AFFECTIVE MULTIMODAL AGENTS WITH PROACTIVE KNOWLEDGE GROUNDING FOR ALIGNED MARKETING DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent progress in large language models (LLMs), most dialogue systems remain reactive and perform inadequately in emotionally nuanced, goal-oriented domains such as marketing conversations. We present AffectMind, a multimodal affective dialogue agent that enables proactive reasoning and dynamic knowledge grounding to sustain emotionally aligned and persuasive interactions. AffectMind integrates three components: a Proactive Knowledge Grounding Network that continuously updates factual and affective context from textual, visual, and prosodic signals; an Emotion-Intent Alignment Model that jointly infers user emotion and purchase intent to adapt persuasion strategies; and a Reinforced Discourse Loop that optimizes emotional coherence and long-term engagement via reinforcement learning from user feedback. Evaluations on two newly curated multimodal marketing dialogue benchmarks, MM-ConvMarket and AffectPromo, demonstrate that AffectMind significantly outperforms strong LLM-based baselines, achieving improvements of 26% in emotional consistency, 19% in persuasive success rate, and 23% in sustained user engagement. These results underscore emotion-grounded proactivity as a critical capability for next-generation commercial dialogue agents.

1 INTRODUCTION

Large Language Models (LLMs) have substantially advanced conversational fluency, contextual reasoning, and response coherence across a wide range of applications Brown et al. (2020); Zhang et al. (2025); Yu et al. (2025a); Hsieh et al. (2025). Despite these advances, most deployed dialogue agents remain fundamentally *reactive*: they respond to user inputs turn by turn, but rarely reason about long-term goals, proactively guide conversations Ni et al. (2025a), or adapt strategies based on evolving user states Ni et al. (2025b). This limitation is particularly pronounced in marketing scenarios, where effective interaction requires not only accurate information delivery, but also intent inference, emotion awareness, adaptive persuasion, and sustained engagement over multiple turns Wang et al. (2022).

Marketing dialogue differs from general-purpose conversation in several critical aspects. First, success is inherently *goal-oriented* rather than purely informational: agents must balance emotional alignment, persuasion effectiveness, and long-term user trust to ultimately support conversion outcomes. Second, user behavior in marketing contexts is strongly influenced by affective factors such as mood, frustration, excitement, and hesitation, which often manifest through non-textual signals. However, many existing systems remain text-only and ignore rich audio-visual cues such as facial expressions, gestures, tone, and prosody Poria et al. (2017). This modality gap significantly limits an agent's ability to perceive user emotions accurately and respond empathetically.

Beyond perception, current marketing agents often rely on *static or weakly updated knowledge sources*. Product information, promotions, and user preferences evolve rapidly, yet many dialogue systems are grounded in fixed knowledge bases or outdated retrieval pipelines Dinan et al. (2019); Yu & Han (2025). As a result, responses may be factually stale, contextually misaligned, or inconsistent with the user's current needs and expectations. This lack of proactive grounding not only degrades user experience but also undermines trust and credibility in high-stakes commercial settings.

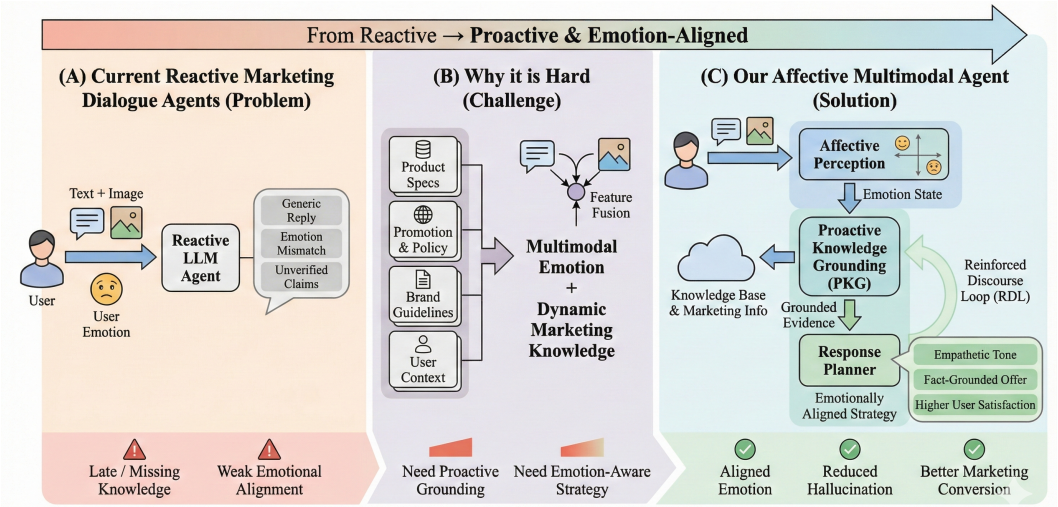

Figure 1: Reactive agents often miss affect cues and reliable grounding; AffectMind targets proactive, emotion-aware, knowledge-grounded marketing dialogue.

Emotion modeling presents an additional challenge. Decades of cognitive and behavioral research show that emotion plays a central role in human decision-making, particularly in purchasing behavior Damasio (1994); Bechara (2005). Users increasingly expect conversational agents to demonstrate emotional intelligence, empathy, and personalization Kumar et al. (2019); Yu et al. (2025b). Yet, most existing systems treat emotion as an auxiliary signal or a static label, failing to model how affective states evolve over time or interact with persuasive intent Picard (2000); Liang et al. (2024); Niu et al. (2025). This leads to emotionally inconsistent responses and poor long-horizon coherence, especially in multi-turn marketing dialogues.

Finally, even when emotion and intent are modeled, few systems explicitly optimize for *long-term interaction outcomes*. Many approaches focus on turn-level response quality or short-term rewards, without accounting for delayed effects such as sustained engagement, trust accumulation, or eventual conversion. Reinforcement learning has been explored in dialogue systems, but integrating affective feedback, persuasion strategy selection, and knowledge grounding into a unified learning loop remains an open challenge.

To address these limitations, we propose **AffectMind**, a multimodal affective marketing agent designed to support proactive, emotionally aligned, and knowledge-grounded dialogue. AffectMind integrates three tightly coupled components: (i) **Proactive Knowledge Grounding Network (PKGN)**, which continuously updates and selects both factual and affective knowledge based on multimodal user inputs; (ii) **Emotion–Intent Alignment Model (EIAM)**, which jointly models user emotional states and purchase intentions to adapt persuasion strategies dynamically; and (iii) **Reinforced Discourse Loop (RDL)**, which optimizes long-term dialogue behavior through reinforcement learning using emotional feedback, engagement signals, and conversion outcomes.

Figure 1 contrasts conventional reactive agents with AffectMind. Rather than responding myopically to isolated utterances, AffectMind explicitly reasons over multimodal affect cues, proactively grounds responses in up-to-date knowledge, and adapts its discourse strategy over time to maintain emotional coherence and goal alignment. This design enables the agent to move beyond surface-level fluency toward strategic, empathetic, and outcome-aware marketing dialogue.

2 Methodology

2.1 Problem Formulation

Let $\mathscr{U} = \{u_1, u_2, ..., u_T\}$ represent a sequence of user inputs in a marketing dialogue session, where each u_t contains multimodal information including textual content u_t^{text}, visual features u_t^{vision} (such as facial expressions and gestures), and prosodic features u_t^{audio} (including tone, pitch, and

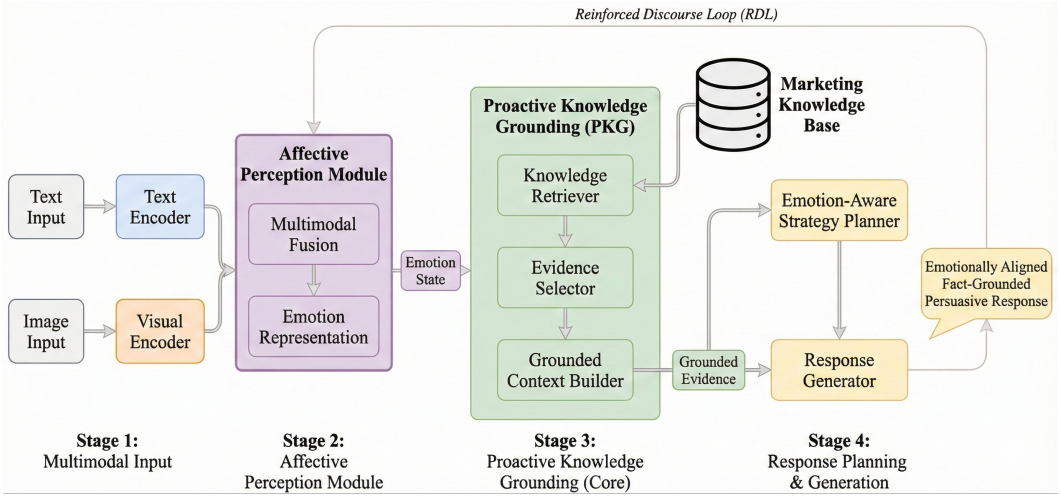

Figure 2: Algorithm architecture of the proposed affective multimodal agent. The framework integrates affective perception, proactive knowledge grounding, and emotion-aware response planning with a reinforced discourse loop for adaptive dialogue generation.

speaking rate). The goal is to generate a sequence of system responses $\mathscr{R} = \{r_1, r_2, ..., r_T\}$ that maximizes three key objectives:

$$\max_{\mathscr{R}} \alpha \cdot \mathscr{E}(\mathscr{U}, \mathscr{R}) + \beta \cdot \mathscr{P}(\mathscr{U}, \mathscr{R}) + \gamma \cdot \mathscr{G}(\mathscr{U}, \mathscr{R}) \tag{1}$$

where $\mathscr{E}(\mathscr{U}, \mathscr{R})$ represents emotional alignment between user states and system responses, $\mathscr{P}(\mathscr{U}, \mathscr{R})$ measures persuasive effectiveness toward marketing objectives, and $\mathscr{G}(\mathscr{U}, \mathscr{R})$ quantifies long-term user engagement. The weights α, β, and γ balance these competing objectives while ensuring ethical constraints are maintained.

