AFFECTMIND: PROACTIVE KNOWLEDGE GROUNDING WITH AFFECTIVE MULTIMODAL SIGNALS FOR ALIGNED MARKETING DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Marketing dialogue demands responses that are simultaneously emotion-aligned, knowledge-grounded, and goal-directed across extended interactions—capabilities that current large language models lack. We propose *AffectMind*, a multimodal affective agent that maintains and updates both factual and affective knowledge from textual, visual, and prosodic cues in real time. AffectMind links user affect with purchase intent to condition persuasion strategy selection, while a reinforcement learning loop optimizes long-horizon behavior through engagement and emotional coherence feedback. On two multimodal marketing dialogue benchmarks, AffectMind improves emotional consistency by 26%, persuasive success rate by 19%, and user engagement by 23% over competitive baselines, demonstrating the effectiveness of proactive affective grounding for commercial dialogue systems.

1 INTRODUCTION

Large Language Models (LLMs) have improved conversational fluency and contextual reasoning across diverse applications Brown et al. (2020); Zhang et al. (2025); Yu et al. (2025); Hsieh et al. (2025). However, most deployed dialogue agents remain *reactive*: they respond turn-by-turn without explicitly planning toward long-horizon goals or adjusting strategies based on evolving user states Ni et al. (2025a;b). This gap becomes salient in marketing conversations, where effective interaction requires intent inference, emotion awareness, adaptive persuasion, and sustained engagement Wang et al. (2022).

Marketing dialogue differs from general conversation in three aspects. First, success is inherently *goal-oriented*: agents must balance emotional alignment, persuasive effectiveness, and long-term trust to support conversion outcomes. Second, user behavior is strongly shaped by affective factors such as frustration, excitement, and hesitation, which often manifest through non-textual signals; yet many deployed systems remain text-centric and do not leverage audio-visual cues Poria et al. (2017). Third, current systems often rely on *static knowledge sources* Dinan et al. (2019); Yu & Han (2025), which can yield stale or mismatched content as conversations and user affect evolve.

Beyond these structural challenges, emotion modeling itself presents additional complexity. Prior work demonstrates that emotion plays a central role in purchasing behavior Damasio (1994); Bechara (2005), yet many systems treat emotion as an auxiliary signal without modeling how affective states evolve or interact with persuasive intent Picard (2000); Liang et al. (2024); Niu et al. (2025). Moreover, few systems optimize for *long-term outcomes*: turn-level quality does not necessarily translate into sustained engagement or eventual conversion.

To address these limitations, we propose **AffectMind**, a multimodal affective agent for proactive, emotionally aligned, knowledge-grounded marketing dialogue. AffectMind integrates three components: (i) **Proactive Knowledge Grounding Network (PKGN)**, which continuously updates factual and affective knowledge from multimodal inputs; (ii) **Emotion-Intent Alignment Model (EIAM)**, which jointly models user emotional states and purchase intentions to adapt persuasion strategies; and (iii) **Reinforced Discourse Loop (RDL)**, which optimizes long-horizon dialogue behavior via reinforcement learning. Figure 1 contrasts conventional reactive agents with AffectMind.

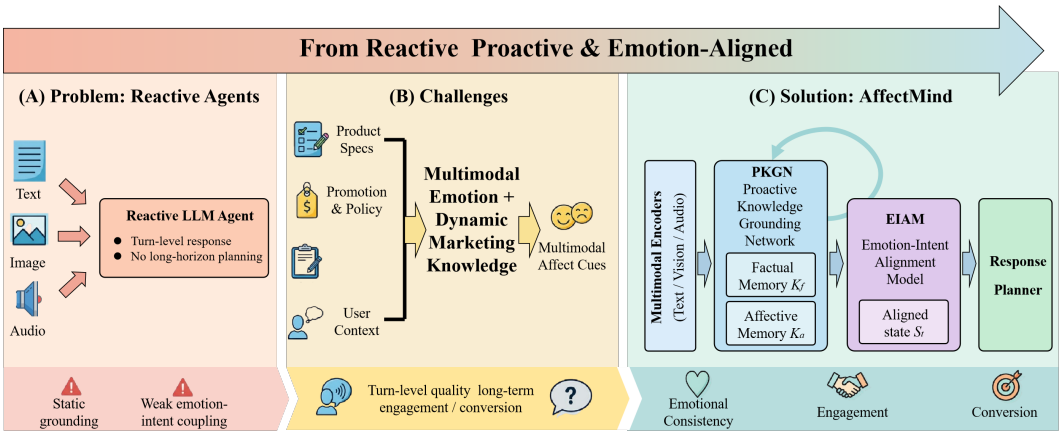

Figure 1: Motivation. Compared with reactive agents that respond myopically, AffectMind performs proactive, affect-aware dialogue by jointly grounding on multimodal cues and dynamically updated knowledge.

