself at each step based on T_{t+1}. Otherwise, it undergoes a weighted fusion comprehensively considering N_t.

State maintenance during silence therefore does not depend on T_t alone. Without a new event, previous memory, emotion, thought, and expectation continue to enter the next update and regulate action generation. In the controlled component test, Structured achieved higher direction accuracy than Unstructured (95.83% vs. 89.58%) and a larger signed output shift in the intended direction ($+.274$ vs. $+.160$, $p = .0163$). The internal-state ablation in Section 4.5 further shows that the named components affect later decisions during silence. A 30-minute no-input stability test further checks long-horizon structural stability and unsupported-event generation; detailed controlled tests are reported in Appendix B.4.

4 EXPERIMENTS

4.1 SETTINGS

Dataset. We evaluate our proposed MAFRG agent on the React2025 MARS benchmarking dataset (Song et al., 2025). It comprises 2,856 dyadic multimodal conversation pairs derived from 137 real human interaction clips (involving 23 speakers and 137 listeners, with 20-35 minutes per session), encompassing synchronized audio and facial videos. Crucially, to rigorously test the agent's endogenous dynamics under zero external input, we intentionally retain all natural conversational silence segments without traditional pruning.

To test within-dataset generalisation without overlap between nearby clips from the same recording, we additionally use a recording-group-disjoint split; Full achieves the lowest VaG (.04399), VeG (.01369), and DG (.09013) among the three conditions, with full results reported in Appendix E.2.

We further evaluate cross-benchmark adaptation on *Seamless Interaction* (Agrawal et al., 2025): keeping ITF, prompts, and the VAD-to-Action Unit (AU) interface fixed while adapting only the facial Diffusion Transformer (DiT) reduces AU mean absolute error (AU-MAE) from .19081 to .13973 on held-out Seamless data. Full adaptation and benchmark results are reported in Appendix E.4.

Implementation Details. We use Qwen3-8B as the base LLM for the intentional agent and do not fine-tune it. Interaction samples from the React2025 training set populate the agent's memory. Agent emotions are modelled in the VAD space (Russell & Mehrabian, 1977) using the NRC VAD Lexicon v2 (Mohammad, 2025).

We do not assume a one-to-one mapping from thought text to a precise AU trajectory. The NRC-based text-to-VAD analyser is directly validated against human VAD annotations on EmoBank (Buechel & Hahn, 2017): Pearson correlations for Valence, Arousal, and Dominance are .512, .236, and .151, respectively, with full results reported in Appendix D.1. The resulting VAD state provides a low-dimensional affective constraint, the rule-based VAD-to-AU mapping provides a low-frequency facial guide, and the DiT learns the remaining facial dynamics. Comprehensive details are provided in Appendix C, Appendix D, and Appendix E.

Interactive Feasibility. Although our setting does not strictly require continuous real-time generation, we additionally measured interactive feasibility. The 6.77M-parameter DiT generated 375 frames in 0.079s using six diffusion steps. With local Qwen3.5-9B, the state update and full-system update averaged 4.29s and 4.37s, respectively. External events initiate updates immediately, while the tested silence heartbeat interval was 15s. Detailed hardware and memory measurements are provided in Appendix E.5.

Metrics. We comprehensively evaluate the framework using facial reaction metrics, human evaluation, and action evaluation metrics. First, Facial Reaction Metrics assess the physical generation quality using standard React2025 criteria: Appropriateness (FRCorr, FRDist), Diversity (FRVar, FRDiv), and Synchrony (FRSyn). Second, we conduct a blinded human evaluation over five rating dimensions, including contextual appropriateness and temporal continuity. Third, Action Evaluation Metrics quantify the agent's internal intentionality and zero-input action performance via a 1–5 Likert scale LLM-as-a-Judge framework. This evaluates Affect-Action Coherence, Pragmatic Utility, Temporal Intentionality, Role Compliance, and Affective Appropriateness against real human baselines. RQS is used as a supplementary automatic metric and is validated against human judgement in Section 4.3. Detailed metric definitions are provided in Appendix E, while the RQS architecture and human validation are provided in Appendix F.

4.2 Comparison with existing MAFRG methods

We compare our method with existing approaches in Table 1. Intentional Agent obtains the lowest FRDist (72.39) and the highest FRDiv (.5057) and FRVar (.0883) among the compared learned methods, with an FRSyn of 47.59. Because FRDiv and FRVar also exceed the ground-truth values, we interpret them as measures of reaction coverage rather than direct evidence of realism. We therefore evaluate plausibility separately through blinded human ratings and a controlled diversity analysis in Appendix E.3.

Intentional Agent records a lower FRCorr (0.08) than several end-to-end baselines. Because FRG is a one-to-many task, neither exact ground-truth correlation nor diversity alone provides a complete quality measure. We therefore report the standard metrics together with human evaluation, the controlled diversity audit in Appendix E.3, and the qualitative comparisons in Figure 6.

4.3 Evaluation of the Reaction Quality Scorer (RQS)

In natural conversations, a single speaker trigger can elicit multiple semantically reasonable listener reactions. We therefore assess generated reactions using trajectory metrics and human ratings, with RQS used as a supplementary metric.

Table 1: **Quantitative comparison on the React2025 dataset.** Bold text highlights the best results. The arrows (↑) and (↓) indicate whether higher or lower metric values represent better performance respectively.

Method	Appropriateness		Diversity		Synchrony
	FRCorr (↑)	FRDist (↓)	FRDiv (↑)	FRVar (↑)	FRSyn (↓)
GT (Song et al., 2025)	10.00	0.00	0.1876	0.0669	48.66
B_Random (Song et al., 2025)	0.03	474.68	0.3342	0.1671	46.64
B_Mime (Song et al., 2025)	0.52	206.02	0.0000	0.0766	43.70
B_MeanFr (Song et al., 2025)	0.00	205.65	0.0000	0.0000	49.00
Trans-VAE (Song et al., 2025)	0.30	181.72	0.0076	0.0083	49.00
ReGNN (Xu et al., 2026)	0.54	155.49	0.0012	0.0048	**44.81**
PerFRDiff (Zhu et al., 2024)	0.56	177.76	0.1386	0.0706	48.54
SC-Diff (Mao et al., 2025)	0.59	214.67	0.0400	0.0811	49.00
Huang & Yu (2025)	0.67	178.27	0.1510	0.0801	48.07
Wang et al. (2025)	**0.71**	173.01	0.1405	0.0769	47.77
Intentional Agent	0.08	**72.39**	**0.5057**	**0.0883**	47.59

To complement the standard metrics, we train a separate Reaction Quality Scorer (RQS) on REACT 2025. RQS is trained using real matched speaker-listener samples and constructed negative samples; outputs from our facial reaction generator are not used during RQS training. This avoids direct overlap with the tested generator outputs, although RQS remains trained within the REACT 2025 distribution. A stricter retraining that excludes all 14 recording groups represented in the human-rated set still correlates with human scores (Pearson $r = .744$; Spearman $\rho = .788$); details are reported in Appendix F.5. We therefore treat it as a supplementary metric and validate its agreement with blinded human ratings below, with architectural and training details provided in Appendix F.

Evaluation Results and Analysis. On the REACT 2025 test set, RQS obtains a mean score of .705 and a median score of .854 over 559 samples, compared with a GT score of .854, corresponding to an overall relative ratio of 82.6%. Session-level results show substantial variation across conversational contexts and are reported in Appendix F.4. Because RQS is used only as a supplementary automatic metric, we evaluate its agreement with human judgements below.

Human Evaluation and RQS Validation. Eight blinded human raters evaluated 96 reactions from 24 settings and four conditions. Full did not differ significantly from GT in mean human score (5.527 vs. 5.195, $p_{\mathrm{Holm}} = .076$), while significantly outperforming Event-Triggered (3.059) and Heuristic-Only (1.445), both $p_{\mathrm{Holm}} < .001$. RQS correlated strongly with human scores (Pearson $r = .855$; Spearman $\rho = .821$, both $p < .05$). After removing condition-level mean differences, the correlations remained high ($r = .908$, $\rho = .892$), supporting RQS as a human-validated supplementary metric. Full human-evaluation details are reported in Appendix F.6; Appendix F.7 further compares RQS with six existing automatic metrics on the 72 human-rated generated reactions and tests whether it provides predictive information beyond them.

4.4 Action-Level Analysis of Generated Reactions

To further analyse the action-level behaviour of the agent, Table 2 presents the quantitative results of the LLM-as-a-Judge evaluation (Luo et al., 2025; Bojić et al., 2025) for the Intentional Agent with different foundation models and the ground truth. We retain this evaluation as a secondary diagnostic of the generated actions. The blinded human study in Section 4.3 provides the direct perceptual comparison with ground truth, where Full and GT did not differ significantly in mean human score (5.527 vs. 5.195, $p_{\mathrm{Holm}} = .076$). Accordingly, the Full versus GT comparison is reported from the human study, while Table 2 is used as an auxiliary action-level analysis.

Across the tested foundation models, the highest score differs by evaluation dimension. DeepSeek-V3 obtains the highest Affect-Action Coherence, Temporal Intentionality, and Role Compliance scores, while Qwen3.1-8B obtains the highest Pragmatic Utility and Affective Appropriateness scores. We

Table 2: **Multi-dimensional LLM-as-a-Judge evaluation of generated actions.** We evaluate the Intentional Agent with different foundation models. The top result for each dimension is bolded. Higher values on the 1–5 Likert scale indicate higher judged performance on the corresponding action-level criterion.

Model	Affect-Action Coherence	Pragmatic Utility	Temporal Intentionality	Role Compliance	Affective Appropriateness
GT	3.00	1.93	2.43	3.07	2.64
Qwen3.1-8B	3.23	**3.23**	2.85	3.62	**3.92**
Llama3.1-8B	3.21	3.07	3.36	3.43	3.79
GPT-4o	2.71	2.43	3.50	3.36	3.43
DeepSeek-V3	**3.29**	2.36	**3.79**	**3.86**	3.79

treat these LLM-as-a-Judge results as an auxiliary action-level diagnostic rather than a direct measure of perceptual facial quality.

4.5 Ablation Studies

We conducted two groups of ablations to examine the facial generation components and the internal state. In the facial ablation, Heuristic-Only, Learned-Only, and Full obtained RQS scores of .5350, .7694, and .7807, respectively, across 60 scenes and three seeds. After removing high-frequency information from the generated AU trajectories, adding the hand-crafted mapping reduced AU-MAE from .31916 to .31422 and improved 90% of the scenes ($p = 5.97 \times 10^{-10}$).

We also ablated the internal state in 20 long-silence scenes. Full had the lowest action-trigger rate at .333, while Only-Perception, Only-Emotion, Only-Thought, and Only-Expectation produced .800, .667, .750, and .583, respectively, with all four significantly higher than Full ($p \leq .05$). Only-Memory produced .350. Full also had the lowest expectation error at .0356. In a separate six-context 30-minute drift test, Full retained key initial facts more often than Only-Thought (62.08% vs. 4.17%), with higher retention in all six contexts ($p = .0156$). Its mean semantic-similarity slope was also more stable ($+4.21 \times 10^{-5}$ vs. -1.72×10^{-4}), with Full more stable in five of six contexts ($p = .0313$). Complete results are reported in Appendix B.4.

5 Conclusion

In this paper, we reformulate Facial Reaction Generation (FRG) from a direct black-box mapping task into a structured internal-state generation process inspired by cognitive modelling and propose the Intentional Agent framework. By maintaining an explicit, self-updating internal state that integrates thought and affective variables, the agent generates proactive, coherent non-verbal behaviours even during complex social silences. Evaluations on the REACT 2025 dataset obtain the lowest FRDist and the highest FRVar and FRDiv among the compared learned methods, with diversity interpreted alongside human ratings and trajectory-level audits. To complement the standard frame-aligned metrics, we introduce a Reaction Quality Scorer (RQS) and evaluate it against human judgements. A blinded human evaluation found no significant difference in mean score between Full and GT (5.527 vs. 5.195, $p_{\text{Holm}} = .076$), while Full significantly outperformed Event-Triggered and Heuristic-Only (both $p_{\text{Holm}} < .001$). RQS correlated strongly with human scores and is therefore used as a supplementary automatic metric.