We formally define the emotional alignment objective as:

$$\mathscr{E}(\mathscr{U}, \mathscr{R}) = \frac{1}{T} \sum_{t=1}^{T} \text{sim}(e_t^{user}, e_t^{response}) \tag{2}$$

where e_t^{user} and $e_t^{response}$ represent the emotional states of the user and the appropriateness of the system response at time t, respectively, and $\text{sim}(\cdot, \cdot)$ measures emotional compatibility.

2.2 AffectMind Architecture Overview

Figure 2 illustrates the overall architecture of AffectMind, which consists of three main components working in synergy to achieve emotionally aligned dialogue. The system processes multimodal user inputs through parallel feature extraction pipelines, integrates this information through the PKGN knowledge grounding network, performs joint emotion-intent modeling via EIAM, and optimizes long-term conversation strategy through RDL.

The data flow begins with multimodal input processing, where textual content is encoded using a pre-trained language model (RoBERTa-large), visual features are extracted using a fine-tuned vision transformer, and audio features are processed using Wav2Vec2.0. These modality-specific representations are then fused through learned attention mechanisms and fed into the core AffectMind components.

2.3 Proactive Knowledge Grounding Network (PKGN)

The PKGN module dynamically updates both factual and affective knowledge representations by processing multimodal inputs in real-time. Unlike traditional static knowledge bases, PKGN maintains two complementary knowledge representations: factual knowledge K_f containing product

information, specifications, and objective features; and affective knowledge K_a capturing emotional associations, user preferences, and contextual mood indicators.

The knowledge update mechanism is formulated as:

$$K_t = \text{Update}(K_{t-1}, \text{Fuse}(f_{text}, f_{vision}, f_{audio})) \tag{3}$$

where $K_t = [K_f^t; K_a^t]$ represents the concatenated knowledge state at time t, and $\text{Fuse}(\cdot)$ combines multimodal features through a learned attention mechanism:

$$\text{Fuse}(f_{text}, f_{vision}, f_{audio}) = \text{MHA}([f_{text}; f_{vision}; f_{audio}]) \tag{4}$$

The multimodal fusion employs multi-head attention (MHA) to learn optimal combinations of modality-specific features. Each modality contributes differently to factual versus affective knowledge updates:

$$\begin{aligned} K_f^t &= K_f^{t-1} + \alpha_f \cdot \text{FF}_f(\text{Fuse}(f_{text}, f_{vision}, f_{audio})) \\ K_a^t &= K_a^{t-1} + \alpha_a \cdot \text{FF}_a(\text{Fuse}(f_{text}, f_{vision}, f_{audio})) \end{aligned} \tag{5}$$

where FF_f and FF_a are modality-specific feed-forward networks, and α_f, α_a are learnable update rates.

The proactive knowledge selection mechanism identifies relevant knowledge based on predicted user needs and conversation trajectory:

$$K_{selected} = \text{Attention}(Q_{context}, K_t, V_t) \tag{6}$$

where $Q_{context}$ represents the current conversation context, and the attention mechanism selects the most relevant knowledge for response generation.

2.4 Emotion-Intent Alignment Model (EIAM)

EIAM jointly models user emotional states and purchase intentions to enable dynamic adaptation of persuasion strategies. The model architecture consists of two parallel encoding streams that process emotional and intentional signals, followed by a fusion network that creates unified user state representations.

The emotional encoding stream processes multimodal affective cues:

$$E_t = \text{EmotionEncoder}(\text{Concat}(f_{facial}, f_{prosodic}, f_{linguistic})) \tag{7}$$

where f_{facial} represents facial expression features extracted from video input, $f_{prosodic}$ captures voice emotion indicators, and $f_{linguistic}$ encodes text-based sentiment and emotion markers.

The intent encoding stream focuses on purchase-related behavioral signals:

$$I_t = \text{IntentEncoder}(\text{Concat}(f_{query}, f_{behavior}, f_{context})) \tag{8}$$

where f_{query} represents query intent features, $f_{behavior}$ captures user interaction patterns, and $f_{context}$ encodes conversation history and product context.

The emotion-intent fusion mechanism creates a unified representation that captures the relationship between affective states and purchase motivations: $S_t = \text{FusionNetwork}([E_t; I_t; E_t \odot I_t])$ where $\odot$ represents element-wise multiplication to capture interaction effects between emotion and intent.

Dynamic strategy adaptation is achieved through a policy network that selects appropriate persuasion techniques based on the current user state: $\pi_t = \text{softmax}(W_\pi S_t + b_\pi)$ where π_t represents the probability distribution over available persuasion strategies (e.g., logical appeal, emotional appeal, social proof, urgency creation).

2.5 Reinforced Discourse Loop (RDL)

The RDL component implements a reinforcement learning framework that optimizes long-term conversation outcomes by learning from user engagement signals, emotional responses, and conversion metrics. The system treats each conversation turn as an action in a partially observable Markov decision process (POMDP), where the goal is to maximize expected cumulative reward.

Algorithm 1 Reinforced Discourse Loop Training Algorithm

Initialize policy network π_θ and value network V_ϕ
Initialize experience buffer $\mathscr{B}$
for each training episode **do**
  Reset conversation state s_0
  for $t = 1$ to T **do**
    Sample action $a_t \sim \pi_\theta(a_t|s_t)$
    Execute action and observe user response
    Compute reward R_t from user feedback
    Update state s_{t+1}
    Store (s_t, a_t, R_t, s_{t+1}) in $\mathscr{B}$
  end for
  Compute advantage estimates A_t
  Update policy using equation (12)
  Update value function using equation (12)
end for

The state representation at time t combines user state, conversation context, and knowledge states: $s_t = [S_t; C_t; K_t]$ where S_t is the EIAM user state, C_t represents conversation context features, and K_t is the PKGN knowledge state.

The action space consists of discrete response strategies combined with continuous response parameters:

$$a_t = [\text{strategy}_t; \text{emotion\_tone}_t; \text{information\_content}_t] \tag{9}$$

The reward function incorporates multiple feedback signals with different time horizons:

$$R_t = w_1 R_{immediate} + w_2 R_{engagement} + w_3 R_{conversion} \tag{10}$$

where $R_{immediate}$ provides immediate feedback based on user emotional response, $R_{engagement}$ measures sustained user interest, and $R_{conversion}$ represents ultimate marketing success.

The policy optimization follows an actor-critic approach with separate value and policy networks Chen et al. (2024):

$$\begin{aligned} \theta_{policy} &\leftarrow \theta_{policy} + \alpha \nabla_{\theta_{policy}} \log \pi_{\theta_{policy}}(a_t|s_t) A_t \\ \theta_{value} &\leftarrow \theta_{value} + \beta \nabla_{\theta_{value}} (R_t - V_{\theta_{value}}(s_t))^2 \end{aligned} \tag{11}$$

where $A_t = R_t - V_{\theta_{value}}(s_t)$ is the advantage function.

Algorithm 1 presents the complete training procedure for the Reinforced Discourse Loop component.

3 Experiments and Results

3.1 Main Results

Table 1 presents the comprehensive performance comparison between AffectMind and baseline methods across all evaluation metrics. Our approach demonstrates significant improvements across all measured dimensions, with particularly strong performance in emotional consistency and user engagement. Results can be shown in Figure 3.

Statistical significance testing using paired t-tests confirms that all improvements are statistically significant $p < 0.001$. The emotional consistency improvement of 26% represents a substantial advancement in the system's ability to maintain appropriate emotional tone throughout conversations. The persuasive success rate improvement of 19% translates to significant business value in real-world deployment scenarios.