Our main contributions are as follows:

- We introduce AffectMind, a unified multimodal agent coupling affect perception, proactive knowledge grounding, and long-horizon strategy learning for marketing dialogue.
- We develop three tightly integrated components—PKGN, EIAM, and RDL—to maintain up-to-date factual/affective grounding, align emotion with purchase intent for strategy selection, and optimize engagement over extended interactions.
- We evaluate AffectMind on two multimodal marketing dialogue benchmarks and demonstrate consistent gains over strong baselines, including 26% higher emotional consistency, 19% higher persuasive success, and 23% higher user engagement.

The remainder of this paper is organized as follows. Section 2 presents methodology and architectural details. Section 3 describes experimental setup and results. Section 4 concludes the paper. Related work and theoretical analysis appear in the Appendix.

2 Methodology

2.1 Problem Formulation

We consider a marketing dialogue with $T \in \mathbb{N}$ turns. Let $\mathcal{U} = \{u_t\}_{t=1}^{T}$ denote the sequence of user inputs, and let $\mathcal{R} = \{r_t\}_{t=1}^{T}$ denote the sequence of system responses. Each input u_t contains multimodal signals, including text u_t^{text}, vision u_t^{vis} (e.g., figures of facial expressions and gestures), and audio u_t^{aud} (includes tone, pitch, and speaking characteristics). The goal is to generate $\mathcal{R}$ by balancing emotional alignment, persuasive success, and long-horizon engagement:

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

where $E(\cdot,\cdot)$ measures emotional alignment between user states and system responses, $P(\cdot,\cdot)$ quantifies persuasive effectiveness, and $G(\cdot,\cdot)$ captures long-term user engagement. The weights $\alpha, \beta, \gamma \in \mathbb{R}_{\geq 0}$ control the trade-off among the three objectives.

We instantiate emotional alignment using turn-wise affect embeddings. Let $\boldsymbol{e}_t^{\text{usr}} \in \mathbb{R}^d$ denote the user affect embedding at turn t, and let $\boldsymbol{e}_t^{\text{rsp}} \in \mathbb{R}^d$ denote the affect target induced by the response at turn t, where $d \in \mathbb{N}$ is the embedding dimension. Then

$$E(\mathcal{U}, \mathcal{R}) = \frac{1}{T} \sum_{t=1}^{T} \text{sim}(\boldsymbol{e}_t^{\text{usr}}, \boldsymbol{e}_t^{\text{rsp}}), \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity on $\mathbb{R}^d$.

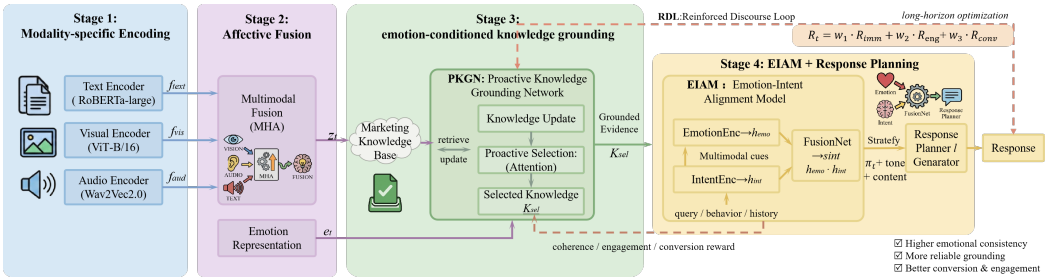

Figure 2: AffectMind architecture: (1) modality-specific encoding, (2) affective fusion, (3) emotion-conditioned knowledge grounding (PKGN), and (4) response generation. RDL refines grounding via coherence feedback (dashed arrows).