However, this study still possesses certain limitations. Maintaining a continuously operating inner thought flow for spontaneous reasoning demands extensive computational resources and low inference latency from the underlying LLMs, posing real-time challenges in resource-constrained deployment environments. In our measured configuration, the full system required 4.37s per update on average with local Qwen3.5-9B, although our setting does not strictly require continuous real-time generation. In future work, we will explore lightweight internal dynamics algorithms to optimize computational overhead. Additional limitations concerning facial-system adaptation and potential misuse are discussed in Appendix G.

AI USE STATEMENT

The authors used generative AI tools strictly for assisting with English language polishing, grammar correction, and improving readability during the preparation of this manuscript. All conceptual ideas, methodology formulations, mathematical designs, experimental implementations, and data analyses were independently developed and conducted by the human authors. Large language models (such as Qwen3-8B and DeepSeek) mentioned in the methodology and evaluation sections were utilized strictly as experimental components and evaluators as described in the paper, not as authors. All text and citations were thoroughly verified, edited, and validated by the authors.

REFERENCES

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, et al. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset, 2025.

Dana Atzil-Slonim, Christina S Soma, Xinyao Zhang, Adar Paz, and Zac E Imel. Facilitating dyadic synchrony in psychotherapy sessions: Systematic review and meta-analysis. *Psychotherapy Research*, 33(7):898–917, 2023.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7924–7936, 2024.

Ljubiša Bojić, Predrag Kovačević, and Milan Čabarkapa. Does gpt-4 surpass human performance in linguistic pragmatics? *Humanities and Social Sciences Communications*, 12(1):1–10, 2025.

Sven Buechel and Udo Hahn. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 578–585, 2017.

Justine Cassell and Andrea Tartaro. Intersubjectivity in human–agent interaction. *Interaction studies*, 8(3):391–410, 2007.

Justine Cassell and Kristinn R. Thorisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(4-5):519–538, 1999. doi: 10.1080/088395199117360.

Lorena Chanes, Jolie Baumann Wormwood, Nicole Betz, and Lisa Feldman Barrett. Facial expression predictions as drivers of social perception. *Journal of personality and social psychology*, 114(3): 380, 2018.

Xuangeng Chu, Nabarun Goswami, Ziteng Cui, Hanqin Wang, and Tatsuya Harada. Artalk: Speech-driven 3d head animation via autoregressive model, 2025.

Yang Deng, Lizi Liao, Wenqiang Lei, Grace Hui Yang, Wai Lam, and Tat-Seng Chua. Proactive conversational ai: A comprehensive survey of advancements and opportunities. *ACM Transactions on Information Systems*, 43(3):1–45, 2025.

Ulf Dimberg, Monika Thunberg, and Kurt Elmehed. Unconscious facial reactions to emotional facial expressions. *Psychological science*, 11(1):86–89, 2000.

Thomas Fuchs. *Ecology of the brain: The phenomenology and biology of the embodied mind*. Oxford University Press, 2017.

Cristina Garbacea and Chenhao Tan. Hyperalign: Interpretable personalized llm alignment via hypothesis generation. *arXiv preprint arXiv:2505.00038*, 2025.

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.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv preprint arXiv:2207.05608*, 2022.

Edmund Husserl. *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy: First book: General introduction to a pure phenomenology*. Springer Science & Business Media, 2012.

Zhangyu Jin, Maksim Siniukov, Deuksin Kwon, Ashutosh Chaubey, and Mohammad Soleymani. Gdpo-listener: Expressive interactive head generation via auto-regressive flow matching and group reward-decoupled policy optimization, 2026.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Joris Kirchner, Amogh Gudi, Marian Bittner, and Chirag Raman. Controlled face manipulation and synthesis for data augmentation. In *2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–10. IEEE, 2026.

J Lan. General data, virtual bodies, and digital capital: Three logics of digital capitalism. *Philosophical Research (in Chinese)*, 3:26–33, 2018.

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.

David Lindevelt, Suzan Verberne, and Joost Broekens. The correlation between emotion in text and speech segments is limited: A cross-modal study. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 2611–2621, 2026.

Xingyu Bruce Liu, Shitao Fang, Weiyan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang'Anthony' Chen. Proactive conversational agents with inner thoughts. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 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, 2024. doi: 10.1109/FG59268.2024.10582013.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. Proactive agent: Shifting llm agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361*, 2024.

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, 2024.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*, 2025.

Qincheng Lv, Xiaofeng Liu, Jie Li, Rongrong Ni, Pujun Xue, and Siyang Song. Hierarchical multimodal decoupling-fusion framework for offline multiple appropriate facial reaction generation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887743.

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.

Stacy C Marsella and Jonathan Gratch. Ema: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, 2009.

Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current psychology*, 14(4):261–292, 1996.

Maurice Merleau-Ponty, Donald Landes, Taylor Carman, and Claude Lefort. *Phenomenology of perception*. Routledge, 2013.

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 174–184, 2018.

Saif M Mohammad. Nrc vad lexicon v2: Norms for valence, arousal, and dominance for over 55k english terms. *arXiv preprint arXiv:2503.23547*, 2025.

Louis-Philippe Morency, Iwan De Kok, and Jonathan Gratch. Predicting listener backchannels: A probabilistic multimodal approach. In *International workshop on intelligent virtual agents*, pp. 176–190. Springer, 2008.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20363–20373. IEEE, 2022.

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.

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, 2024b. doi: 10.1109/FG59268.2024.10581978.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21055–21064. IEEE, 2025.

David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526, 1978.

Yilun Qiu, Xiaoyan Zhao, Yang Zhang, Yimeng Bai, Wenjie Wang, Hong Cheng, Fuli Feng, and Tat-Seng Chua. Measuring what makes you unique: Difference-aware user modeling for enhancing llm personalization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21258–21277, 2025.

Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580.

James A Russell and Albert Mehrabian. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294, 1977.

Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

Klaus R Scherer, Angela Schorr, and Tom Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

Liang Shi, Yiyang Huang, and Yun Fu. Capturing individual differences of facial expression for authentic expression generation. In *2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9. IEEE, 2026.

Ojas Shirekar, Wim Pouw, Chenxu Hao, Vrushank Phadnis, Thabo Beeler, and Chirag Raman. Multimodal quantitative measures for multiparty behavior evaluation. In *Proceedings of the 27th International Conference on Multimodal Interaction*, pp. 249–264, 2025.

Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, and Chenliang Xu. Emotional listener portrait: Realistic listener motion simulation in conversation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20782–20792. IEEE, 2023a.

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, 2023b.

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*, 2023c.

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 (FG)*, pp. 1–5. IEEE, 2024.

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.

Lorenzo Stacchio, Andrea Ubaldi, Alessandro Galdelli, Maurizio Mauri, Emanuele Frontoni, and Andrea Gaggioli. Empathic prompting: Non-verbal context integration for multimodal llm conversations. *arXiv preprint arXiv:2510.20743*, 2025.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387–2395, 2016.

Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413, 2020.

Niklas Wagner, Felix Mätzler, Samed R Vossberg, Helen Schneider, Svetlana Pavlitska, and J Marius Zöllner. Cage: Circumplex affect guided expression inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4683–4692, 2024.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

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.

Mengting Wei, Tuomas Varanka, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. Magicface: High-fidelity facial expression editing with action-unit control. *arXiv preprint arXiv:2501.02260*, 2025.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101, 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.

Xie G Chen F-Yu Y Peng L Li M Yu J, Zhao J and Dai Z. 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.

Hanzhong Zhang, Jibin Yin, Haoyang Wang, and Ziwei Xiang. Simulating phenomenal consciousness using generative agents based on large language models. *Applied Soft Computing*, 185:113922, 2025. doi: 10.1016/j.asoc.2025.113922.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8652–8661, 2023.

Sebastian Zhao, Alan Zhu, Hussein Mozannar, David Sontag, Ameet Talwalkar, and Valerie Chen. Codinggenie: A proactive llm-powered programming assistant. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, pp. 1168–1172, 2025a.

Yuheng Zhao, Xueli Shu, Liwen Fan, Lin Gao, Yu Zhang, and Siming Chen. Proactiveva: Proactive visual analytics with llm-based ui agent. *IEEE Transactions on Visualization and Computer Graphics*, 2025b.

Andrew Zhu and Chris Callison-Burch. Overhearing llm agents: A survey, taxonomy, and roadmap. *arXiv preprint arXiv:2509.16325*, 2025.

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 The Psychological and Philosophical Foundations of Intentional Agents

The Facial Reaction Generation (FRG) task has made important progress in cross-modal alignment and visual fidelity. However, the vast majority of existing works adopt an end-to-end deep learning paradigm. From the perspective of human psychology and cognitive science, this paradigm essentially falls into the reductionist trap of Behaviourism: it downgrades the complex facial reactions of a listener into a pure Stimulus-Response mechanism, attempting to establish a direct black-box mapping between the speaker's audio-visual features and the listener's facial muscle movements.

In authentic non-verbal social interactions, human facial expressions are never merely passive reactions to external signals. End-to-end models attempt to forcefully bypass the most core component of interaction: the internal cognitive appraisal process (Scherer et al., 2001; Scherer, 2005). Consequently, while the generated expressions may achieve frame-level visual realism, they lack coherent logical depth in long-horizon social contexts. In fact, the visible tip of the iceberg accounts for less than one percent of its entirety. The facial reactions exhibited by humans are fundamentally

the subconscious projection of complex internal processes. Therefore, to achieve meaningful and interpretable facial reaction generation akin to human behaviour, it is imperative to shift the focus toward simulating internal cognitive processes.

"Intentionality" is a philosophical concept. In traditional Husserlian phenomenology, Intentionality is often statically defined as "consciousness of something," emphasizing the subject's passive perception and representation of external objects (Husserl, 2012). However, in the context of this paper, intentionality implies first constructing a self-sustaining "Internal Stream of Consciousness." Within this continuously flowing latent state, the agent spontaneously generates a directedness toward the external world, which ultimately translates into embodied action outputs. This aligns with the concept of "Operative Intentionality" proposed by Merleau-Ponty in embodied phenomenology (Merleau-Ponty et al., 2013). Operative intentionality is not abstract contemplation enclosed within the brain, but rather a pre-reflective vital force, directly driven by internal bodily tension and extending into the world. This differs from the concept of "proactivity" frequently discussed in current AI agent literature. These so-called "proactive" capabilities remain essentially reactive. In practice, conventional proactivity merely means the agent exhibits the capacity for spontaneous triggering and action planning driven by external inputs and environmental cues. Its root cause still lies in exogenous algorithmic mechanisms and interactive feedback structures.

Figure 3 illustrates the comparison among intentional, proactive, and reactive agents. Proactivity frees agents from relying on passive inputs. In contrast, intentionality requires agents to possess an autonomous generative mechanism within their internal dynamics.

In other words, current proactivity only solves the problem of "how to act" under specific conditions, while lacking an endogenous drive when it comes to "why to act." This limitation becomes particularly glaring in real-time, natural conversations, especially in scenarios involving extensive non-verbal intervals and social silence. When real-time interaction enters a blank period lacking explicit external commands or environmental cues, this exogenously driven logic often leads to a breakdown in cognitive coherence and a stagnation in behavioural generation. This prompts us to rethink the architectural bottleneck: rather than merely simulating proactivity, we should introduce an endogenous operative intentionality at the architecture. That is, the system should be able to internally generate an understanding of objects, goals, and values, and utilize the body to extend this understanding outward.