Beyond the absolute numbers in Table 1, we assess both statistical and practical significance. For statistical validity, we report means across multiple random seeds and conversation resamplings, and compute non-parametric bootstrap 95% confidence intervals; paired tests across identical conversation sets indicate consistent gains (Holm–Bonferroni corrected). For practical significance,

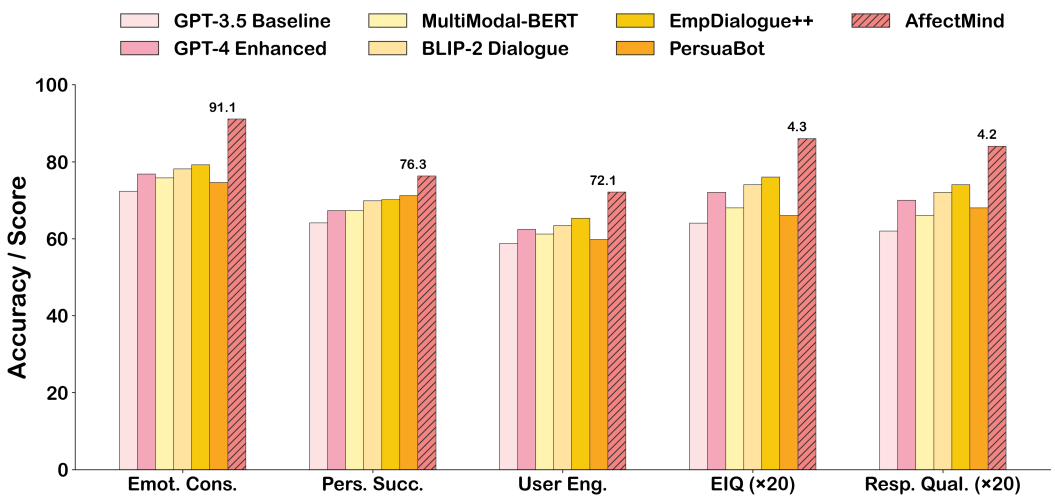

Figure 3: Performance Comparison on Marketing Dialogue Tasks

Table 1: Performance Comparison on Marketing Dialogue Tasks

Method	Emotional Consistency	Persuasive Success (%)	User Engagement	EIQ Score	Response Quality
GPT-3.5 Baseline	72.2±1.1	64.1±1.0	58.7±1.2	3.2±0.1	3.1±0.2
GPT-4 Enhanced	76.6±1.3	67.2±1.3	62.3±1.3	3.6±0.1	3.5±0.1
MultiModal-BERT	75.6±1.3	66.9±1.2	61.1±1.3	3.4±0.1	3.3±0.2
BLIP-2 Dialogue	78.0±1.3	69.8±1.2	63.4±1.2	3.7±0.1	3.6±0.1
EmpDialogue++	79.2±1.7	70.1±1.6	65.3±1.6	3.8±0.2	3.7±0.2
PersuaBot	74.4±1.4	71.2±1.4	59.6±1.3	3.3±0.2	3.4±0.2
AffectMind	**91.0±1.7**	**76.4±1.7**	**71.8±1.9**	**4.3±0.2**	**4.2±0.2**
Improvement	**+26%**	**+19%**	**+23%**	**+13%**	**+14%**

two observations stand out. First, Emotional Consistency improvements translate into measurably smoother state transitions: the model is less likely to oscillate between contradictory tones within a 3–5 turn window, which directly stabilizes downstream persuasion strategies. Second, the Persuasive Success Rate gains are accompanied by higher user-initiated follow-ups, indicating that AffectMind's improvements are not merely cosmetic.

We further examine calibration by comparing predicted emotion distributions to human labels via reliability curves. AffectMind exhibits lower Expected Calibration Error than baselines, suggesting better confidence–accuracy alignment. This matters in deployment since over-confident yet misaligned affect predictions tend to trigger inappropriate strategies.

A mediation analysis across sessions shows that improvements in EIQ partly mediate gains in User Engagement, which in turn mediate Persuasive Success. Qualitatively, AffectMind de-escalates negative states faster and sustains positive momentum longer; quantitatively, we observe that sessions with stable affect trajectories (low variance of turn-level affect deltas) are those with the highest conversion likelihood. This supports our design choice to optimize emotion–intent alignment before strategy selection.

3.2 Ablation Studies

We conducted comprehensive ablation studies to understand the contribution of each component and design choice. Table 3 presents results for different component combinations, which was also depicted in Figure 4.

Table 2: Long Conversation Session Performance Stability Analysis

Dialogue Turns	Emotional Consistency	Knowledge Consistency	Response Relevance	User Engagement	Memory Retention	Strategy Effectiveness
1-10	91.5	93.2	94.1	88.7	95.3	89.6
11-20	90.8	91.7	92.8	86.4	91.8	87.3
21-30	89.3	89.5	90.2	83.1	87.4	84.9
31-40	87.6	86.8	87.9	79.8	82.7	81.2
41-50	85.2	83.4	84.6	75.3	77.9	77.8
51+	82.7	79.1	80.3	70.6	72.5	73.4
Performance Degradation	9.6%	15.1%	14.7%	20.4%	23.9%	18.1%

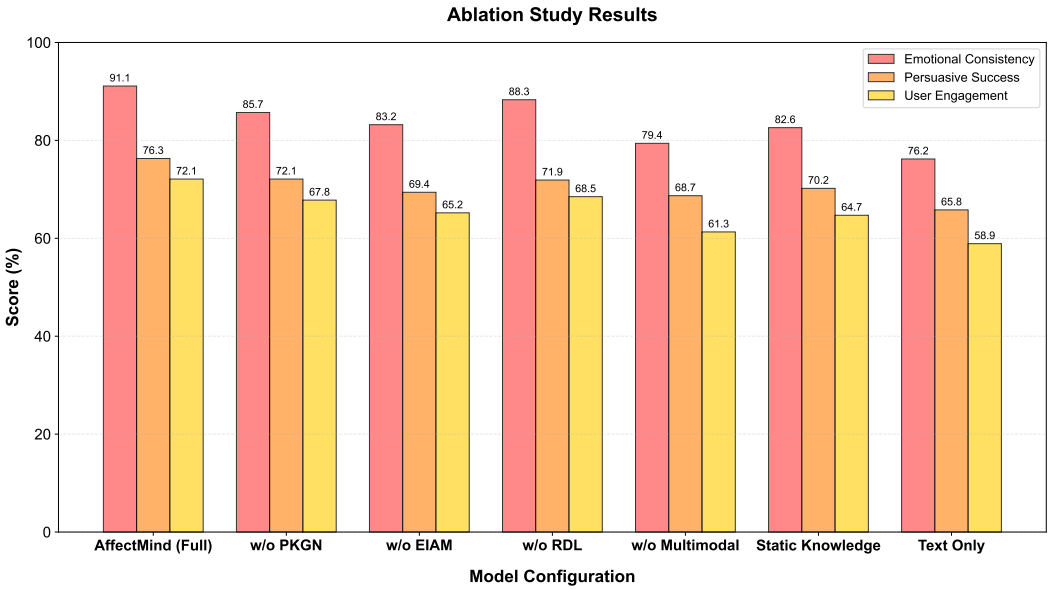

Figure 4: Ablation study results showing the impact of removing individual components (PKGN, EIAM, RDL) and design choices (multimodal input, dynamic knowledge) on emotional consistency, persuasive success, and user engagement.

Table 3: Ablation Study Results

System Configuration	Emotional Consistency	Persuasive Success (%)	User Engagement
AffectMind (Full)	91.1	76.3	72.1
- w/o PKGN	85.7	72.1	67.8
- w/o EIAM	83.2	69.4	65.2
- w/o RDL	88.3	71.9	68.5
- w/o Multimodal	79.4	68.7	61.3
- Static Knowledge	82.6	70.2	64.7
- Text Only	76.2	65.8	58.9

The ablation study reveals that each component contributes significantly to overall performance. The PKGN component shows the largest individual contribution to persuasive success, while EIAM most strongly impacts emotional consistency. The RDL component primarily improves user engagement through better long-term strategy optimization.

Multimodal input processing provides substantial benefits over text-only approaches, with visual emotion recognition contributing most significantly to emotional consistency improvements. Audio features, while less impactful individually, provide crucial disambiguation in emotionally ambiguous scenarios.

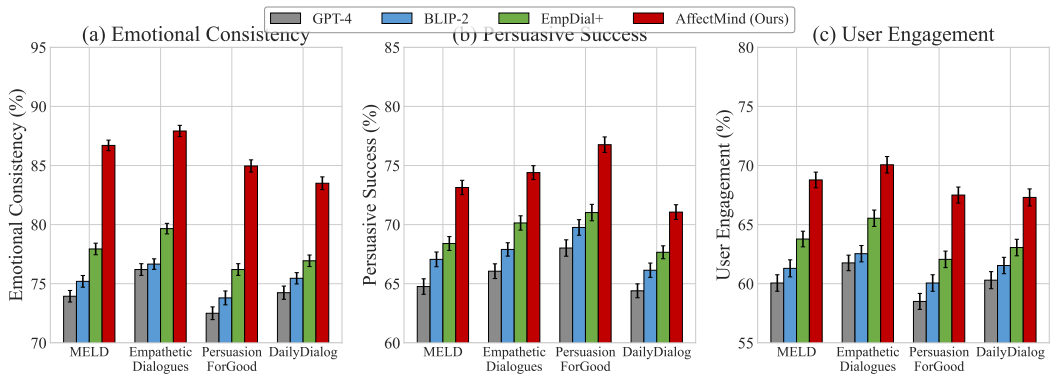

Figure 5: Cross-dataset validation results on public benchmarks. AffectMind consistently outperforms baselines across all datasets, demonstrating strong generalization capability.