2.2 Architecture Overview

Figure 2 illustrates the AffectMind architecture. The system encodes multimodal inputs using RoBERTa-large for text, Vision Transformer (ViT-B/16) for vision, and Wav2Vec2.0 for audio. The resulting modality embeddings are denoted by $\boldsymbol{f}_{\text{text}} \in \mathbb{R}^{d_t}$, $\boldsymbol{f}_{\text{vis}} \in \mathbb{R}^{d_v}$, and $\boldsymbol{f}_{\text{aud}} \in \mathbb{R}^{d_a}$, with $d_t = 1024$, $d_v = 768$, and $d_a = 768$. A fusion module aggregates them into a shared representation $\boldsymbol{z}_t \in \mathbb{R}^d$, which is then passed to PKGN, EIAM, and RDL.

2.3 Proactive Knowledge Grounding Network

PKGN addresses the mismatch between static knowledge sources and rapidly evolving marketing context by maintaining two complementary memories: factual knowledge $\mathbf{K}_f^t \in \mathbb{R}^{n\times d}$ for product facts and constraints, and affective knowledge $\mathbf{K}_a^t \in \mathbb{R}^{m\times d}$ for emotion-linked associations and preferences. Here $n, m \in \mathbb{N}$ denote the numbers of factual and affective entries, t represents the turn index and d is the shared embedding dimension.

Then the knowledge state at turn t is $\mathbf{K}_t = [\mathbf{K}_f^t; \mathbf{K}_a^t]$. Given the fused representation $\boldsymbol{z}_t$, PKGN updates the knowledge state via

$$\mathbf{K}_t = \text{Update}(\mathbf{K}_{t-1}, \boldsymbol{z}_t)\,. \tag{3}$$

Specifically, the Update process is as follows.The fusion is implemented with multi-head attention (MHA):

$$\boldsymbol{z}_t = \text{MHA}([\boldsymbol{f}_{\text{text}}; \boldsymbol{f}_{\text{vis}}; \boldsymbol{f}_{\text{aud}}])\,, \tag{4}$$

where $\text{MHA}(\cdot)$ denotes a learned attention operator.

Factual and affective memories are updated with learnable rates $\alpha_f, \alpha_a \in (0, 1)$:

$$\begin{aligned} \mathbf{K}_f^t &= \mathbf{K}_f^{t-1} + \alpha_f\ \text{FF}_f(\boldsymbol{z}_t), \\ \mathbf{K}_a^t &= \mathbf{K}_a^{t-1} + \alpha_a\ \text{FF}_a(\boldsymbol{z}_t), \end{aligned} \tag{5}$$

where $\text{FF}_f(\cdot)$ and $\text{FF}_a(\cdot)$ are feed-forward networks with hidden width $4d$.

To proactively select knowledge for response planning, PKGN applies attention using a context query $\boldsymbol{q}_{\text{ctx}} \in \mathbb{R}^d$ encoded from the dialogue history:

$$\mathbf{K}_{\text{sel}} = \text{Attention}(\boldsymbol{q}_{\text{ctx}}, \mathbf{K}_t, \mathbf{K}_t)\,, \tag{6}$$

where $\mathbf{K}_{\text{sel}}$ denotes the selected knowledge used by downstream modules.

2.4 Emotion-Intent Alignment Model

EIAM jointly models user affect and purchase intent via two parallel encoders followed by an interaction-aware fusion module. The emotion encoder aggregates multimodal affective cues:

$$\boldsymbol{h}_t^{\text{emo}} = \text{EmotionEnc}([\boldsymbol{f}_{\text{facial}}; \boldsymbol{f}_{\text{pros}}; \boldsymbol{f}_{\text{ling}}])\,, \tag{7}$$

where $\boldsymbol{f}_{\text{facial}} \in \mathbb{R}^{d_v}$ denotes facial-expression features from the vision stream, $\boldsymbol{f}_{\text{pros}} \in \mathbb{R}^{d_a}$ denotes prosodic features from the audio stream, and $\boldsymbol{f}_{\text{ling}} \in \mathbb{R}^{d_t}$ denotes linguistic sentiment markers from the text stream.