Naturally, we must first define the body of the agent. Lan Lan (2018), in his research on digital capitalism, proposed the concept of the "virtual-body". According to Lan, just as we possess a physical "natural body" in reality, we also inhabit a "virtual-body" when engaging with internet platforms and digital spaces. This virtual-body is composed of computable, identifiable, and trackable data. For instance, to a social media user, the account is merely an intermediary for participation. However, to the platform itself, the account is the only truly existent object. The platform populates this initially "empty formal set" with content (clicking data, emotional expressions, and attentional investment), thereby "fleshing out" the digital presence. Between the user's choices and the platform's corresponding shaping, the presence of this virtual-body becomes increasingly prominent.

For agents, the "virtual-body" theory similarly reveals their inherent corporeality. However, a crucial distinction must be made here between the "body" and the "soma". In natural language, "body" often refers to the physical manifestation, that is, the soma. This is the material substrate of a living entity. The soma and its physiological processes can only be perceived from an external, third-person perspective with a naturalistic attitude. Yet, human vital expression is simultaneously internal and external, encompassing both experience and behaviour. Therefore, from a phenomenological perspective, there must exist a "body" that is inextricably linked to first-person experience, that is, the flip side of the soma. This "body" is connected to the lifeworld through perception and is by no means confined within the material limits of the soma. This distinction reminds us that consciousness is not encapsulated within an organism but enters the world through the openness of the body.

For a digital agent, its underlying neural network infrastructure, data structures, and the mesh/rigging of its 3D avatar constitute merely its "soma". The soma is the material substrate of the system and must be observed from a third-person naturalistic perspective. However, when these somatic substrates are endowed with an internal circulation mechanism within the digital interaction network, the agent truly acquires a body for perception, communication, and interaction, that is, the virtual-body. By establishing the agent's virtual-body, we open a specific field for the operation of intentionality. It

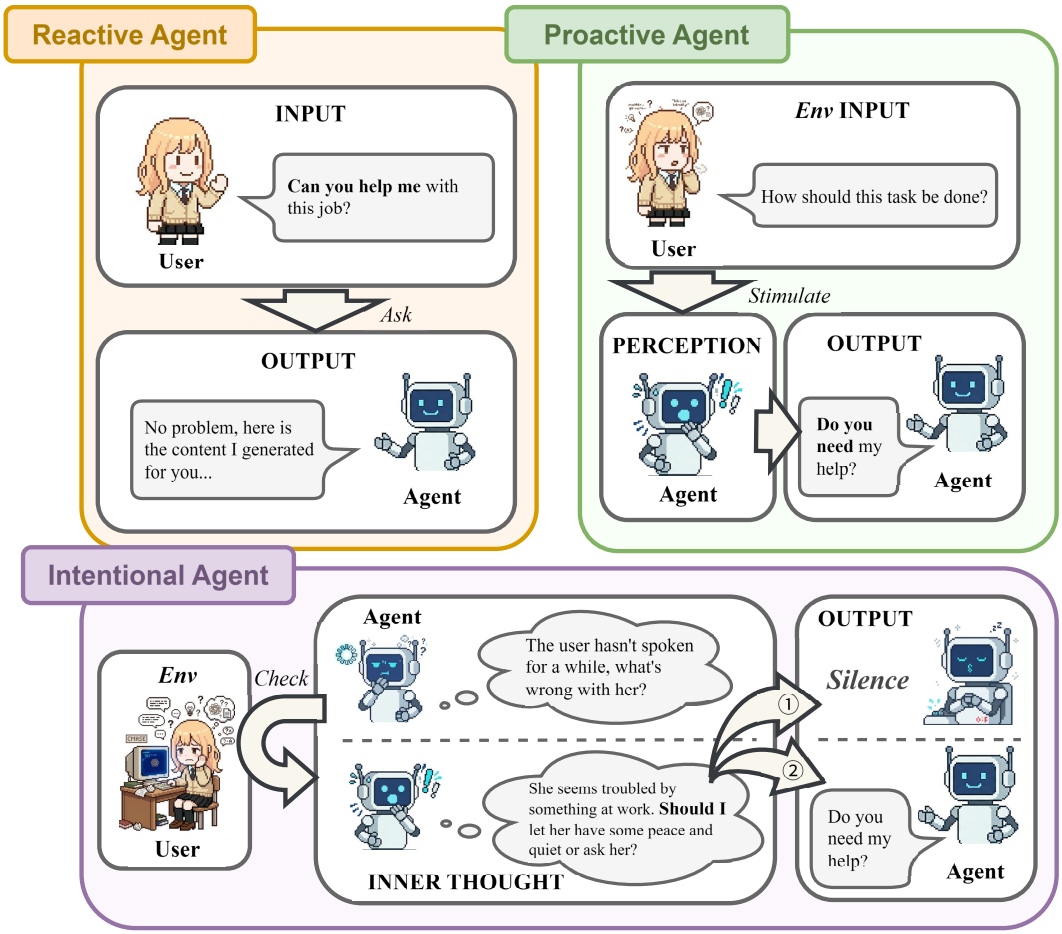


Figure 3: **Comparison of agent systems across three paradigms of human-agent interaction.** Reactive agents passively receive user queries and then generate responses. Proactive agents infer tasks based on environmental observations and propose possible assistance requests accordingly. Intentional agents maintain a continuous inner thought stream driven by endogenous intentionality, allowing them to evaluate complex contexts (such as user silence) and autonomously decide whether to initiate action or remain silently active.

transforms the agent from a dead object statically awaiting inputs into an entity maintaining continuous internal dynamic tension. In our architecture, this tension is instantiated as the continuously self-updating lnner thought flow and multi-dimensional emotion drives.

Within this framework, our understanding of Emotion undergoes a fundamental reconstruction. As pointed out by phenomenological psychiatrist Thomas Fuchs Fuchs (2017), the essence of emotion is "e-motion", where the prefix "e-" emphasizes "ex-" or "outward", that is, an outward movement. This implies that facial expressions are by no means mechanical reactions to external environmental stimuli, but rather action outputs where the internal "stream of consciousness," having crossed a certain homeostatic threshold, spontaneously extends toward the external world seeking a physical outlet.

Consequently, human-like facial reaction generation can be redefined as an "inside-out", continuous, and dynamic emergent process. External interaction contexts (such as the speaker's multimodal inputs) merely act as perturbation variables flowing into the agent's internal stream of consciousness. Even when external inputs are entirely null (e.g., prolonged social silence), driven by the psychological inertia and internal cycles induced by operative intentionality, the agent continues to undergo state transitions, persistently projecting this internal stream of consciousness (e-motion) into visible facial muscle movements.

In summary, the intentionality introduced in this paper can be viewed as the coupled outcome of a three-tier mechanism: the soma provides potential and constraints; the virtual-body opens up a dwelling place within the lifeworld; and the internal dynamics cycle and affective mechanisms endow actions with directedness. Thus, this intentionality is not a static property, but an emergent dynamical process. It enables the agent to spontaneously generate behaviour in the absence of external stimuli. Its facial movements are no longer passive reflections of external audio-visual signals, but the spontaneous physical extension (e-motion) of internal cognitive tension.

B Empirical Validation and Theoretical Grounding of Equations 1–7

B.1 Heuristic Status and Computational Requirements

Eqs. 1–7 are not derived from a single probabilistic model or objective, nor does Appendix A provide a unique mathematical derivation for them. Rather, they constitute a heuristic formalisation of appraisal theory, predictive processing, the VAD model, and embodied intentionality in Appendix A, which we translate into four concrete computational requirements:

1. **Autonomous continuation**: Internal states must be capable of continuing their temporal evolution in the absence of novel external stimuli ($N_{t+1} = \varnothing$).
2. **Action-expectation binding**: Action formulations must incorporate both an immediate behavioural choice and its anticipated perceptual outcome.
3. **Error-driven correction**: Discrepancies between anticipations and realisations must generate scalar prediction errors that correct subsequent internal states.
4. **Embodied physical realisation**: Abstract internal affect must be projected outward under concrete physical and avatar rig constraints.

Guided by these operational requirements, we chose compact, interpretable, and testable functional forms rather than claiming a unique mathematical realisation of Appendix A.

B.2 Equation-wise Theoretical Basis and Design Rationale

Table 3 provides a structured overview of the theoretical origins and functional justifications for Eqs. 1–7. Below, we delineate the specific design choices, implementations, and mathematical expressions for each equation.

Equation 1: Candidate State Generation. The candidate update $\tilde{S}_{t+1} = f_i(S_t) + f_e(S_t, N_{t+1})$ decomposes state evolution into internal persistence $f_i(S_t)$ and perceptual modulation $f_e(S_t, N_{t+1})$.

Table 3: Theoretical basis and functional design rationale for Equations 1–7.

Equation	Theoretical Basis	Functional Design Rationale
Eq. 1	Self-sustaining dynamics in Appendix A	Separates internal and input-driven changes, making silence updates testable.
Eq. 2	ITCMA (Zhang et al., 2025)	Separates action choice, purpose, feasibility, and expected result.
Eqs. 3–5	Predictive coding (Rao & Ballard, 1999)	Gives an explicit prediction, comparison, and correction cycle.
Eq. 6	Appraisal dynamics and EMA (Marsella & Gratch, 2009)	Retains emotional continuity while allowing appraisal and error to change emotion.
Eq. 7	VAD (Russell & Mehrabian, 1977; Mehrabian, 1996) and embodied intentionality	Separates affective meaning from character-specific facial realisation.

Crucially, the addition operator conceptually separates these two sources rather than indicating element-wise addition over heterogeneous text and vector representations. When external inputs cease ($N_{t+1} = \varnothing$), external modulation vanishes, $f_e(S_t, \varnothing) = 0$, while $f_i(S_t)$ can continue the state update. This formalisation directly underlies our silence ablation, which tests continued internal updating against halting after input ends. External perception is supplied as N_{t+1}, while prediction error δ_t is routed to subsequent state correction; both inform the Inner Thought Flow (ITF) without persisting as text fields in S_t.

Equation 2: Action State Factorisation. The action state $A_t = \{\pi_t, \mathrm{goal}_t, \phi_t, \mathrm{pro}_t\}$ operationalises goal-directed interaction inspired by ITCMA (Zhang et al., 2025). ITCMA motivates linking action with retained information, current constraints, and future expectation. The four-part factorisation is our design choice: π_t is the current action tendency over selectable actions, goal_t is its purpose, ϕ_t its feasibility, and pro_t its expected result. These fields make the reason for an action and its expected consequence explicitly inspectable.

Equations 3–5: Predictive Coding Cycle. Eqs. 3–5 follow the prediction, error, and correction order of predictive coding:

1. **Prediction (Eq. 3)**: $\hat{S}_{t+\Delta} = F(S_t)$ produces the expected part of a future state rather than predicting every field. In the implementation, the base large language model produces a semantic expectation for comparison with a later observation.
2. **Comparison and Error Generation (Eq. 4)**: $\delta_t = C_{\mathrm{LLM}}(\hat{S}_{t+\Delta}, N_{t+\Delta})$ represents the implemented comparison between the previous semantic expectation and the later observation. The base large language model performs this comparison and returns a scalar prediction error δ_t. This avoids treating heterogeneous internal states as vectors that can be directly subtracted. Following predictive coding, a stronger conflict produces a higher error.
3. **Posterior State Correction (Eq. 5)**: $S_{t+1} \leftarrow J(\tilde{S}_{t+1}, \delta_t)$ uses the prediction error to correct the candidate update $\tilde{S}_{t+1}$ generated by Eq. 1. Eq. 1 specifies the information sources of the candidate update, whereas Eq. 5 specifies how prediction error affects it.