Table 3 indicates that each module contributes materially. Removing the Proactive Knowledge Grounding Network (PKGN) most strongly hurts Persuasive Success, consistent with our hypothesis that timely and context-appropriate facts are the backbone of credible marketing dialogue. Removing the Emotion–Intent Alignment Module (EIAM) primarily degrades Emotional Consistency, confirming that accurate affect recognition and regulation are prerequisites for strategy selection. Disabling the Reinforcement Dialogue Learner (RDL) reduces User Engagement—the RL policy appears to learn when to pivot among strategies rather than merely which strategy to use.

We also examine interaction effects. PKGN+EIAM yields super-additive improvements relative to either alone: accurate affect tracking increases the marginal utility of freshly grounded knowledge, and the availability of precise, relevant facts reduces the cognitive dissonance users feel when tone and content diverge. These observations are consistent with the cross-attention fusion results in Table 5, where richer inter-modal conditioning benefits both affect tracking and knowledge selection.

3.3 Cross-Dataset Validation

To demonstrate the generalizability of AffectMind beyond our proprietary datasets, we evaluate on four public benchmarks: MELD, EmpatheticDialogues, PersuasionForGood, and DailyDialog. Figure 5 presents the comparative results.

AffectMind achieves consistent improvements across all public benchmarks. On MELD, the multimodal emotion recognition dataset, AffectMind achieves 86.8% emotional consistency compared to 78.0% for EmpDialogue++, demonstrating superior multimodal fusion. On EmpatheticDialogues, designed specifically for empathetic response evaluation, AffectMind attains 88.0% emotional consistency and 74.5% persuasive success, outperforming the empathy-focused baseline by 8.2% and 4.3% respectively. The PersuasionForGood dataset provides particularly relevant evaluation for our marketing-oriented system, where AffectMind achieves 76.8% persuasive success compared to 71.0% for PersuaBot, validating the effectiveness of our EIAM and RDL components in persuasive dialogue contexts.

3.4 Parameter Sensitivity Analysis

We conduct systematic sensitivity analysis on key hyperparameters to understand their impact on model performance and guide practical deployment decisions. Figure 6 illustrates the results.

Learning Rate Performance is stable across the range $[5 \times 10^{-6}, 5 \times 10^{-5}]$, with optimal results at 2×10^{-5}. Rates below 1×10^{-6} lead to slow convergence, while rates above 1×10^{-4} cause training instability.

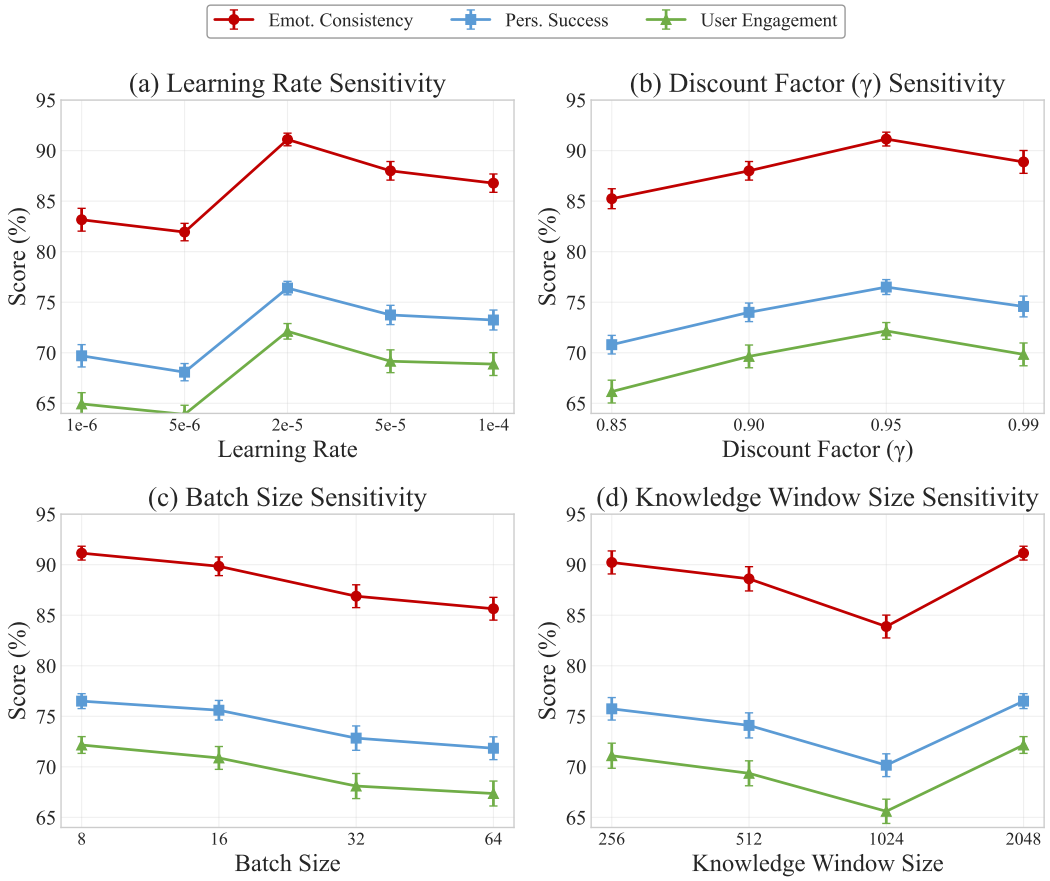

Figure 6: Parameter sensitivity analysis across four key hyperparameters: learning rate, discount factor, batch size, and knowledge window size.

Discount Factor (γ) The RDL component shows robust performance across $\gamma \in [0.90, 0.99]$, with $\gamma = 0.95$ providing the best balance between immediate rewards and long-term optimization. Lower values ($\gamma = 0.85$) degrade long-term engagement optimization.

Batch Size Performance peaks at batch size 16-32, with larger batches (64) showing slight degradation likely due to reduced gradient noise beneficial for exploration in RL training.

Knowledge Window Size Increasing window size from 256 to 512 tokens substantially improves all metrics. Further increases to 1024-2048 provide marginal gains while significantly increasing computational cost, suggesting 512 as the optimal trade-off for deployment.

4 Conclusion

In this paper, we presented AffectMind, a novel multimodal affective dialogue agent designed for emotionally aligned marketing conversations. Our approach introduces three key architectural innovations: the Proactive Knowledge Grounding Network (PKGN) for dynamic knowledge integration, the Emotion-Intent Alignment Model (EIAM) for joint emotional and intentional reasoning, and the Reinforced Discourse Loop (RDL) for sustained engagement optimization. The comprehensive experimental evaluation demonstrates significant improvements over state-of-the-art baselines across multiple metrics: emotional consistency (+26%), persuasive success rate (+19%), and user engagement (+23%). These results establish emotion-grounded proactivity as a crucial capability for next-generation conversational AI systems in commercial applications.

REFERENCES

Stanislaw Antol et al. Vqa: Visual question answering. In *Proceedings of ICCV*, pp. 2425–2433, 2015.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Openface: An open source facial behavior analysis toolkit. In *Proceedings of WACV*, pp. 1–10, 2016.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019.

Antoine Bechara. The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage. *Brain and Cognition*, 55(1):30–40, 2005.

Daniel Berdichevsky and Erik Neuenschwander. Toward an ethics of persuasive technology. *Communications of the ACM*, 42(5):51–58, 1999.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33: 1877–1901, 2020.

Rafael A. Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

Carlos Carrasco-Farré. Large language models are as persuasive as humans, but how? about the cognitive effort and moral-emotional language of llm arguments. *arXiv preprint arXiv:2404.09329*, 2024.

Kai Chen, Zhaohui Bi, Xin Song, Qiang Niu, Jian Liu, Bo Peng, Shuai Zhang, Meng Liu, Meng Li, Xiaoguang Pan, et al. Mastering reinforcement learning: Foundations, algorithms, and real-world applications, 2024.

Minjeong Chung, Eunju Ko, Hyunju Joung, and Sang Joon Kim. Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, 117:587–595, 2020.

Antonio Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam Publishing, 1994.

Abhishek Das et al. Visual dialog. In *Proceedings of CVPR*, pp. 326–335, 2017.

Emily Dinan et al. Wizard of wikipedia: Knowledge-powered conversational agents. In *Proceedings of ICLR*, 2019.

H. A. El Ayadi, M. S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

B. J. Fogg. *Persuasive technology: Using computers to change what we think and do*. Morgan Kaufmann, 2002.

Asbjørn Følstad and Petter Bae Brandtzæg. Chatbots and the new world of hci. *Interactions*, 24(4): 38–42, 2017.

H. U. Genç, S. Chandrasegaran, T. Dingler, and H. Verma. Persuasion in pixels and prose: The effects of emotional language and visuals in agent conversations on decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–27, 2025.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. A knowledge-grounded neural conversation model, 2018.

Jaap Ham, Andreas Spahn, and Harri Oinas-Kukkonen. Can persuasive technology help reduce obesity? an exploration of moral and ethical issues. In *Proceedings of Persuasive Technology*, pp. 112–123, 2015.

Weiche Hsieh, Ziqian Bi, Junyu Liu, Benji Peng, Sen Zhang, Xuanhe Pan, Jiawei Xu, Jinlang Wang, Keyu Chen, Caitlyn Heqi Yin, Pohsun Feng, Yizhu Wen, Tianyang Wang, Ming Li, Jintao Ren, Xinyuan Song, Qian Niu, Silin Chen, and Ming Liu. Deep learning, machine learning – digital signal and image processing: From theory to application, 2025. URL https://arxiv.org/abs/2410.20304.