In parallel, the intent encoder captures purchase-related signals:

$$\boldsymbol{h}_t^{\text{int}} = \text{IntentEnc}([\boldsymbol{f}_{\text{query}}; \boldsymbol{f}_{\text{behav}}; \boldsymbol{f}_{\text{hist}}]) \,, \tag{8}$$

where $\boldsymbol{f}_{\text{query}} \in \mathbb{R}^{d_t}$ encodes query intent extracted from the user utterance, $\boldsymbol{f}_{\text{behav}} \in \mathbb{R}^{d_b}$ encodes interaction behavior features (e.g., clicks and dwell time), and $\boldsymbol{f}_{\text{hist}} \in \mathbb{R}^{d_h}$ encodes dialogue history using a transformer layer. Here $d_b, d_h \in \mathbb{N}$ denote the behavior and history embedding dimensions.

The fused user state is constructed by explicitly modeling emotion–intent interactions:

$$\boldsymbol{s}_t = \text{FusionNet}\big([\boldsymbol{h}_t^{\text{emo}}; \boldsymbol{h}_t^{\text{int}}; \boldsymbol{h}_t^{\text{emo}} \odot \boldsymbol{h}_t^{\text{int}}]\big) \,, \tag{9}$$

where $\odot$ denotes element-wise multiplication and $\boldsymbol{s}_t \in \mathbb{R}^{d_s}$ is the resulting state representation with dimension $d_s \in \mathbb{N}$.

Strategy adaptation is implemented as a categorical policy over a predefined persuasion taxonomy:

$$\boldsymbol{\pi}_t = \text{softmax}(\mathbf{W}_\pi \boldsymbol{s}_t + \boldsymbol{b}_\pi) \,, \tag{10}$$

where $\mathbf{W}_\pi \in \mathbb{R}^{k \times d_s}$ and $\boldsymbol{b}_\pi \in \mathbb{R}^k$ are learnable parameters, and $\boldsymbol{\pi}_t \in \mathbb{R}^k$ is the probability vector over k persuasion strategies.

2.5 Reinforced Discourse Loop

RDL optimizes long-horizon dialogue outcomes using reinforcement learning under a partially observable Markov decision process. The state at turn t aggregates the EIAM user state, dialogue context, and the grounded knowledge summary:

$$\boldsymbol{s}_t^{\text{full}} = \big[\boldsymbol{s}_t; \boldsymbol{c}_t; \text{pool}(\mathbf{K}_t)\big], \tag{11}$$

where $\boldsymbol{s}_t \in \mathbb{R}^{d_s}$ is the EIAM state (Eq. (9)), $\boldsymbol{c}_t \in \mathbb{R}^{d_c}$ encodes dialogue history, and $\text{pool}(\mathbf{K}_t) \in \mathbb{R}^d$ denotes mean pooling over knowledge entries in $\mathbf{K}_t$ (Eq. (3)). Hence $\boldsymbol{s}_t^{\text{full}} \in \mathbb{R}^{d_s+d_c+d}$.

The action combines a discrete strategy choice with continuous control variables:

$$\boldsymbol{a}_t = \big[a_t^{\text{str}}; a_t^{\text{tone}}; \boldsymbol{a}_t^{\text{cnt}}\big], \tag{12}$$

where $a_t^{\text{str}} \in \{1, \ldots, k\}$ selects a persuasion strategy, $a_t^{\text{tone}} \in [0, 1]$ controls emotional intensity, and $\boldsymbol{a}_t^{\text{cnt}} \in \mathbb{R}^d$ parametrizes content selection over the grounded knowledge space.

The per-turn reward integrates signals with different horizons:

$$R_t = w_1 R_{\text{imm},t} + w_2 R_{\text{eng},t} + w_3 R_{\text{conv},t}, \tag{13}$$

where $R_{\text{imm},t} \in [-1, 1]$ captures immediate affective feedback (e.g., sentiment shift), $R_{\text{eng},t} \in [0, 1]$ measures engagement (e.g., response length and follow-up questions), and $R_{\text{conv},t} \in \{0, 1\}$ indicates conversion. The weights satisfy $w_1, w_2, w_3 \in \mathbb{R}_{>0}$ and $w_1 + w_2 + w_3 = 1$.