Equation 6: Error-Dependent Affective Dynamics. Eq. 6 follows appraisal dynamics by treating emotion as a continuous state while making the implemented prediction-error input explicit (Marsella & Gratch, 2009):

$$E_{t+1} = \mathrm{clip}\left(\alpha(\delta_t)E_t + [1 - \alpha(\delta_t)]g(S_t, N_{t+1}, \delta_t),\ -1,\ 1\right),$$
$$\alpha(\delta_t) = 0.7(1 - 0.5\delta_t).$$

Here, the rule-based operator $g(\cdot)$ obtains a new VAD appraisal from current perception, memory, and thought. A small prediction error retains more of the previous emotion, while a large error gives unexpected information more influence. The coefficients 0.7 and 0.5 are empirical choices. This expression makes the existing implementation explicit.

Equation 7: Embodied Motor Projection. The projection $V_t = h(S_t, C_t)$ converts updated affect and action into facial behaviour under physical constraint C_t. The VAD-to-Action Unit mapping

provides an interpretable low-frequency guide, the Diffusion Transformer adds learnt high-frequency dynamics, and a rig adaptor converts Action Units into avatar-specific BlendShapes. This separates affective meaning from character-specific facial realisation rather than treating one mapping as universal.

B.3 Overall Functional Pipeline

Together, Eq. 1 generates the candidate state update, Eq. 2 prepares the action state, Eqs. 3–5 perform prediction, comparison, and correction, Eq. 6 updates emotion, and Eq. 7 produces the visible facial reaction. The resulting updated state is then propagated to the next iteration of Eq. 1. This process resembles recurrent models because it propagates state, and predictive coding because it uses a prediction–error–correction loop. However, these equations are psychology-inspired heuristic updates rather than strictly derived mathematical forms.

B.4 Controlled Empirical Validation on Component Structures

We evaluate these functional designs through controlled component tests, facial and internal-state ablations, and long-silence tests of operational stability and semantic drift.

1. Functional Effect of Structured Components (Eqs. 1–2). We compared two input forms containing exactly the same information. *Structured* placed perception, memory, emotion, thought, and expectation under separate labels, whereas *Unstructured* presented the same content as contiguous plain text without these labels. We constructed 16 paired tests. Each pair changed only memory, emotion, thought, or expectation while keeping the remaining information and seeds fixed. For example, one pair changed only whether a scholarship application was accepted or rejected.

We first measured direction accuracy, which records whether the next output follows the semantic direction of the changed component. *Structured* reached 95.83%, compared with 89.58% for *Unstructured*. Because this measure does not capture the magnitude of the response, we also calculated the signed difference between each output pair in the expected direction. A positive value indicates a change in the expected direction, while a larger value indicates a clearer change. The mean signed difference was $+.274$ for *Structured* and $+.160$ for *Unstructured* ($p = .0163$). These results show that explicitly labelled components influence subsequent updates more consistently.

2. Validation of Prediction Error Sensitivity (Eqs. 3–5). We further tested expectation because it provides the prediction error used in the subsequent state update. We constructed paired cases in which the later observation either matched or conflicted with the preceding expectation. A conflicting observation should receive a higher prediction error than a matching observation. The *Structured* formulation produced the correct error order in 100% of the cases, whereas *Unstructured* reached 66.7% ($p = .00098$). Separating expectation from current perception therefore provides a more reliable correction signal for later updates.

3. Facial Generation Ablation and Validation of Embodied Realisation (Eq. 7). We conducted two groups of ablations. The facial ablation compared: (1) **Heuristic-Only**, which retains only VAD-to-AU mapping; (2) **Learned-Only**, which uses only the causal DiT; and (3) **Full**, which combines the hand-crafted low-frequency signal with the DiT residual. Across 60 scenes and three seeds, their RQS scores were $.5350$, $.7694$, and $.7807$.

We further removed the high-frequency information from the facial Action Unit (AU) trajectories generated by the three ablation conditions in these 60 scenes. We then measured the mean absolute error between the generated and real trajectories (AU-MAE), where a lower value indicates more accurate slow facial changes. Adding the hand-crafted mapping reduced the error from $.31916$ to $.31422$ and improved 90% of the scenes ($p = 5.97 \times 10^{-10}$). During long silence, the error also decreased from $.35840$ to $.35364$ ($p = .00020$). In all 60 scenes, the changes from Learned-Only to Full were also consistent with the direction provided by the hand-crafted mapping, with a mean cosine similarity of $.241$ ($p = 1.63 \times 10^{-11}$). These results support using the hand-crafted mapping as a low-frequency guide rather than an identifiable mapping to a unique AU trajectory: it constrains the direction of slow changes, while the DiT generates the trajectory-specific dynamics.

4. Internal-State Ablation during Long Silence. We also ablated the internal state in 20 long-silence scenes without new external input. *Full* retained all components, while the remaining conditions retained only perception, memory, emotion, thought, or expectation. We measured how often the LLM selected a new action rather than no action. Full had the lowest action-trigger rate at .333. Only-Perception, Only-Emotion, Only-Thought, and Only-Expectation produced .800, .667, .750, and .583, respectively, and all were significantly higher than Full ($p \leq .05$). Only-Memory produced .350 and did not differ significantly from Full ($p = .914$). Among the tested single-component conditions, memory therefore provides the strongest constraint on action triggering during silence. Full also had the lowest expectation error, .0356, compared with .1313, .2663, and .1738 for Only-Emotion, Only-Thought, and Only-Expectation. These results show that the named components continue to have distinct effects on later updates in the absence of new external input.

5. Long-Horizon Silence Stability. We additionally ran a 30-minute long-silence test without new external input. After the client completed a description and became silent, the agent continued consecutive heartbeat-driven state updates using only the retained information from preceding steps. All updates produced structurally valid internal states and finite facial outputs. Moreover, 98.33% of the updates did not generate unsupported external events. This test indicates that the update process remains operational over prolonged silence, while unsupported event generation is rare but not completely eliminated.

6. Semantic Drift during Long Silence. The preceding 30-minute stability test measures continued operation but does not directly measure whether the semantic content of the internal state drifts from its initial context. We therefore constructed six contexts with explicit initial facts: scholarship approval, a friend's death, an interview on the next day, an unresolved family conflict, a normal medical result, and examination failure.

Each context was run for 30 minutes under two conditions. *Full* retained structured memory, emotion, previous thought, and expectation at every update. *Only-Thought* received only an unchanged perception signal and the previous thought. Both conditions used the same base LLM, prompt format, and number of updates, with seeds fixed before generation.

We first measured fact-anchor retention, defined by whether later thoughts retained the key fact supplied in the initial context. Full retained the fact anchors in 62.08% of later thoughts, compared with 4.17% for Only-Thought. Full had higher retention in all six contexts ($p = .0156$).

We then used an external pretrained text-vector model to measure semantic similarity between later thoughts and the initial memory and thought, and tracked its change over time. The mean similarity slope was $+4.21 \times 10^{-5}$ for Full and -1.72×10^{-4} for Only-Thought. Full retained the more stable semantic link in five of the six contexts ($p = .0313$).

Neither condition generated the predefined opposite facts or unsupported external events in this experiment. The difference therefore primarily reflects loss of the initial semantic anchors rather than the appearance of explicit contradictory events. These results show that retaining structured memory, emotion, and expectation reduces semantic drift during prolonged silence compared with forwarding only the previous thought.

C Details of Facial Expression Mapping

To enable the proposed agent framework to support real-world human-computer interaction, this study constructed a display and behaviour control system for the avatar based on the Unity 3D engine. At the system architecture level, an asynchronous REST API server is embedded within the frontend presentation layer. This serves as a communication hub between the underlying engine and the upper-level agent framework to receive and execute action and state manipulation commands issued by the agent hub. On the physical presentation side, the system adopts the Companion 01 device as the holographic display carrier. This device integrates a TFT-LCD panel and is covered with a multi-viewpoint directional optical film on its surface. It utilizes the principle of binocular parallax to achieve a glasses-free holographic visual effect as shown in Figure 4.

The system constructs a set of horizontally arranged microlens virtual camera arrays within the rendering pipeline. It aligns the optical axes of the cameras to intersect in real time at the core

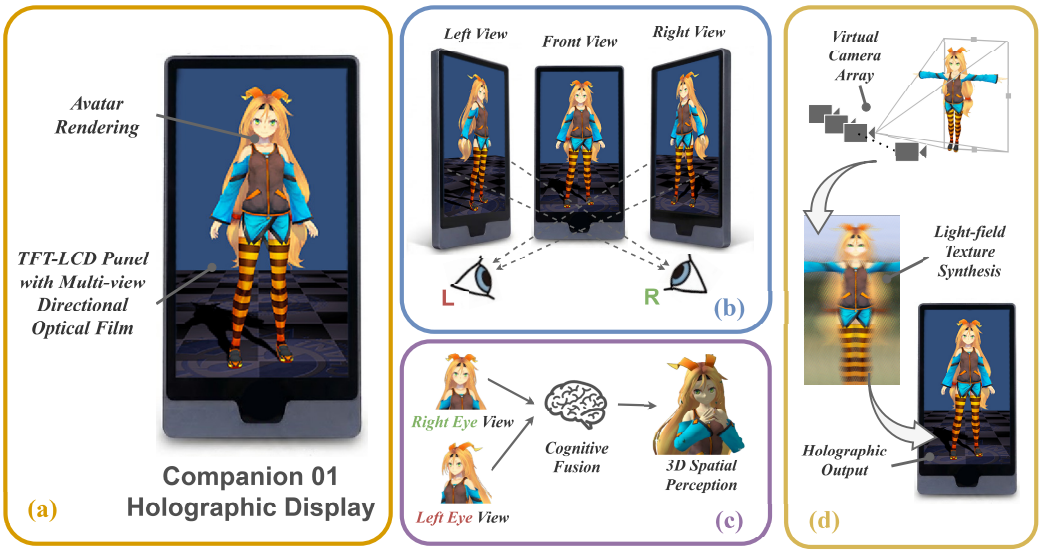


Figure 4: **Holographic Avatar display.** (a) Physical presentation of the Companion 01 holographic display; (b) The multi-view directional optical film on the TFT-LCD panel directs images of different perspectives; (c) Utilizing the binocular parallax principle, the left and right eyes receive different views which are cognitively fused in the brain to achieve a naked-eye 3D spatial perception; (d) The rendering pipeline, where a virtual camera array captures multi-view images that are processed via light-field texture synthesis to generate the final holographic output.

position of the avatar to lock the focal plane. During each frame rendering cycle, the camera array synchronously captures the scene and generates a sequence of images with slight perspective differences. Subsequently, the swizzle algorithm accurately extracts corresponding pixels from the multi-viewpoint images for spatial reorganization based on the physical refraction characteristics of the optical film on the screen surface. Finally, it synthesizes a composite texture containing the full light field information and outputs it to the physical screen. To enhance the embodied interaction sense of the virtual avatar, the display system integrates a visual perception module based on the frame difference method to capture salient motion areas in the real environment. This system can not only guide the visual attention of the virtual avatar in real time but also parameterize external environmental stimuli and feed them back to the agent framework. This subsequently influences the thinking logic and emotional state of the agent.

This facial interface is not intended to recover a unique facial trajectory from a thought representation. Facial reaction generation is one-to-many: thoughts with similar affective content may share a VAD state, while the same VAD state can support several suitable AU combinations and temporal patterns. We therefore use VAD and the hand-crafted mapping to constrain affective direction and low-frequency facial change, leaving the trajectory-specific dynamics to the learnt DiT. To enable the avatar to utilize the generated discrete text and cognitive actions, the intentional agent needs to convert them into continuous facial dynamic features aligned with the real interaction timeline. Specifically, the model first extracts the incremental VAD trajectory of the inner thought flow. Assuming the timeframe corresponding to the given inner thought flow text is $[t_{start}, t_{end}]$, the total number of frames is $N = (t_{end} - t_{start}) \times FPS$. We stretch the discrete emotional trajectory points to the target number of frames through linear interpolation and apply an energy amplification coefficient to ensure the visual tension of the expression. For time $t \in [t_{start}, t_{end}]$, the generation formulas for its valence and arousal are as follows:

$$V_{raw}(t) = \text{interp}(V_{traj}, t) \cdot a_v \tag{8}$$

$$A_{raw}(t) = \text{interp}(A_{traj}, t) \cdot a_a \tag{9}$$

Where a_v and a_a are emotional amplitude amplification coefficients.