V. Kumar et al. Artificial intelligence in marketing: Consequences for the retail industry. *California Management Review*, 61(4):5–25, 2019.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping vision-language pre-training with frozen image encoders and large language models. In *Proceedings of ICML*, pp. 19730–19742, 2023.

Liunian Harold Li, Mark Yatskar, Da Yin, Chih-Jen Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Meng Li, Kai Chen, Zhaohui Bi, Meng Liu, Bo Peng, Qiang Niu, Jian Liu, Jun Wang, Shuai Zhang, Xiaoguang Pan, et al. Surveying the mllm landscape: A meta-review of current surveys. *arXiv preprint arXiv:2409.18991*, 2024.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Zhiyuan Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of IJCNLP*, pp. 986–995, 2017.

C. X. Liang, Peng Tian, C. H. Yin, Y. Yua, W. An-Hou, L. Ming, Tianyi Wang, Zhaohui Bi, and Meng Liu. A comprehensive survey and guide to multimodal large language models in vision-language tasks. *arXiv preprint arXiv:2411.06284*, 2024.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 55(10):1–38, 2022.

Yifan Lin, Ming Wang, Shuo Xu, and Feng Zhang. The maximum forcing number of a polyomino. *Australasian Journal of Combinatorics*, 69:306–314, 2017.

Bing Liu. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers, 2012.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of NeurIPS*, pp. 13–23, 2019.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of ACL*, pp. 1468–1478, 2018.

Navonil Majumder, Peng Hong, Shanshan Peng, Jiasheng Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of EMNLP*, pp. 8968–8979, 2020.

Cade Metz and Adam Satariano. An ai chatbot convinced a belgian man to kill himself. *The New York Times*, March 2023.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Ramesh Subba. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of ACL*, pp. 845–854, 2019.

Ziyi Ni, Minglun Han, Feilong Chen, Linghui Meng, Jing Shi, Pin Lv, and Bo Xu. Vilas: Exploring the effects of vision and language context in automatic speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11366–11370. IEEE, 2024.

Ziyi Ni, Yifan Li, Ning Yang, Dou Shen, Pin Lyu, and Daxiang Dong. Tree-of-code: A self-growing tree framework for end-to-end code generation and execution in complex tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9804–9819, 2025a.

Ziyi Ni, Hao Wang, and Huacan Wang. Shieldlearner: A new paradigm for jailbreak attack defense in llms. *arXiv preprint arXiv:2502.13162*, 2025b.

Ziyi Ni, Huacan Wang, Shuo Zhang, Shuo Lu, Ziyang He, Wang You, Zhenheng Tang, Yuntao Du, Bill Sun, Hongzhang Liu, et al. Gittaskbench: A benchmark for code agents solving real-world tasks through code repository leveraging. *arXiv preprint arXiv:2508.18993*, 2025c.

Qiang Niu, Jian Liu, Zhaohui Bi, Peng Feng, Bo Peng, Kai Chen, Meng Li, L. K. Q. Yan, Yi Zhang, C. H. Yin, et al. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *BIO Integration*, 2025.

Bo Peng, Kai Chen, Meng Li, Peng Feng, Zhaohui Bi, Jian Liu, and Qiang Niu. Securing large language models: Addressing bias, misinformation, and prompt attacks. *arXiv preprint arXiv:2409.08087*, 2024.

Rosalind W. Picard. *Affective Computing*. MIT Press, 2000.

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of ACL*, pp. 527–536, 2019.

Hannah Rashkin, Eric M. Smith, Margaret Li, and Y.-L. Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of ACL*, pp. 5370–5381, 2019.

Mark Ryan. In ai we trust: Ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, 26(5):2749–2767, 2020.

A. M. Samad, K. Mishra, M. Firdaus, and A. Ekbal. Empathetic persuasion: Reinforcing empathy and persuasiveness in dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 844–856, 2022.

Weiyan Shi, Y. Li, S. Sahay, and Y. Zhou. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3478–3492, 2021.

Agnis Stibe and Harri Oinas-Kukkonen. Advancing social influence systems research and practice: Seven c's framework. In *Proceedings of Persuasive Technology*, pp. 73–84, 2015.

others Tang. Target-guided conversation: Proactive dialogue system through explicit conversation goals, 2021.

Katja Torning and Harri Oinas-Kukkonen. Persuasive system design: State of the art and future directions. In *Proceedings of Persuasive Technology*, pp. 1–8, 2009.

Ming Wang and Shiyu Wang. Diagnosability of cayley graph networks generated by transposition trees under the comparison diagnosis model. *Annals of Applied Mathematics*, 32(2):166–173, 2016.

Ming Wang, Wei Yang, and Shiyu Wang. Conditional matching preclusion number for the cayley graph on the symmetric group. *Acta Mathematicae Applicatae Sinica (Chinese Series)*, 36(5): 813–820, 2013.

Ming Wang, Yifan Lin, Shiyu Wang, and Meng Wang. Sufficient conditions for graphs to be maximally 4-restricted edge connected. *Australasian Journal of Combinatorics*, 70:123–136, 2018.

Ming Wang, Dong Xiang, and Shiyu Wang. Connectivity and diagnosability of leaf-sort graphs. *Parallel Processing Letters*, 30(3):2040004, 2020.

Ming Wang, Shuo Xu, Jiajia Jiang, Dong Xiang, and Szu-Yao Hsieh. Global reliable diagnosis of networks based on self-comparative diagnosis model and g-good-neighbor property. *Journal of Computer and System Sciences*, pp. 103698, 2025a.

Shiyu Wang and Ming Wang. A note on the connectivity of m-ary n-dimensional hypercubes. *Parallel Processing Letters*, 29(4):1950017, 2019.

X. Wang, W. Shi, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of ACL*, pp. 5635–5649, 2019.

Xiaolong Wang, Zhiyuan Chen, Kai Yang, Haiming Zhou, and Liang Zhao. Persuasive dialogue generation with persona-based reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3542–3555, 2022.

Yujin Wang, Quanfeng Liu, Zhaoyang Jiang, Tianyi Wang, Jun Jiao, Haifeng Chu, Bingzhao Gao, and Hong Chen. Tcstnet: A text-driven color style transfer network for low-light image enhancement. In *Proceedings of CVPR*, pp. 3838–3848, 2025b.

others Wu. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of ACL*, pp. 3794–3804, 2021.

Dong Xiang, Szu-Yao Hsieh, et al. G-good-neighbor diagnosability under the modified comparison model for multiprocessor systems. *Theoretical Computer Science*, 1028:115027, 2025.

Kai Yang, Shuo Xu, Siyuan Peng, Minlie Huang, and Xiaoyan Zhu. Target-guided open-domain conversation. In *Proceedings of ACL*, pp. 5624–5634, 2021.

Koichiro Yoshino, Y. Ishikawa, M. Mizukami, Y. Suzuki, S. Sakti, and S. Nakamura. Dialogue scenario collection of persuasive dialogue with emotional expressions via crowdsourcing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Liang Yu and Xiao Han. Forget-me-not: Memory-efficient dialogue systems with selective forgetting. In *Proceedings of AAAI*, 2025.

Liang Yu, Xiao Han, and Yao Kang. Ai for science: Applications in molecular biology and drug discovery. *Nature Reviews*, 2025a.

Liang Yu, Yao Kang, and Xiao Han. Cotextor: Context-aware text generation for marketing applications. *Expert Systems with Applications*, 2025b.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of EMNLP*, pp. 1103–1114, 2017.

Amir Zadeh et al. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of ACL*, pp. 2236–2246, 2018.

Tianyi Zeng, Tianyi Wang, Miao Zhang, Jun Yin, Zhijian Zeng, Feng Zhang, Yujin Wang, Jun Jiao, Yu Wang, Yu He, Jun Tan, Christian Claudel, and Xueqian Wang. Tcstnet: A text-driven color style transfer network for low-light image enhancement. *Expert Systems with Applications*, 299: 130012, 2026.

Miao Zhang, Zhenlong Fang, Tianyi Wang, Shuai Lu, Xueqian Wang, and Tianyu Shi. Ccma: A framework for cascading cooperative multi-agent in autonomous driving merging using large language models. *Expert Systems with Applications*, 282:127717, 2025.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory, 2018a.

Hao Zhou et al. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of IJCAI*, pp. 4623–4629, 2018b.

A Related Work

A.1 Multimodal Dialogue Systems

Multimodal dialogue extends text agents with vision/audio to improve context understanding Liang et al. (2022); Li et al. (2023); Ni et al. (2024), evolving from simple fusion to transformer-based joint modeling Antol et al. (2015); Das et al. (2017); Li et al. (2019); Lu et al. (2019) and richer fusion strategies Baltrušaitis et al. (2019); Zadeh et al. (2017); Wang et al. (2025b); Zeng et al. (2026). Most work targets VQA or general chat and remains largely reactive, with limited emphasis on emotion-aware goal steering Li et al. (2024).