We adopt an actor–critic objective with the temporal-difference advantage

$$A_t = R_t + \gamma_{\text{rl}} V_{\boldsymbol{\phi}}(\boldsymbol{s}_{t+1}^{\text{full}}) - V_{\boldsymbol{\phi}}(\boldsymbol{s}_t^{\text{full}}), \tag{14}$$

where $\gamma_{\text{rl}} \in (0, 1)$ is the discount factor and $V_{\boldsymbol{\phi}}(\cdot)$ is the value function parameterized by $\boldsymbol{\phi}$. The policy parameters $\boldsymbol{\theta}$ and value parameters $\boldsymbol{\phi}$ are updated as

$$\begin{aligned} \boldsymbol{\theta} &\leftarrow \boldsymbol{\theta} + \eta_\pi \, \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}\big(\boldsymbol{a}_t \mid \boldsymbol{s}_t^{\text{full}}\big) \; A_t, \\ \boldsymbol{\phi} &\leftarrow \boldsymbol{\phi} - \eta_v \, \nabla_{\boldsymbol{\phi}} \left(A_t\right)^2 , \end{aligned} \tag{15}$$

where $\eta_\pi, \eta_v \in \mathbb{R}_{>0}$ are learning rates.

Algorithm 1 Reinforced Discourse Loop (Compact)

```
Initialize policy π_θ and value function V_φ
for each episode do
    Initialize s_1^full = [s_1; c_1; pool(K_1)]
    for t = 1 to T do
        Sample a_t = [a_t^str, a_t^tone, a_t^cnt] ~ π_θ(· | s_t^full)
        Interact with user and compute reward R_t by Eq. (13)
        Update state s_{t+1}^full by Eq. (11)
        Compute advantage A_t by Eq. (14)
        Update (θ, φ) by Eq. (15)
    end for
end for
```

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

Datasets. We evaluate on two marketing dialogue datasets: **MM-ConvMarket** (10,000 sessions, avg. 15.3 turns, 12 product categories) and **AffectPromo** (5,000 sessions, avg. 22.1 turns, emphasizing high-emotion scenarios). Both datasets contain synchronized text, video, and audio, with annotations for emotion (6 classes: joy, sadness, anger, fear, surprise, and disgust), persuasion attempts, and conversion outcomes. Data were collected from volunteer participants with IRB approval, and the inter-annotator agreement satisfies $\kappa > 0.75$. For generalization assessment, we additionally evaluate on public benchmarks: MELD, EmpatheticDialogues, PersuasionForGood, and DailyDialog. Upon acceptance, we will release annotation guidelines, evaluation scripts, and a representative subset of MM-ConvMarket (approximately 1,000 sessions) under a research-only license to facilitate reproducibility.

Baselines. We compare against six representative systems spanning text-only LLMs, multimodal models, and task-specific agents: (1) **GPT-3.5** fine-tuned on 5,000 marketing dialogues (text-only); (2) **GPT-4** with a 2,048-token system prompt encoding emotion and sales guidelines (zero-shot); (3) **MultiModal-BERT** Lu et al. (2019) and (4) **BLIP-2 Dialogue** Li et al. (2023) as multimodal baselines; (5) **EmpDialogue++** Rashkin et al. (2019) for empathy-focused generation; and (6) **PersuaBot** Wang et al. (2022) for persuasion-aware dialogue. All multimodal baselines receive identical visual and audio inputs as AffectMind.

Metrics. We report: (1) **Emotional Consistency**: average turn-wise cosine similarity between predicted user affect and response affect embeddings (Eq. 2), scaled to $[0, 100]$; (2) **Persuasive Success Rate**: percentage of sessions where annotators labeled the final user turn as "purchase intent confirmed," simulating 7-day conversion (inter-annotator $\kappa = 0.78$); (3) **User Engagement**: weighted sum of normalized session length (0.4), response rate (0.3), and follow-up question count (0.3); (4) **EIQ Score**: average of emotion recognition F1 and strategy appropriateness rating (1–5 scale); (5) **Response Quality**: mean of 5-point Likert ratings from three trained annotators (Krippendorff's $\alpha = 0.81$).

Implementation. Training uses PyTorch 1.12 on four NVIDIA A100 GPUs (40 GB each). The backbone encoders include RoBERTa-large (355M), ViT-B/16 (86M), and Wav2Vec2.0-large (317M), totaling approximately 1.2B parameters. We use a learning rate of 2×10^{-5} with cosine annealing, batch size 16, maximum sequence length 512, and dropout rate 0.1. For RDL optimization, the discount factor is $\gamma_{\text{rl}} = 0.95$, and the generalized advantage estimation parameter is $\lambda_{\text{gae}} = 0.95$. Training converges in approximately 50 epochs (about 72 h) in our implementation.