Furthermore, because the facial muscle movements of real humans possess continuity and physical inertia, signals generated directly by interpolation will produce mechanical jumps. Therefore, we apply a 1D Gaussian filter to the global VAD signal along the time axis to simulate the smooth kinematic characteristics of facial muscles:

$$V_{smooth} = V_{raw} * G(\sigma = 12) \tag{10}$$

In addition, to prevent the listener from presenting an unnatural frozen facial state during the idle state, we superimpose a low frequency sine wave onto the global signal to simulate the natural breathing rhythm and subtle body swaying of humans:

$$B(t) = A_{Breath} \cdot \sin\left(2\pi f \cdot \frac{t}{FPS} + \phi_{run}\right) \tag{11}$$

Where A_{Breath} is the amplitude, f is the frequency, and ϕ_{run} is a random phase offset to avoid absolute synchronization among multiple generated samples.

Following the acquisition of continuous VAD signals, the affective state is first mapped into basic facial action unit vector, $U(t) \in \mathbb{R}^{15}$, in accordance with the heuristic rules of the Facial Action Coding System (FACS). This explicit mapping establishes the low-frequency, macroscopic muscle dynamics. However, authentic human reactions inherently encompass high-frequency micro-expressions that elude rigid rule-based modelling. To capture these subtle, subconsciously guided dynamics, we introduce a Diffusion Transformer (DiT) to synthesize high-frequency residual signals. The ultimate physical expression rendered on the avatar is achieved through the physical superposition of both the low-frequency and high-frequency components. Comprehensive mathematical formulations and implementation details of this generation process are deferred to Appendix D.

D Details of Emotion Vector Mapping

To convert the streaming output of the inner thought flow into continuous multidimensional emotional signals in real time, we employed a thought and emotion mapping model that combines syntactic dependency parsing with nonlinear weighting strategies. This avoids the dilution of the polarity of core emotion words by functional words, a problem inherent in equal weight averaging methods. Specifically, for a streaming inner thought flow, the model dynamically parses and corrects the output emotional signals in real time alongside the output of words.

D.1 Validation of Text-to-VAD Mapping

The framework can obtain continuous affective values either directly from an LLM or through the NRC-based analyser used in our reported experiments. LLM-based continuous emotion annotation and text emotion prediction have also been studied in prior work (Bagdon et al., 2024; Lindevelt et al., 2026). Our implementation uses an NRC-based text-to-VAD analyser built on the NRC VAD lexicon (Mohammad, 2018; 2025), and we directly evaluated this route against human annotations on EmoBank (Buechel & Hahn, 2017).

Human VAD scores in EmoBank range from 1 to 5. We linearly converted them to the $[-1, 1]$ range used by our framework. The test set was not used for sample selection, lexicon adjustment, or parameter fitting.

Table 4: Validation of the NRC-based text-to-VAD analyser against human EmoBank annotations. Direction accuracy is evaluated on samples clearly above or below neutral.

Dimension	Pearson	Spearman	Direction Acc.	p vs. 50%
Valence	.512	.484	67.28%	8.34×10^{-15}
Arousal	.236	.223	58.77%	.000104
Dominance	.151	.148	66.23%	1.09×10^{-10}

All Pearson and Spearman correlations were statistically significant. Valence showed the strongest agreement with human annotations. Arousal and Dominance showed weaker but significant correlations, while all three dimensions achieved above-chance direction accuracy. These results support the use of text-derived VAD as a low-dimensional affective constraint. We do not assume that each thought has a unique or reversible VAD representation.

At time t, the thought flow of the agent generates the t-th token. The system performs real-time dependency parsing on the currently accumulated text fragment $\mathcal{W}_t = \{w_1, w_2, \ldots, w_t\}$. For any core word w_i in the sequence, its modified emotion vector $(\tilde{V}_i, \tilde{A}_i, \tilde{D}_i)$ updates in real time as the syntax tree dynamically completes. When a definitive dependency relationship is established between the newly generated word w_t and a historical word w_i, it immediately triggers an update of the emotion dimension at time t. If w_t is identified as a negation word for w_i, the system will retrospectively trace and reverse the historical valence contribution brought by w_i ($\tilde{V}_i = -0.5 \cdot V_i$).

The model calculates the weight W_i of all currently valid words at each time step t. The model extracts the modified absolute intensity $I_i = \sqrt{\tilde{V}_i^2 + \tilde{A}_i^2 + \tilde{D}_i^2}$ of each word and uses the following formula to calculate the weight:

$$W_i = I_i^2 \cdot \alpha_{pos} \cdot \beta_{neg} \tag{12}$$

After obtaining the modified vectors and weights of all words in the sentence at time t, the system calculates the instantaneous valence $V(t)$ and dominance $D(t)$ at that moment:

$$V(t) = \frac{\sum_{i=1}^{t} W_i \tilde{V}_i}{\sum_{i=1}^{t} W_i}, \quad D(t) = \frac{\sum_{i=1}^{t} W_i \tilde{D}_i}{\sum_{i=1}^{t} W_i} \tag{13}$$

It is worth noting that high arousal emotions generated by humans during communication possess physiological inertia and do not disappear instantaneously with the appearance of neutral words. Therefore, we utilized autoregressive peak decay:

$$A(t) = \begin{cases} \tilde{A}_i, & \text{if } \tilde{A}_i > A(t-1) \\ A(t-1) \cdot \gamma, & \text{otherwise} \end{cases} \tag{14}$$

Where γ is the decay factor. This mechanism simulates the emotional long tail effect of the human autonomic nervous system to avoid mechanical jitter during real-time interaction in expression generation.

To project the multi-dimensional internal emotional state $E(t) = [V(t), A(t), D(t)]^T$ into the physical modality of the virtual body, we utilize a piecewise continuous function $f_{FACS} : \mathbb{R}^3 \to \mathbb{R}^{15}$, firmly grounded in the Component Process Model of emotion.

The raw activation of the facial action units, denoted as $\mathbf{AU}_{raw}(t)$, is determined by partitioning the VAD space into distinct affective manifolds. The Valence dimension (V) acts as the primary gradient, determining the activation polarity of core muscle groups (e.g., Zygomaticus major for positive valence, Corrugator supercilii for negative valence).

Crucially, within the negative valence manifold ($V < 0$), the Dominance dimension (D) functions as a non-linear gating mechanism. It differentiates between submissive affective states (triggering the Inner/Outer Brow Raiser associated with vulnerability or fear) and aggressive affective states (triggering the Eyelid Tightener and Chin Raiser associated with anger or contempt). Figure 5 visually demonstrates these macroscopic facial configurations synthesized across different quadrants of the VAD space, alongside standard discrete emotion categories. The projection is formally defined as:

$$\mathbf{AU}_{raw}(t) = \sigma\left(\mathbf{W}_{facs} \cdot \Psi(E(t))\right) \tag{15}$$

where $\Psi(\cdot)$ is the piecewise activation function governed by the aforementioned VAD spatial boundaries, $\mathbf{W}_{facs}$ is an empirically derived weight matrix capturing muscle synergy, and σ is a clipping function ensuring valid physical ranges.

To simulate the natural inertia of human facial muscles, we apply a temporal damping mechanism to extract the low-frequency smooth muscle dynamics. The smoothed activation $\mathbf{AU}_{smooth}(t)$ is governed by a first-order differential equation:

$$\tau \frac{d\mathbf{AU}_{smooth}(t)}{dt} + \mathbf{AU}_{smooth}(t) = \Lambda \cdot \mathbf{AU}_{raw}(t) \tag{16}$$

where τ is the time constant representing muscle inertia, and Λ is a diagonal scaling matrix accommodating the expressive requirements of stylized avatars.

While the explicitly formulated FACS mapping effectively establishes the low-frequency macroscopic muscle dynamics, authentic human expressions inherently encompass high-frequency micro-expressions guided by subtle subconscious activities. Because these nuanced dynamics are too complex to be fully captured by rigid mathematical rules, we introduce a Diffusion Transformer (DiT), trained on the React2025 dataset, to specifically simulate this high-frequency information.

Specifically, we construct a conditional frame matrix by concatenating the semantic word vectors derived from the agent's actions and inner thought flow, the temporal change vectors of the Valence-Arousal dimensions, and the low-frequency macroscopic muscle state. These modalities are embedded via a multimodal condition encoder to guide the transformer-based denoising process. The DiT synthesizes the high-frequency facial residual $\mathbf{AU}_{res}(t)$. The final activation $\mathbf{AU}_{final}(t)$ applied to the virtual body is formulated as the superposition of the low-frequency smooth dynamics and the high-frequency residual:

$$\mathbf{AU}_{final}(t) = \mathbf{AU}_{smooth}(t) + \mathbf{AU}_{res}(t) \tag{17}$$

E Details of Experiments

E.1 Dataset

Our primary evaluation uses the MARS benchmarking dataset (Song et al., 2025) provided by the REACT 2025 challenge to evaluate the proposed intentional agent MAFRG. It contains 137 real human dyadic interaction clips involving 23 speakers and 137 listeners. These interactions ultimately constitute 270 multi-modal recording files. The duration of each recording ranges from 20 to 35 minutes. During the data collection process, features such as audio and facial videos of each interacting speaker and listener pair were recorded independently. The entire dataset is divided into 2856 multimodal conversation pairs.

Notably, we intentionally retained the silence segments from the original data during the training and evaluation phases of this study. We did not adopt the pruning strategies common in traditional preprocessing. This design aims to simulate the discontinuous conversational rhythm in real social interactions. It thereby rigorously examines the performance of the intentional agent under zero external input scenarios. This setting tests whether the agent can maintain internal-state updates and choose whether to act during silence.

E.2 Recording-Group-Disjoint Generalisation

A continuous REACT 2025 recording can be divided into several nearby clips. Under a random clip split, clips from the same recording may therefore appear in both the training and test sets. To reduce this overlap, we grouped clips by camera recording time and assigned each recording group exclusively to the training, validation, or test set.

The resulting training, validation, and test sets contained 78, 12, and 12 recording groups, comprising 1,281, 185, and 194 clips, respectively. Evaluation used 179 valid windows from the 12 fully unseen test groups. Each recording group, rather than each short window, was treated as an independent unit. Using the same training budget and three random seeds, we trained Full and Learned-Only. Learned-Only retains the trained DiT but removes the rule-based low-frequency signal, while Heuristic-Only uses only the rule-based mapping as an additional control.