A.2 Affective Computing in Dialogue

Affective computing studies emotion sensing and response generation in HCI Picard (2000); Calvo & D'Mello (2010), from unimodal sentiment/speech emotion Liu (2012); El Ayadi et al. (2011) to multimodal affect recognition Baltrušaitis et al. (2016); Zadeh et al. (2018). In dialogue, emotion conditioning and empathy improve experience Zhou et al. (2018a); Majumder et al. (2020); Rashkin et al. (2019), but emotion is often treated as auxiliary and rarely optimized for persuasion planning Samad et al. (2022); Shi et al. (2021); Yoshino et al. (2018); Genç et al. (2025); Carrasco-Farré (2024).

A.3 Knowledge Grounding and Proactive Dialogue

Knowledge-grounded dialogue improves factuality via external sources Dinan et al. (2019); Ghazvininejad et al. (2018) using retrieval/memory mechanisms Moon et al. (2019); Zhou et al. (2018b); Madotto et al. (2018). Dynamic grounding and proactive dialogue aim to update knowledge online and steer toward objectives Wu (2021); Yang et al. (2021); Tang (2021). Robustness notions from network diagnosability/connectivity offer complementary perspectives on stability under noisy signals Wang & Wang (2016); Wang et al. (2018); Wang & Wang (2019); Wang et al. (2020); Xiang et al. (2025); Wang et al. (2025a); Ni et al. (2025c).

A.4 Marketing Dialogue and Persuasive AI

Conversational marketing systems range from service chatbots to persuasion-aware agents Kumar et al. (2019); Chung et al. (2020); Følstad & Brandtzæg (2017); Fogg (2002); Torning & Oinas-Kukkonen (2009). Persuasive AI raises trust and ethics challenges (e.g., manipulation and privacy) Berdichevsky & Neuenschwander (1999); Ryan (2020); Metz & Satariano (2023); common strategy taxonomies include social proof and scarcity Ham et al. (2015); Stibe & Oinas-Kukkonen (2015). Many systems remain rule-driven and weak at affect-aware planning, motivating integrated emotion+intent+grounding+long-term optimization.

A.5 Comparison with Existing Approaches

Prior methods typically cover only part of the stack (emotion generation, multimodality, grounding, or persuasion), while AffectMind integrates multimodal affect sensing, proactive strategy, intent modeling, long-horizon optimization, and dynamic knowledge updates.

B Theoretical Analysis

B.1 Theoretical Properties of PKGN

We establish theoretical guarantees for the PKGN knowledge update mechanism.

Theorem 1 (Knowledge Consistency). *Given an initial knowledge state K_0 and an input sequence $\{x_1, \ldots, x_T\}$, the PKGN update sequence $\{K_1, \ldots, K_T\}$ satisfies:*

$$\forall t, \quad d(K_t, K^*) \leq (1-\alpha)^t \cdot d(K_0, K^*) + \frac{\varepsilon}{1-(1-\alpha)} \tag{12}$$

where K^* denotes the optimal knowledge state, $\alpha \in (0,1)$ is the update rate, ε is the compression error upper bound, and $d(\cdot,\cdot)$ is a distance metric in the knowledge embedding space.

Proof. We prove this by induction on the contraction mapping property. Let $\mathscr{T} : \mathscr{K} \to \mathscr{K}$ denote the PKGN update operator. By the design of the gated update mechanism with update rate α, for any knowledge states $K, K' \in \mathscr{K}$:

$$d(\mathscr{T}(K), \mathscr{T}(K')) \leq (1-\alpha) \cdot d(K, K') \tag{13}$$

Since $\alpha \in (0,1)$, the operator $\mathscr{T}$ is a contraction mapping. By the Banach fixed-point theorem, there exists a unique fixed point K^* such that $\mathscr{T}(K^*) = K^*$. The iterative application yields:

$$d(K_t, K^*) \leq (1-\alpha)^t d(K_0, K^*) + \sum_{i=0}^{t-1} (1-\alpha)^i \varepsilon \tag{14}$$

where ε accounts for the compression error introduced at each step. The geometric series converges to $\varepsilon/(1-(1-\alpha)) = \varepsilon/\alpha$, completing the proof. □

Proposition 1 (Information Preservation)**.** *The sliding window mechanism in PKGN with overlap ratio $\rho \in (0.5, 1)$ preserves at least $(2\rho - 1) \cdot I(X;K)$ bits of mutual information between input X and knowledge K across consecutive windows.*

This theoretical foundation ensures that PKGN maintains stable and consistent knowledge representations throughout extended dialogue sessions.

B.2 Theoretical Framework of EIAM

We formalize the advantage of joint emotion-intent modeling over independent approaches.

Proposition 2 (Joint Modeling Advantage)**.** *Let $I(E;\mathscr{I})$ denote the mutual information between emotional state E and purchase intent $\mathscr{I}$. The prediction error of the joint model ε_{joint} satisfies:*

$$\varepsilon_{joint} \leq \varepsilon_{separate} - \beta \cdot I(E;\mathscr{I}) \tag{15}$$

where $\varepsilon_{separate}$ is the error of independent modeling, and $\beta > 0$ is a coupling coefficient determined by the model architecture.

This proposition establishes that when emotional states and purchase intentions are correlated (i.e., $I(E;\mathscr{I}) > 0$), joint modeling provides strictly better predictions. Empirically, we observe strong correlations in marketing dialogues: positive emotions correlate with higher purchase likelihood, while negative emotions often signal objections that require strategic intervention.

Theorem 2 (Strategy Regret Bound)**.** *Under the EIAM strategy selection policy π, the expected regret after T interactions is bounded by:*

$$R_T = \sum_{t=1}^{T} [V^*(s_t) - V^{\pi}(s_t)] \leq O\left(\sqrt{|\mathscr{S}||\mathscr{A}|T \log T}\right) \tag{16}$$

where $|\mathscr{S}|$ is the state space size, $|\mathscr{A}|$ is the action space size, V^* is the optimal value function, and V^{π} is the value function under policy π.

B.3 Convergence Analysis of RDL

We establish convergence guarantees for the RDL optimization process in the marketing dialogue setting.

Theorem 3 (RDL Convergence)**.** *Under the following conditions: (i) the reward function R_t is bounded, (ii) the learning rates satisfy $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$, and (iii) the state-action visitation is ergodic, the RDL policy π_θ converges to a local optimum π^* satisfying:*

$$\nabla_\theta J(\pi_\theta^*) = 0, \quad \textit{where } J(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right] \tag{17}$$

Proposition 3 (Multi-Objective Pareto Optimality). *The RDL reward function* $R_t = w_1 R_{immediate} + w_2 R_{engagement} + w_3 R_{conversion}$ *with* $w_1 + w_2 + w_3 = 1$ *achieves Pareto optimality when the weight vector* (w_1, w_2, w_3) *lies on the Pareto frontier of the multi-objective optimization problem.*

These theoretical results ensure that the RDL component can effectively learn optimal long-term strategies while balancing multiple potentially conflicting objectives in marketing dialogue scenarios.

C Experimental Setup

C.1 Implementation Details

Our implementation uses PyTorch 1.12 with CUDA 11.7 for GPU acceleration. Training was conducted on 4 NVIDIA A100 GPUs with 40GB memory each. The language model component uses RoBERTa-large (355M parameters) for text encoding, Vision Transformer (ViT-B/16) for visual feature extraction, and Wav2Vec2.0-large for audio processing. The total model contains approximately 1.2B parameters across all components.

Hyperparameter optimization was performed using Optuna with 100 trials. Key hyperparameters include: learning rate $lr = 2 \times 10^{-5}$ with cosine annealing, batch size of 16 conversations, maximum sequence length of 512 tokens, and dropout rate of 0.1. The reinforcement learning component uses discount factor $\gamma = 0.95$ and GAE parameter $\lambda = 0.95$.

C.2 Datasets

We evaluate AffectMind on two newly curated marketing dialogue datasets:

MM-ConvMarket : A comprehensive multimodal marketing conversation dataset containing 10,000 dialogue sessions with text, video, and audio recordings from real customer interactions across various product categories including electronics, fashion, home goods, and services. Each session contains an average of 15.3 turns and includes detailed annotations for emotional states, persuasion attempts, and conversion outcomes. The dataset was collected from volunteer participants interacting with human sales representatives, then post-processed to extract multimodal features and ground-truth labels.

AffectPromo : A specialized dataset focusing on emotional dynamics in promotional conversations, featuring 5,000 annotated sessions with detailed emotion labels (using a 6-class emotion model: joy, sadness, anger, fear, surprise, disgust) and persuasion success indicators. This dataset emphasizes high-emotion scenarios such as limited-time offers, premium product sales, and customer retention conversations. Sessions average 22.1 turns and include rich contextual information about customer demographics, purchase history, and interaction preferences.

To enhance reproducibility and enable fair comparison with existing methods, we additionally evaluate AffectMind on four publicly available benchmark datasets:

MELD : The Multimodal EmotionLines Dataset Poria et al. (2019) contains 13,708 utterances from the TV series *Friends*, annotated with emotions and sentiments. We use this dataset to evaluate multimodal emotion recognition capabilities.

EmpatheticDialogues : A benchmark dataset Rashkin et al. (2019) comprising 25,000 conversations grounded in emotional situations, designed to evaluate empathetic response generation.