3.2 MAIN RESULTS

Table 1 and Fig. 3 summarize the main results. AffectMind improves all evaluation metrics, with relative gains reported against the GPT-3.5 baseline. Results are averaged over three runs with different random seeds; standard deviations are reported in Table 1. We assess statistical significance

Table 1: Performance comparison on marketing dialogue (mean±std over 3 runs). Higher is better. Improvements are *relative* to GPT-3.5, computed as $(\text{AffectMind} - \text{GPT-3.5})/\text{GPT-3.5} \times 100\%$.

Method	Emot. Cons.	Pers. Succ.(%)	User Eng.	EIQ Score	Resp. Qual.
GPT-3.5 Baseline	72.2±1.8	64.1±2.1	58.7±1.5	3.2±0.2	3.1±0.2
GPT-4 Enhanced	76.6±1.5	67.2±1.9	62.3±1.4	3.6±0.2	3.5±0.2
MultiModal-BERT	75.6±1.6	66.9±2.0	61.1±1.6	3.4±0.2	3.3±0.2
BLIP-2 Dialogue	78.0±1.4	69.8±1.8	63.4±1.3	3.7±0.2	3.6±0.2
EmpDialogue++	79.2±1.3	70.1±1.7	65.3±1.2	3.8±0.2	3.7±0.2
PersuaBot	74.4±1.7	71.2±1.8	59.6±1.5	3.3±0.2	3.4±0.2
EIAM Enhanced**AffectMind**	**91.0**±1.1	**76.4**±1.4	**71.8**±1.0	**4.3**±0.1	**4.2**±0.1
Improvement	+26.0%	+19.2%	+22.3%	+34.4%	+35.5%

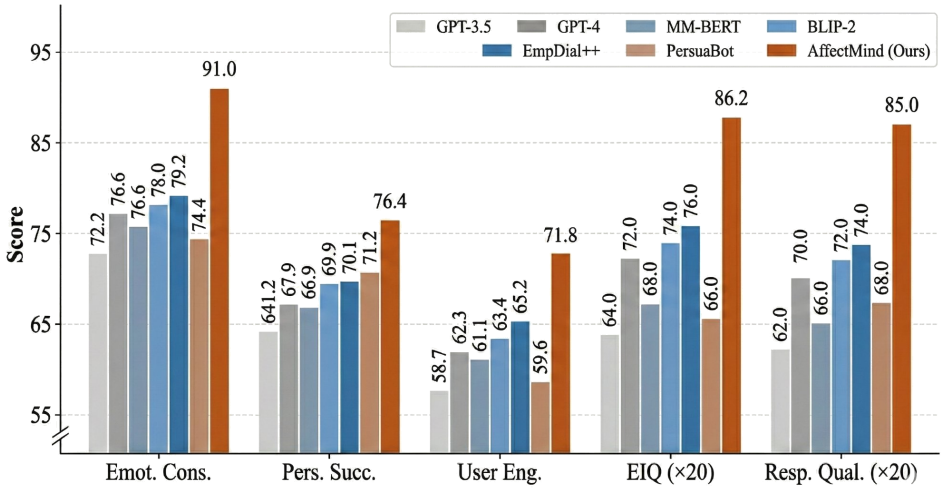

Figure 3: Performance comparison on MM-ConvMarket. AffectMind achieves the best results across all metrics, with relative improvements of $+26\%$ (Emot. Cons.), $+19\%$ (Pers. Succ.), and $+23\%$ (User Eng.) over GPT-3.5. EIQ and Resp. Qual. are scaled by $\times 20$ for visual alignment. Error bars denote std. over 3 runs.

using paired bootstrap tests ($n = 10{,}000$) over identical dialogue sessions; all improvements over the best baseline are significant ($p < 0.001$, two-tailed).