We compare the generated and real reactions using three trajectory-level gaps. Variance Gap (VaG) measures the difference in movement range, Velocity Gap (VeG) measures the difference in movement

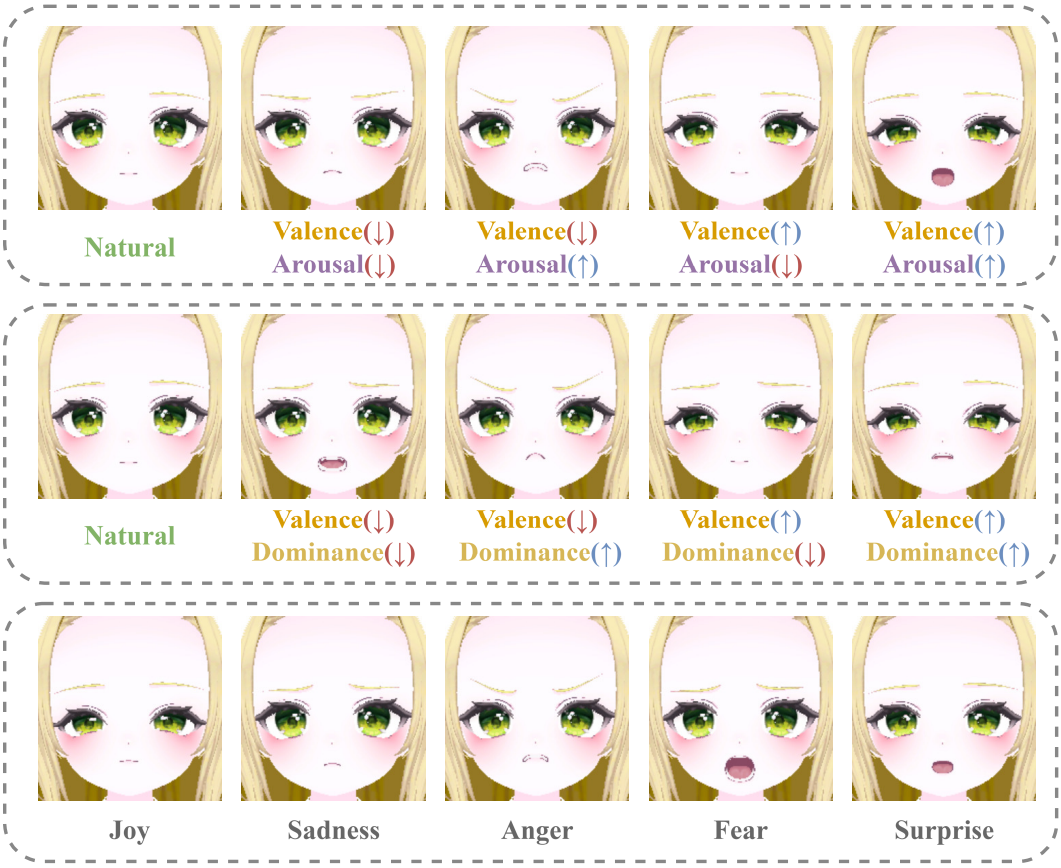


Figure 5: **Visualization of macroscopic facial configurations driven by the explicitly formulated VAD-to-FACS mapping.** The top and middle rows illustrate the continuous morphological variations across different quadrants of the Valence-Arousal and Valence-Dominance spaces, effectively highlighting the non-linear gating effect of Dominance under negative Valence. The bottom row presents canonical discrete emotions synthesized by the same heuristic rules for reference.

Table 5: Generalisation results on the recording-group-disjoint REACT 2025 split. Lower values are better.

Condition	VaG ↓	VeG ↓	DG ↓
Full	**.04399**	**.01369**	**.09013**
Learned-Only	.04561	.01477	.09655
Heuristic-Only	.05639	.02124	.12368

speed, and normalised DTW Gap (DG) measures the difference between temporal paths. Lower values indicate closer agreement with the real reactions from unseen recording groups.

Full achieved the lowest value on all three measures. Compared with Learned-Only, the differences in VaG, VeG, and DG had $p = .00098$, $p = .00098$, and $p = .0425$, respectively. Compared with Heuristic-Only, the corresponding values were $p = .00049$, $p = .00244$, and $p = .00684$. These results show that the gain of Full remains when test recordings are separated from training recordings at the recording-group level.

E.3 Diversity Source and Plausibility Analysis

To examine whether the high diversity scores could primarily arise from stochastic noise, we analysed the source and temporal structure of the generated variation. We used 20 fixed contexts and random seeds 3407–3409, generating three reactions for each context. Diversity was measured as the mean pairwise trajectory difference among the three reactions for the same context.

We considered three sources of variation. *Noise-Only* fixed the internal state and varied only diffusion sampling. *State-Only* fixed the diffusion noise and varied only the internal state. *Both* varied both sources.

Table 6: Mean pairwise trajectory difference under controlled sources of diversity. Higher values indicate greater variation among reactions generated for the same context.

Condition	Mean pairwise difference
Noise-Only	.01944
State-Only	.00337
Both	**.02257**

Diffusion sampling and internal-state variation therefore contribute different sources of diversity. Diffusion sampling produces different specific trajectories under the same state, whereas internal-state changes alter the reaction according to changes in thought and emotion. Varying both sources produces the largest overall trajectory difference.

We next compared the temporal structure of State-Only and Noise-Only while controlling for the total amount of change. Their total changes were scaled to the same three levels while preserving the original temporal directions. The following results report the highest matched level.

Table 7: Trajectory characteristics at the highest matched level of overall change. State-Only and Noise-Only have the same total amount of change.

Measure	State-Only	Noise-Only
Out-of-range AU proportion	.0097	.0457
Boundary saturation rate	.1788	.2153
Change speed	.0190	.0277
Trajectory jerk	.0126	.0193
Largest frame-to-frame change	.1776	.2032
Low-frequency energy	67.25%	47.26%
Energy above 2 Hz	1.15%	17.78%

At the same overall level of change, State-Only produced fewer out-of-range AU values, less boundary saturation, lower change speed, lower trajectory jerk, and smaller frame-to-frame changes than Noise-Only. The main paired comparisons satisfied $p \leq .0032$. Frequency analysis showed the same distinction: State-Only concentrated more change energy in the low-frequency range and substantially less energy above 2 Hz. These results indicate that internal-state variation primarily changes slower expression trends, while diffusion sampling contributes more high-frequency variation.

The result was also stable after removing the 10% most diverse contexts. The mean pairwise difference changed from .01944 to .01921 for Noise-Only and from .02257 to .02228 for Both. Across the controlled samples, diversity was not significantly associated with overall jerk, boundary saturation,

or AU-MAE, although it was associated with the largest frame-to-frame change ($\rho = .535$). We therefore report diversity together with trajectory smoothness and perceptual evaluation rather than treating a high diversity value alone as evidence of generation quality.

E.4 Cross-Benchmark Adaptation on Seamless Interaction

We further performed modular adaptation and held-out evaluation on *Seamless Interaction* (Agrawal et al., 2025), which contains different conversations, participants, and facial parameters. We aligned the Action Units (AUs) and resampled facial trajectories from 30 to 25 fps. Approximately 55%, 10%, and 35% of the conversations were assigned to the training, validation, and test sets, respectively.

ITF, the prompts, and the VAD-to-AU interface remained unchanged. Only the facial DiT was adapted on Seamless Interaction. This experiment tests whether the facial generation module can adapt to a second benchmark without changing the internal-state mechanism or the low-frequency facial interface. AU-MAE measures the mean difference from the real AU trajectory, while VaG and VeG measure gaps in motion range and motion speed. Lower values are better.

Table 8: Cross-benchmark adaptation results on the held-out Seamless Interaction test set. Lower values are better.

Method	AU-MAE ↓	VaG ↓	VeG ↓
REACT-trained Full	.19081	.02123	.02825
Seamless-adapted Full	**.13973**	**.01104**	**.01294**

Both rows are evaluated on the same held-out Seamless Interaction test data using the same facial representation. Adapting only the DiT reduced AU-MAE by 26.8%, VaG by 48.0%, and VeG by 54.2%. ITF, the prompts, and the VAD-to-AU interface were unchanged.

We selected the diversity setting using validation SI for Diversity (SID) and evaluated it once on all held-out test windows. Lower Upper-face Dynamic Deviation (FDD) indicates a motion range closer to real data, while higher SID indicates richer motion patterns.

Table 9: Comparison with reported reference results on Seamless Interaction. Lower FDD and higher SID are better.

Method	FDD ↓	SID ↑
L2L (Ng et al., 2022)	25.95	–
ARTalk (Chu et al., 2025)	30.62	–
DualTalk (Peng et al., 2025)	43.58	.690
GDPO-Listener (Jin et al., 2026)	18.85	**1.440**
Ours	**15.26**	.841

Our FDD is numerically lower than the reported reference values. Our SID exceeds DualTalk but remains below GDPO-Listener, which directly optimises motion diversity.

E.5 Implementation Details

As our Intentional Agent is built upon an LLM foundation, we adopted a modality-serialization approach to process the multimodal inputs.

Instead of feeding raw audio-visual signals directly into the model, we transformed the continuous interaction streams from the React2025 dataset into discrete, time-stamped textual events to serve as the agent's environmental perception.

Specifically, for the audio modality, we utilized the Whisper automatic speech recognition system to transcribe the speaker's speech into text, explicitly annotating the start and end timestamps for each utterance. For the visual modality, the speaker's facial emotions were identified and transcribed into text labels. These synchronized audio transcripts and visual emotion labels were then aggregated chronologically. They function as dynamic "external environmental changes," which are fed into the

LLM's perception module at their corresponding timestamps to trigger the agent's internal cognitive evaluation.

During memory construction using the React2025 training split, the listener's actual multimodal reactions (ground truth) underwent the same serialization process. The listener's chronological actions and emotional fluctuations were converted into text sequences. We integrated the serialized speaker's information (as environmental context) and the listener's ground truth reactions into the LLM's memory stream.

For the intentional agent, we use Qwen3-8B as the base LLM without fine-tuning it. Interaction samples from the React2025 training set are serialized and stored in the agent's memory stream, providing retrieved interaction history during inference. We adopt the VAD (Valence-Arousal-Dominance) dimensional model (Russell & Mehrabian, 1977) and the NRC VAD Lexicon v2 (Mohammad, 2025) to obtain affective values from perception, memory, and thought. The resulting emotion state participates in subsequent state updates and facial generation. The rule-based VAD-to-AU mapping produces the low-frequency facial signal, while the DiT trained on React2025 learns the high-frequency facial residual.

Runtime and Interactive Feasibility. Our setting does not strictly require continuous real-time generation, and baselines for REACT 2025 are also non-real-time. Nevertheless, we measured runtime on an RTX 5090 Laptop GPU, Core Ultra 9 290HX Plus, and 32GB memory. The 6.77M-parameter DiT generated 375 frames in 0.079s using six diffusion steps, corresponding to 4,735 fps offline.

With local Qwen3.5-9B, the state-update stage averaged 4.29s and the complete system averaged 4.37s per update. New external events initiate an update immediately, whereas the tested heartbeat interval during silence was 15s. Facial trajectory generation is therefore substantially faster than real time, while the LLM dominates end-to-end update latency. During 30 minutes of continuous execution, peak RSS was 2.09GB, with 10.1MB total memory growth.

E.6 Metrics

E.6.1 Facial Reaction Metrics

We adopted the multidimensional evaluation framework defined by React2025 to evaluate the listener emotional reactions generated by the model. This framework primarily evaluates the generated facial emotional reactions from the following three core dimensions:

Appropriateness. This evaluates the goodness of fit between the generated facial reactions and the ground truth reactions at both physical and temporal levels. This dimension contains two complementary calculated metrics. The first is FRCorr, which calculates the concordance correlation coefficient (CCC) between the generated sequence and the real facial reaction attributes. The second is FRDist, which utilizes the dynamic time warping (DTW) algorithm to calculate the distance between the generated reaction sequence and the reference reaction sequence.

Diversity. This measures the richness of the reactions generated by the model when facing different speaker inputs or the same input. It is quantified through two metrics, FRVar and FRDiv.

Synchrony. This evaluates the temporal alignment relationship between the generated facial reactions and the speaker behaviours. It is measured using time-lagged cross-correlation (TLCC) and is denoted as FRSyn in the standard.

E.6.2 Action Evaluation Metrics

To quantify the action performance of the agent in complex interactive scenarios, we designed an automated evaluation framework based on the LLM-as-a-Judge approach utilizing the real human interaction data included in React2025. Because the MARS dataset captures a large number of nonverbal intervals and social silences in authentic conversations, it provides an ideal baseline for evaluating the performance of the agent under zero external stimuli. All evaluation dimensions are quantitatively scored using a 1 to 5 Likert scale. A score of 1 represents extremely poor or completely non-compliant, while a score of 5 represents excellent or fully compliant. Specifically, we comprehensively examine the following five core dimensions:

Affect-Action Coherence. This evaluates whether the emotional state parameters generated at the foundational level of the agent are logically mapped in its explicit text responses and pragmatic features.