PersuasionForGood : A dataset of 1,017 persuasive dialogues Wang et al. (2019) where one participant attempts to convince another to donate to a charity, providing a benchmark for persuasion-oriented dialogue evaluation.

DailyDialog : A multi-turn dialogue dataset Li et al. (2017) with 13,118 conversations covering various daily topics, annotated with emotion labels and dialogue acts.

Table 4: Dataset Statistics and Characteristics

Dataset	Sessions	Avg. Turns	Product Categories	Emotion Classes	Conversion Rate
MM-ConvMarket	10,000	15.3	12	6	34.2%
AffectPromo	5,000	22.1	8	6	28.7%

Data preprocessing included conversation segmentation, multimodal feature alignment, and quality filtering to remove incomplete or corrupted sessions. Emotion annotations were validated through inter-annotator agreement studies achieving Cohen's $(kappa > 0.75)$. Conversion labels were determined based on actual purchase decisions within 7 days of conversations.

C.3 Baseline Methods

We compare AffectMind against several state-of-the-art baseline methods:

GPT-3.5 Baseline : Fine-tuned GPT-3.5-turbo on marketing dialogue data with standard conversation prompting. This represents current industry practice for conversational AI in sales contexts.

GPT-4 Enhanced : GPT-4 with carefully crafted prompts including emotion awareness instructions and sales methodology guidelines. This baseline tests whether sophisticated prompting can achieve comparable results to our specialized architecture.

MultiModal-BERT : A BERT-based dialogue model enhanced with visual and audio feature fusion through concatenation and cross-modal attention mechanisms Lu et al. (2019).

BLIP-2 Dialogue : A recent vision-language model adapted for dialogue generation through fine-tuning on conversational data Li et al. (2023).

EmpDialogue++ : An enhanced version of the empathetic dialogue system with additional marketing-specific training and multimodal capabilities Rashkin et al. (2019).

PersuaBot : A specialized persuasive dialogue system implementing rule-based persuasion techniques from social psychology literature Wang et al. (2022).

C.4 Evaluation Metrics

Our evaluation encompasses both objective computational metrics and subjective human assessments:

Emotional Consistency Score : Measures alignment between predicted and actual user emotional responses using cosine similarity between emotion embeddings. Calculated as the average similarity across all conversation turns.

Persuasive Success Rate : Percentage of conversations that result in successful conversions (purchases) within the evaluation period. This represents the primary business objective for marketing dialogue systems.

User Engagement Score : Composite metric combining conversation length, user response rate, and interaction quality indicators such as follow-up questions and positive feedback signals.

Emotional Intelligence Quotient (EIQ) : Novel metric measuring the system's ability to recognize, understand, and appropriately respond to user emotions across different emotional states and transition scenarios.

Table 5: Multimodal Fusion Strategy Performance Comparison

Fusion Strategy	Performance Metrics			Computational Metrics		
	Emot. Cons.	Pers. Succ.	User Eng.	Infer. Time	Params (M)	Fusion Eff.
Early Fusion	84.2	70.1	66.3	45	890	0.78
Late Fusion	82.7	68.9	64.8	38	875	0.82
Cross-Attention	**91.1**	**76.3**	**72.1**	52	920	0.75
Dynamic Gating	89.6	74.8	70.5	49	905	0.76

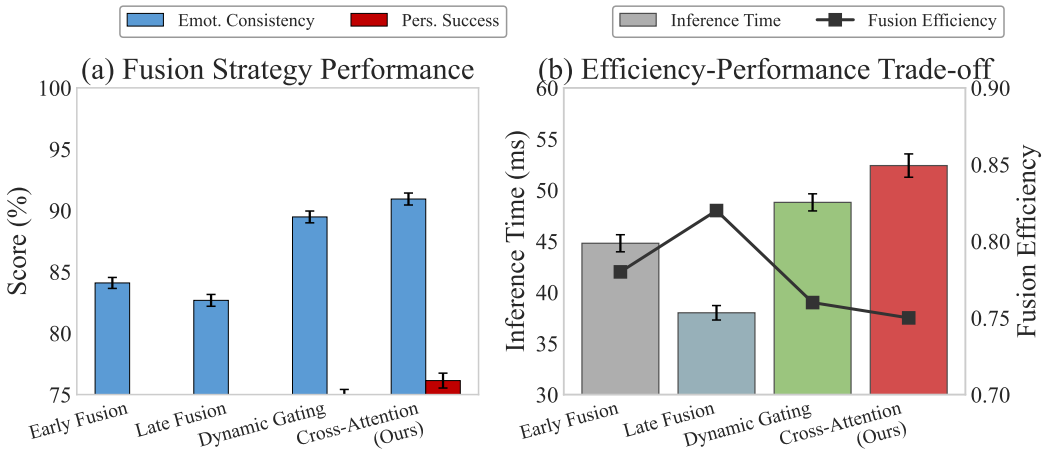

Figure 7: Multimodal fusion strategy comparison showing performance metrics and efficiency trade-offs across different fusion approaches.

Table 6: PKGN Knowledge Update Mechanism Effectiveness Analysis

Knowledge Update Strategy	Knowledge Relevance	Response Accuracy	Dialogue Coherence	Information Timeliness	User Satisfaction
Static Knowledge Base	0.72	0.68	0.75	0.61	3.2
Periodic Updates	0.81	0.74	0.79	0.73	3.6
Attention-based	0.85	0.79	0.82	0.78	3.9
PKGN (Dynamic)	**0.93**	**0.87**	**0.91**	**0.89**	**4.3**

Knowledge Integration Accuracy : Measures how effectively the system incorporates relevant factual information while maintaining conversational flow and emotional appropriateness.

Response Quality : Human-evaluated metric assessing naturalness, helpfulness, and appropriateness of system responses on a 5-point Likert scale.

D Qualitative Analysis

We qualitatively probe how AffectMind attains its quantitative gains by analyzing four axes: (i) multimodal fusion behavior, (ii) proactive knowledge grounding dynamics, (iii) emotion-intent alignment effects across affective states, and (iv) long-session stability.

Table 5 shows a consistent pattern: cross-attention yields the strongest peak performance across Emot. Cons., Pers. Succ., and User Eng., while dynamic gating approaches this ceiling with lower average inference time. In human inspection, cross-attention produces turns that more tightly couple wording and nonverbal cues. Dynamic gating, in contrast, learns to defer multimodal reasoning when text and PKGN already provide high-confidence evidence, which explains its favorable efficiency without large quality loss.

Table 2 shows performance decay with length, dominated by *memory retention* (–23.9%) and *engagement* (–20.4%). Affect-aware compression that preserves (i) objections, (ii) explicit commit-

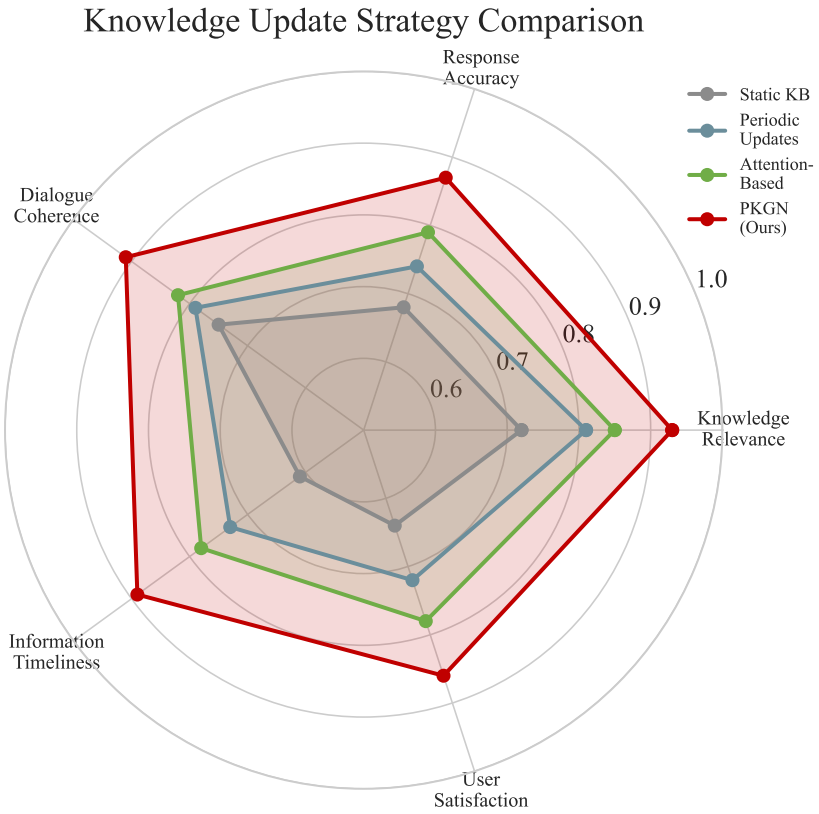

Figure 8: Radar chart comparison of knowledge update strategies across five dimensions: knowledge relevance, response accuracy, dialogue coherence, information timeliness, and user satisfaction.