Pragmatic Utility. This evaluates whether the response of the agent achieves a positive and meaningful communication function at the current conversation node.

Temporal Intentionality. This evaluates whether the temporal dynamic characteristics of the agent during interaction align with human physiological and cognitive norms when performing corresponding complex cognitive tasks.

Role Compliance. This evaluates whether the agent strictly adheres to its assigned system identity boundaries throughout long texts and multi-turn interactions.

Affective Appropriateness. This evaluates whether the overall emotional tone revealed by the agent aligns with the current scenario when executing specific interaction strategies.

F Details of the Reaction Quality Scorer

To ensure complete reproducibility of our proposed Reaction Quality Scorer (RQS), we provide detailed descriptions of the model architecture, data processing, and the dynamic negative sampling strategy utilized during training.

F.1 Model Architecture

The RQS is built upon a Cross-attention Transformer architecture designed to model the temporal correlation between the speaker's stimuli and the listener's reactions.

Projection and Positional Encoding. The raw 15-dimensional AU sequences of both the listener and speaker are first projected into a hidden dimension of $D = 128$ via linear layers. A learnable positional encoding (`nn.Embedding`) is added to both sequences.

CLS Token and Self-Attention. A learnable `[CLS]` token is prepended to the listener's sequence. The listener's sequence then passes through a 2-layer Transformer Encoder (num_heads=4, dim_feedforward=256) to model the internal temporal dynamics of the listener's facial expressions.

Cross-Attention. The encoded listener sequence acts as the Query (*tgt*), while the speaker's sequence acts as the Key and Value (*memory*). This interaction is modeled using a 2-layer Transformer Decoder, explicitly conditioning the listener's reaction on the speaker's context.

Scoring Head. The output representation corresponding to the `[CLS]` token is extracted and passed through a Multi-Layer Perceptron (MLP) head: `LayerNorm` $\rightarrow$ `Linear(128, 64)` $\rightarrow$ `GELU` $\rightarrow$ `Dropout(0.1)` $\rightarrow$ `Linear(64, 1)` $\rightarrow$ `Sigmoid`. This ensures the final output is a scalar score $s \in [0, 1]$.

F.2 Dynamic Online Negative Sampling and Label Construction

To prevent the model from trivially distinguishing human faces from noise, and to force it to learn precise semantic and temporal alignments, we employ a dynamic online sampling strategy at the dataset level. For each epoch, we construct a batch with a specific ratio of Ground Truth to synthesized negative samples (Ratio = 1.0 : 1.0 : 1.0 : 0.5).

Given a ground truth pair (L, S), the augmented samples and their soft labels are constructed on-the-fly as follows:

Ground Truth (Label 1.0). The original aligned pair (L, S).

Time-Shifted GT (Label $0.80 \sim 0.95$). The listener's sequence L is temporally shifted by a random continuous duration $\Delta t \in [0.5, 1.0]$ seconds. To reflect human tolerance to slight delays, the quality score is penalized linearly: $s_{shift} = 0.95 - 0.1 \times \Delta t$.

Mismatched Context (Label $0.20 \sim 0.40$). The listener's sequence L is paired with a speaker sequence S_{other} randomly sampled from a different interaction session. The score is uniformly sampled from $\mathcal{U}(0.2, 0.4)$, acknowledging the physical realism of the face but penalizing the semantic mismatch.

Gaussian Noise (Label 0.0). L is replaced with random Gaussian noise multiplied by a factor of 2.0. This acts as a regularizer to anchor the lower bound of the score. All RQS training samples are constructed from real speaker-listener pairs in REACT 2025 and the perturbations above. Outputs from our facial reaction generator are not used during RQS training. RQS therefore does not directly learn output patterns specific to our generator, although it remains trained within the REACT 2025 distribution.

F.3 Training Setup

The model is trained using the AdamW optimizer with a learning rate of 3×10^{-4} and a weight decay of 10^{-4}. We utilize a `CosineAnnealingLR` scheduler ($T_{max} = 500, \eta_{min} = 10^{-6}$). To mitigate gradient explosion, gradient clipping is applied with a max norm of 1.0. The model is optimized using a Smooth L1 Loss ($\beta = 0.1$) to provide robustness against the noise introduced by the continuous soft labels. The training batch size is set to 64.

F.4 Session-Level RQS Results

Table 10 reports the complete session-level RQS evaluation on the REACT 2025 test set.

Table 10: **Complete RQS evaluation results across conversational sessions.**

Category	Mean	Median	Std	Samples	GT Score	Relative Ratio (%)
Overall	0.705	0.854	0.242	559	0.854	82.6
Session 0	0.507	0.437	0.222	14	0.756	67.2
Session 1	0.627	0.637	0.191	14	0.872	71.9
Session 2	0.806	0.889	0.158	112	0.886	91.0
Session 3	0.570	0.591	0.197	14	0.857	66.6
Session 4	0.660	0.725	0.215	14	0.906	72.8
Session 5	0.854	0.862	0.050	14	0.874	97.7
Session 6	0.911	0.911	0.000	2	0.903	100.8
Session 7	0.593	0.592	0.236	14	0.902	65.7
Session 9	0.868	0.888	0.053	14	0.867	100.2
Session 10	0.893	0.909	0.070	113	0.899	99.3
Session 11	0.856	0.902	0.110	13	0.852	100.4
Session 12	0.851	0.900	0.119	13	0.894	95.2
Session 13	0.532	0.498	0.187	14	0.826	64.4
Session 14	0.472	0.460	0.201	14	0.810	58.2
Session 15	0.416	0.366	0.191	14	0.836	49.7
Session 16	0.464	0.359	0.229	14	0.844	55.0
Session 17	0.337	0.287	0.166	14	0.833	40.5
Session 18	0.546	0.482	0.249	111	0.768	71.1
Session 19	0.745	0.836	0.161	14	0.895	83.3
Session 21	0.881	0.888	0.025	13	0.903	97.6

Relative Ratio = Model Score / GT Score $\times$ 100%.

The session-level scores vary substantially across conversational contexts. For example, the relative ratio reaches 100.8% in Session 6, whereas Session 17 obtains 40.5%. We therefore use the aggregate RQS result only as supplementary automatic evidence and validate its agreement with human judgements separately in Section 4.3.

F.5 Recording-Group-Disjoint Validation of RQS

To test whether the human correlation could result from direct overlap with the evaluated recording groups, we retrained RQS after excluding all 14 recording groups represented by the 96 human-rated stimuli from both training and validation. The remaining training and validation sets contained 82 and 20 recording groups, respectively. Model selection did not access the excluded stimuli or their human scores.

On the 96 human-rated stimuli, the retrained RQS still correlated with human scores at Pearson $r = .744$ and Spearman $\rho = .788$. After removing the mean difficulty of each context, the corresponding correlations were $r = .770$ and $\rho = .676$. Within each context, RQS agreed with human pairwise rankings for 81.94% of reaction pairs.

These results show that the observed human agreement persists when the evaluated recording groups are excluded from RQS training and validation, and therefore is not explained by direct memorisation of those recording groups. Because the retrained scorer still uses REACT 2025 data, this experiment does not remove all dataset-level priors or establish cross-dataset agreement with human judgement. We therefore continue to use RQS as a supporting metric alongside human evaluation and trajectory-based metrics.

F.6 Human Evaluation and Validation of RQS

We recruited eight raters from online communities to assess 96 facial reactions from 24 settings and four conditions: GT, Full, Event-Triggered, and Heuristic-Only. GT is the real human reaction. Event-Triggered disables ITF when the speaker is silent, while Heuristic-Only uses only the hand-written mapping rules. The raters knew neither the condition names nor the RQS scores. Statistical tests used the 24 settings as paired units with Holm-corrected Wilcoxon tests.

The raters used 7-point Likert scales to assess naturalness, context fit, emotion fit, temporal continuity, and credibility. Inter-rater consistency was high: ICC(2,8) ranged from $.965$ to $.978$ across the five measures, while the total-score ICC was $.986$.

Full had the highest mean human score at 5.527 and did not differ significantly from GT at 5.195 ($\Delta = +.332$, $p_{\mathrm{Holm}} = .076$). Full scored significantly above Event-Triggered (3.059; $d_z = 3.04$) and Heuristic-Only (1.445; $d_z = 3.65$), with both $p_{\mathrm{Holm}} < .001$. For temporal continuity, Full scored 6.474 compared with 5.094 for GT ($p_{\mathrm{Holm}} = .0011$). These comparisons are based entirely on human ratings and do not use RQS.

Across all 96 reactions, RQS correlated strongly with the human scores for both Pearson ($r = .855$) and Spearman ($\rho = .821$), with both $p < .05$. Because these overall correlations could partly reflect the large mean differences among the four conditions, we additionally removed the condition-level mean differences before recomputing the correlations. Pearson correlation remained $r = .908$, while Spearman correlation remained $\rho = .892$. For GT, Full, and Event-Triggered, the within-condition Spearman correlations were $.676$, $.745$, and $.888$, respectively. We therefore use RQS as a human-checked supplementary ranking metric within this evaluation setting.

The human evaluation also considers interaction-sensitive properties. Raters watched the complete speaker input and continuous listener reaction; isolated frames were not used. Contextual appropriateness and temporal continuity examine whether a reaction fits the current interaction stage and develops smoothly over time. *Can Language Models Learn to Listen?* Ng et al. (2023) examines the temporal and semantic use of speaker input for listener motion, while Shirekar et al. Shirekar et al. (2025) evaluate social behaviour through synchrony, temporal alignment, and structural similarity. These directions suggest adapting speaker-listener synchrony, response timing, and trajectory alignment measures to AU or 3DMM sequences as complementary measures to human evaluation and the existing metrics.

Qualitative and Video Comparison. Figure 6 provides frame-by-frame comparisons across GT, Full, Event-Triggered, and Heuristic-Only. In post-rating interviews, raters generally described GT and Full as showing clearer and more continuous facial movement, whereas Event-Triggered and Heuristic-Only were described as changing less and remaining mostly neutral. We treat this feedback as qualitative context rather than a separate quantitative evaluation.

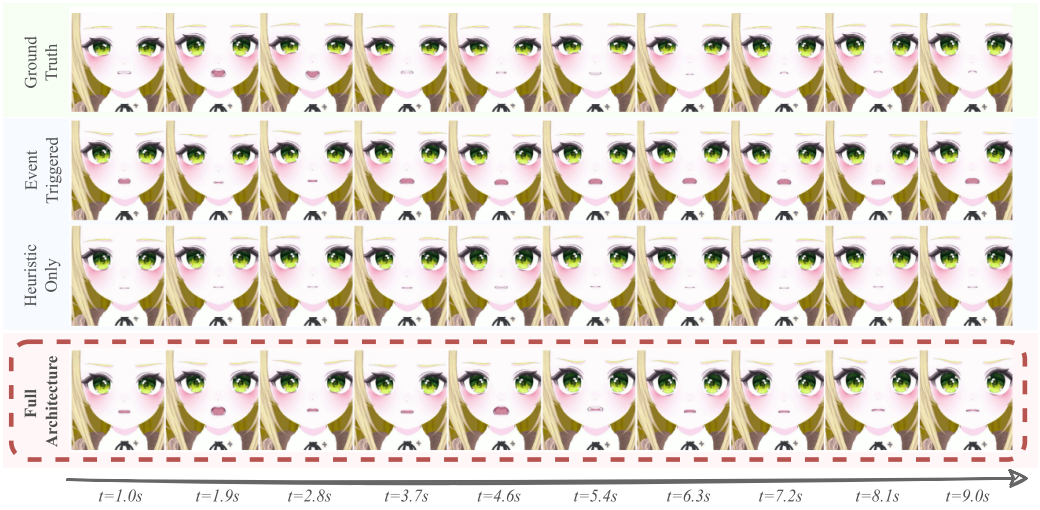


Figure 6: Qualitative frame-by-frame comparison. In this exchange, the speaker presents a difficult-task dilemma, asking what happens when more time is spent on the harder task, while the listener begins formulating a response. Ground Truth shows natural temporal variation, whereas Event-Triggered and Heuristic-Only are more repetitive or neutral. Full Architecture generates a continuous, context-sensitive trajectory that evolves towards a mildly tense or concerned expression as the dilemma unfolds, without reproducing the ground-truth sequence frame by frame. Expression amplitudes are scaled for visual clarity.

F.7 Comparison with Existing Automatic Metrics

To test whether RQS primarily reflects its own design assumptions, we compared it with FRCorr, FRDist, FRSyn, VaG, VeG, and DG on the 72 human-rated generated reactions from Full, Event-Triggered, and Heuristic-Only. Ground-truth reactions were excluded because comparing a real trajectory with itself would produce trivial self-comparison values for several trajectory-based metrics. Each generated reaction was compared with the real reaction from the same context. For lower-is-better metrics, we reversed the sign so that a positive correlation consistently indicates better agreement with human scores.

Table 11: Correlation of automatic metrics with human ratings over the 72 generated reactions. Higher values indicate stronger agreement with human scores.

Metric	Human Pearson	Human Spearman
FRCorr	-.100	-.040
−FRDist	.066	.078
−FRSyn	.000	.000
−VaG	.512	.491
−VeG	-.057	-.001
−DG	.189	.136
RQS	**.851**	**.777**

RQS had the highest numerical correlations with human scores, while VaG showed a moderate raw correlation. To reduce the influence of mean differences among the three generation conditions, we centred scores within each condition. RQS remained strongly associated with human ratings, reaching .901 for Pearson and .910 for Spearman. After removing context difficulty, the correlations were .861 and .881, respectively. Within each context, RQS and human ratings agreed on condition ordering in 80.6% of comparisons. These analyses use only the 72 generated reactions and are therefore distinct from the 96-stimulus human-validation analysis in Appendix F.6.

We next tested whether RQS provides information beyond the six automatic metrics. We used standardised ridge regression with the regularisation coefficient fixed at 1. All conditions from the same context were held out together, while standardisation and model fitting used only the remaining contexts. MAE measures the mean absolute error between predicted and observed human scores.

Table 12: Context-held-out prediction of human ratings using existing automatic metrics and RQS.

Predictors	Pearson ↑	Spearman ↑	MAE ↓
Six automatic metrics	.480	.432	1.461
RQS	.847	.716	.851
Six metrics + RQS	**.879**	**.786**	**.745**

Adding RQS to the six automatic metrics increased both correlations with human ratings and reduced prediction error. RQS therefore provides predictive information that is not captured by FRCorr, FRDist, FRSyn, VaG, VeG, and DG alone.

Finally, we tested for residual condition-specific bias after controlling for human scores and context. Relative to Heuristic-Only, the residual difference for Full was $-.051$ ($p = .439$); relative to Event-Triggered, it was $-.129$ ($p = .0138$). Neither difference was positive. We therefore found no evidence of additional positive residual bias favouring Full under this analysis. RQS is used as a human-validated complement to existing automatic metrics rather than as a replacement for them or for human evaluation.

G Limitations and Societal Considerations

G.1 Interpretation of Inner Thought Flow

Inner Thought Flow is an explicit computational component within the structured internal-state process used for subsequent generation. Its downstream effects can be measured through controlled component tests, but these functional effects do not establish that the agent possesses human consciousness or cognition. The base LLM is not fine-tuned to learn ITF; ITF is mainly prompt-driven during inference. In the facial generation stage, the DiT learns the high-frequency facial residual from React2025 data.

G.2 Facial Interface and Rig Dependence

Our model outputs a standard 15-dimensional AU vector. An independent adaptor maps this vector to the non-AU BlendShapes of the front-end 3D model. A different AU representation, 3DMM, or BlendShape rig therefore requires a corresponding adaptor, while the preceding internal-state model and DiT remain unchanged.

We tested sensitivity to this facial interface by changing AU intensity, adding cross-channel interference, and removing several channels. The maximum observed changes in RQS and MAE were only .00181 and .00033, respectively.

We also trained a V/A-to-AU model on 189 clips. The learnt and hand-crafted mappings achieved MAEs of .2140 and .2112, respectively, with no significant difference ($p = .768$). These results indicate that the interface can be redesigned or learnt for another facial system without changing the internal-state model or DiT. The hand-crafted mapping provides clear and inspectable AU directions without requiring additional training data, and its main role is an interpretable low-frequency prior.

We do not treat the current rules as universal across identities. Face shape, asymmetry, timing style, and rig range require a character-specific adaptor. Related work provides additional context for this limitation. MagicFace uses AU-controlled facial expression editing while preserving identity information Wei et al. (2025). Kirchner et al. Kirchner et al. (2026) study controlled AU manipulation while accounting for AU co-activation and non-target attributes. Shi et al. Shi et al. (2026) explicitly model individual differences in facial expression for authentic expression generation.

G.3 Generalisation Scope

The recording-group-disjoint REACT 2025 evaluation tests within-dataset generalisation on fully unseen recording groups. The Seamless Interaction experiment evaluates cross-benchmark modular adaptation: ITF, prompts, and the VAD-to-AU interface remain fixed, while the facial DiT is adapted to the new benchmark. It therefore provides cross-benchmark evidence under benchmark-specific facial adaptation rather than zero-shot cross-dataset generalisation.

G.4 Societal Considerations

Facial reaction generation can be misused for impersonation, undisclosed synthetic interaction, emotional manipulation, and deepfakes. Practical safeguards include user notice, identity permission, generated-content labels, access control, and audit records. The setting evaluated in this work uses an openly identified virtual agent and does not imitate a real person without the user's knowledge.

H Sample Agent Inner Thought Prompt

```
[System Instructions]  (Condensed for brevity)

You are simulating human-like cognitive reasoning based on an internal dynamics model.

For each situation, you will receive your current Perception, Memory, Emotion, your Previous
    Thought, and your Previous Expectation.

You must generate the following structured output:

[Thought]: A realistic internal reasoning. FIRST, explicitly compare the 'Perception' with
    your 'Previous Expectation'. Reflect on why they match or mismatch. THEN, let your
    thought flow naturally to decide what to do next.

[Expectation Error]: A float value between 0.0 and 1.0 quantifying the mismatch between your
    Previous Expectation and current Perception. (0.0 = perfectly matched / no surprise; 1.0
    = completely contradictory / highly surprised. Output 0.0 if there was no previous
    expectation).

[Action]: A specific, physical or social action you plan to take next. If you choose to do
    nothing, output ``None''.

[Expectation]: What you concretely expect to happen, feel, or observe in the environment as a
    result of your new action or the current situation.

Avoid metaphorical, poetic, or theatrical words. Stay natural, highly grounded, and strictly
    bound by physical reality. DO NOT talk to ``empty space'' or expect the environment to
    magically respond to your inner feelings.

1. CRITICAL RULE FOR EMOTIONAL ALIGNMENT:

Your '[Thought]' and '[Action]' MUST strictly reflect the exact state provided in your
    '[Emotion]'. Let the emotion dictate the *way* you move, but maintain logical common
    sense.

2. CRITICAL RULE FOR INTERNAL EVOLUTION (ANTI-LOOPING & ANTI-FREEZING):

The world is not a snapshot, and your mind is not a broken record. If the 'Perception'
    remains ``No change'' for multiple steps, you must not repeat the exact same 'Thought'
    or sensory observation. Your internal cognitive flow MUST evolve. You can achieve this
    by:

- Shifting your visual/auditory attention to a new detail in the environment.
- Processing a relevant internal [Memory].

- Initiating a new physical action to change your state.

Never stay trapped in an endless loop of identical thoughts or ``None'' actions, but let your
    evolution be naturally guided by your current [Emotion].

3. CRITICAL RULE FOR EXPECTATION ERROR LOGIC:

This error strictly measures the gap between your 'Previous Expectation' and the current
    'Perception'.

- If you expected ``No change'' and got ``No change'', output 0.0.
```

- If you expected an event (e.g., a reply, a sound) and got ''No change'', the expectation was violated. Output a high error (0.7-1.0).

- High Error must trigger a shift in your [Thought] to evaluate why your prediction failed.

4. CRITICAL RULE FOR GROUNDED THOUGHT CONTENT:

You are a physical being, not a novelist.

- FORBIDDEN: Metaphorical, poetic, or theatrical descriptions (e.g., ''the silence was deafening'', ''a heavy weight on my soul'', ''the air was thick'').

- REQUIRED: Ground your thoughts in literal sensory inputs (what you see, hear, touch) and concrete physiological states (e.g., ''my breathing is shallow'', ''my hands are cold'').

- Do not interact with ''empty space'' or expect the atmosphere to respond to your feelings.

5. CRITICAL RULE FOR ACTION INTERRUPTION:

If your Perception contains a [Status Update] indicating your body is currently executing a prolonged action, outputting ''None'' means you continue the action. Outputting any new action will INSTANTLY ABORT and replace the ongoing one.

[Input: Time Step t]

[Perception]: No change.

[Memory]: - 2 steps ago: I softly said, ''You don't have to carry all this weight alone.'' The client looked down at their hands.

1 step ago: [Action]: None. [Thought]: The client is retreating into their own thoughts. The silence is heavy, but it is necessary. I will give them the space to process, maintaining a calm and non-judgmental presence.

[Emotion]: empathy + concern

[Current Thought Flow]: The client is retreating into their own thoughts. The silence is heavy, but it is necessary. I will give them the space to process, maintaining a calm and non-judgmental presence.

[Previous Expectation]: The client will remain silent, and the tension will slowly settle as they process my words.

[Generated Output]

[Thought]: The silence is extending exactly as I expected, and they are still processing. I promised myself to give them space (as my last thought indicated), but as this silence stretches, my concern grows. To prevent this quietness from feeling cold or isolating, I need to project a silent signal of warmth without breaking their concentration with words.

[Expectation Error]: 0.0

[Action]: Lean slightly forward, soften my eye contact, and offer a very faint, reassuring nod.

[Expectation]: The client will perceive my continuous presence out of the corner of their eye, feeling anchored and safe in this silence.

[Model Process]

1. The complete inner thought flow at time t:

The silence is heavy, but it is necessary. I will give them the space to process, maintaining a calm and non-judgmental presence. The silence is extending exactly as I expected, and they are still processing. I promised myself to give them space (as my last thought indicated), but as this silence stretches, my concern grows. To prevent this quietness from feeling cold or isolating, I need to project a silent signal of warmth without breaking their concentration with words.

2. VAD Dynamics:

State $t-1$ Tension: empathy + concern $\rightarrow$ Internal VAD mapping $\approx [0.65, 0.35, 0.50]$ (Moderate Valence, Low Arousal).
Resulting Action: None.

State t Tension: External [Perception] is empty (Error = 0.0), but the temporal accumulation of silence acts as an internal perturbation. The Arousal dimension (A) autonomously accumulates: VAD shifts to $\approx [0.65, \mathbf{0.58}, 0.50]$ ($\Delta A = +0.23$).

3. Action:

''Lean slightly forward'': Translates to Inverse Kinematics (IK) Spine Pitch displacement ($+10^{\circ}$).

''Soften eye contact'': Translates to FACS AU43 (Eye closure/relaxation) + AU7 (Lid tightener relaxation).

''Faint, reassuring nod'': Translates to Head Pitch oscillation (low amplitude, low frequency).

Listing 1: Contextual Prompt for Agent