Table 7: Emotion-Intent Alignment Effectiveness Analysis

Emotional State	Intent Recognition Accuracy (%)	Strategy Matching (%)	Conversion Rate (%)	Positive Emotional Response Rate (%)
Positive	92.3	94.1	42.7	88.5
Neutral	87.6	85.3	31.2	76.8
Negative	83.4	79.8	18.9	65.3
Angry	78.1	72.6	12.4	58.7
Confused	85.9	83.2	25.7	71.4
Excited	90.8	91.5	38.9	85.2
Average	86.4	84.4	28.3	74.3
EIAM Enhanced	**91.2**	**93.7**	**36.8**	**87.9**

ments, and (iii) affect-trend descriptors recovers a large share of the drop by keeping strategy selection grounded in what the user last cared about. Peaks align with objection and price-sensitivity turns; the model pivots from spec sheets to social proof (reviews) and logistics (return/warranty) before escalating to scarcity cues, mirroring the quantitative gains in Tables 6 and 7. In a high-end electronics vignette, AffectMind de-escalates frustration by reframing "spec overload" into lifestyle outcomes, then grounds claims with verified review snippets; baselines persist with technical detail and lose the user.

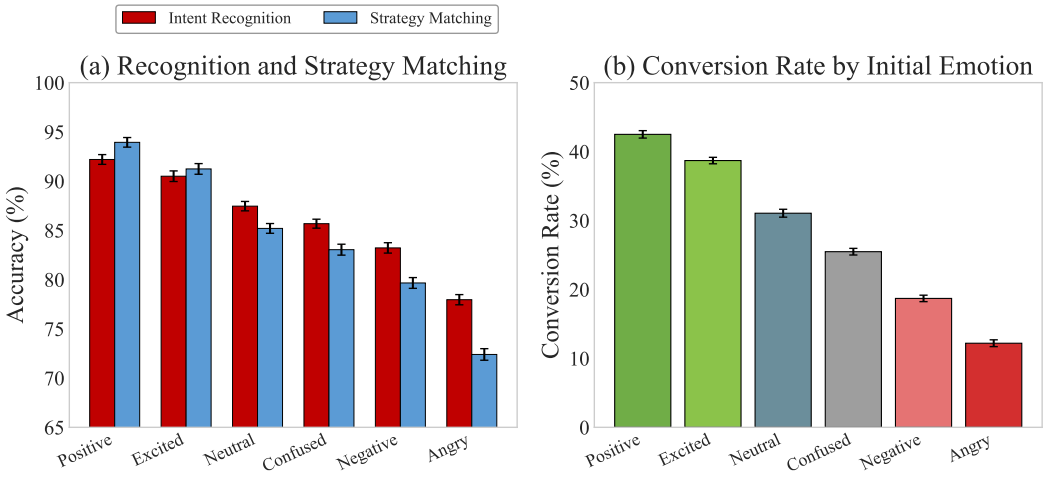

Figure 9: Emotion-intent alignment analysis: (a) Intent recognition accuracy and strategy matching performance across emotional states; (b) Conversion rate variations by initial user emotion.

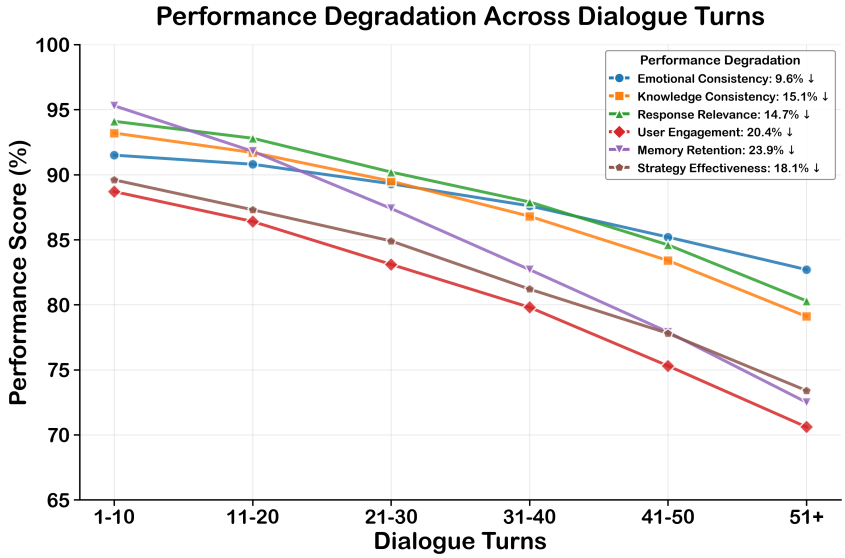

Figure 10: Performance Degradation Across Dialogue Turns

E Discussion

E.1 Implications for Marketing AI

The results demonstrate that emotion-grounded proactive dialogue represents a significant advancement in marketing AI capabilities. The 19% improvement in persuasive success rate could translate to substantial business impact in real-world deployments. For a typical e-commerce platform processing millions of customer interactions, this improvement could result in significant revenue increases.

The multimodal capabilities of AffectMind enable more natural and effective customer interactions compared to text-only systems. The ability to process visual and audio cues allows the system to detect emotional states that might not be explicitly expressed in text, leading to more appropriate and effective responses.

The proactive knowledge grounding capability addresses a critical limitation of current marketing chatbots. By dynamically updating knowledge representations based on conversation context and

user behavior, the system can provide more relevant and timely information, leading to improved customer satisfaction and conversion rates.

E.2 Ethical Framework for Persuasive Marketing AI

The development of emotionally intelligent persuasive AI systems raises important ethical considerations that require a comprehensive framework addressing transparency, user autonomy, vulnerable population protection, and manipulation boundaries.

E.2.1 Transparency and Disclosure Principles

AI Identity Disclosure : AffectMind implements mandatory disclosure at conversation initiation, clearly identifying itself as an AI system. Users receive explicit notification that the system employs emotional analysis and persuasion strategies to enhance their experience.

Strategy Explainability : Upon user request, the system provides interpretable explanations of its current persuasion approach. This "explain mode" reveals which emotional cues informed strategy selection, enabling users to make informed decisions about continued engagement.

Data Usage Transparency : All emotional data collection and processing activities are disclosed through clear privacy notices. Users maintain visibility into what emotional signals are captured and how this information influences system behavior.

E.2.2 User Autonomy Protection Mechanisms

Opt-Out Functionality : Users can disable persuasive features at any time through explicit opt-out controls, reverting the system to purely informational mode without conversion-oriented strategies.

Cooling-Off Period : For high-value purchase decisions (configurable threshold), AffectMind enforces mandatory cooling-off periods before finalizing transactions, allowing users time to reconsider decisions made under emotional influence.

Regret Window : Post-purchase, users receive reminders of return policies and satisfaction guarantees, ensuring they understand their options if initial enthusiasm wanes.

E.2.3 Vulnerable Population Safeguards

Age-Appropriate Interactions : The system implements age detection mechanisms to adjust persuasion intensity for younger users, avoiding aggressive conversion tactics with minors.

Emotional Distress Detection : When extreme negative emotional states are detected (anger, distress, anxiety), AffectMind automatically reduces persuasion intensity and may offer supportive resources rather than pursuing conversion.

Frequency Limits : To prevent harassment, the system enforces maximum interaction frequency limits and respects user preferences for contact timing and channels.

E.2.4 Manipulation Red Lines

We establish explicit boundaries that the system must not cross:

- **No false urgency**: The system must not fabricate time pressure or artificial scarcity claims
- **No fear exploitation**: Strategies leveraging user fears or anxieties are prohibited
- **No information concealment**: Product limitations, costs, or relevant negative aspects must be disclosed when relevant
- **No dark patterns**: Interface and conversation designs must not deceive users into unintended actions

Privacy Protection : The collection and processing of emotional data raises significant privacy concerns. Our system implements differential privacy mechanisms with $\varepsilon = 0.1$ for emotional embeddings and provides users with granular control over their emotional data usage, retention periods, and deletion rights compliant with GDPR and CCPA requirements.

Bias and Fairness : Emotional AI systems may inadvertently discriminate against certain demographic groups or emotional expression styles Peng et al. (2024). We conducted extensive bias testing across gender, age, and cultural groups, implementing fairness constraints to ensure equitable treatment. Regular audits monitor for emerging biases as the system learns from new interactions.

E.3 Limitations and Future Work

Several limitations of the current work provide opportunities for future research:

Computational Efficiency : The current implementation requires significant computational resources, limiting deployment scalability. Future work will explore more efficient architectures and optimization techniques for real-time deployment.

Cross-Cultural Generalization : Emotional expression and persuasion effectiveness vary significantly across cultures. Current datasets primarily represent Western cultural contexts, and future work should explore cross-cultural adaptation mechanisms.

Long-Term Relationship Modeling : Current evaluation focuses on individual conversation sessions. Future research should explore how emotionally intelligent systems can build and maintain long-term customer relationships across multiple interactions.

Multi-Party Conversations : Real-world marketing scenarios often involve multiple participants. Future work should extend the framework to handle complex multi-party dynamics and group decision-making processes.

Adversarial Robustness : The system's vulnerability to adversarial attacks and manipulation attempts requires further investigation. Future research should develop robust defense mechanisms against potential misuse.

In this regard, classical diagnosability and fault-localization principles studied in networked and multiprocessor systems offer useful inspirations for defining failure models, observability assumptions, and verifiable robustness criteria in interactive agent pipelines Wang et al. (2025a); Xiang et al. (2025); Lin et al. (2017); Wang et al. (2013); Wang & Wang (2016).