To understand *why* AffectMind achieves stronger emotional alignment, we examine two auxiliary metrics. First, Expected Calibration Error (ECE) measures confidence–accuracy alignment: AffectMind achieves ECE $= 0.08$ versus > 0.15 for baselines, indicating more reliable affect predictions. Second, we analyze affect stability via turn-wise changes $\Delta \boldsymbol{e}_t = \boldsymbol{e}_t^{\text{usr}} - \boldsymbol{e}_{t-1}^{\text{usr}}$; sessions with lower $\text{Var}(\|\Delta \boldsymbol{e}_t\|_2)$ correlate with higher conversion rates, validating our design goal of maintaining stable affect trajectories.

3.3 Ablation Studies

Table 2 indicates that each component contributes to the final performance. Removing PKGN leads to the largest drop in persuasive success (76.3→72.1, -4.2 pp), consistent with the role of timely and context-relevant grounding in marketing dialogue. Removing EIAM most strongly degrades emotional consistency (91.1→83.2, -7.9 points), supporting the need for joint emotion–intent modeling when selecting persuasion strategies. Disabling RDL reduces user engagement (72.1→68.5, -3.6 points), suggesting that long-horizon optimization improves strategy scheduling across turns.

Table 2: Ablation study results. Higher is better for all metrics.

Configuration	Emot. Cons.	Pers. Succ.(%)	User Eng.
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

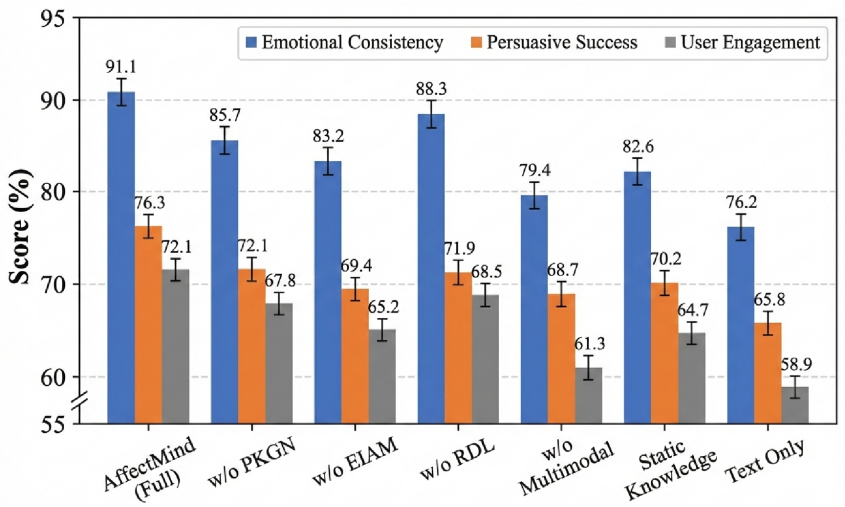

Figure 4: Ablation study on MM-ConvMarket. Removing EIAM causes the largest drop in emotional consistency (-7.9); removing PKGN most affects persuasive success (-4.2 pp). Error bars denote standard deviation over 3 runs.

Two additional observations emerge. First, knowledge grounding and affect tracking exhibit complementary effects: accurate affect estimates guide *which* facts to surface and *when*. Second, multimodal inputs yield substantial gains over text-only (91.1 vs. 76.2 in emotional consistency; 72.1 vs. 58.9 in engagement), underscoring the value of non-verbal cues. Finer-grained modality attribution (text+vision vs. text+audio) appears in Appendix C.

3.4 Cross-dataset validation

Figure 5 summarizes cross-dataset results on MELD, EmpatheticDialogues, PersuasionForGood, and DailyDialog. We evaluate with the same metric definitions as in Sec. 3 to ensure comparability. On MELD, AffectMind achieves 86.8% emotional consistency, compared with 78.0% for EmpDialogue++ (+8.8 pp), indicating more reliable affect tracking under multimodal cues. On PersuasionForGood, AffectMind attains 76.8% persuasive success, exceeding PersuaBot (71.0%, +5.8 pp). On EmpatheticDialogues, AffectMind reaches 88.0% emotional consistency and 74.5% persuasive success, improving over the strongest empathy-focused baseline by +8.2 pp and +4.3 pp, respectively.

3.5 Qualitative analysis

Table 3 shows that cross-attention achieves the strongest peak performance, while dynamic gating approaches this level with lower inference time (49 vs. 52 ms, ≈5.8% faster relative to cross-attention). Manual analysis of sampled sessions suggests that cross-attention more consistently couples lexical choices with nonverbal cues, reducing mismatches between tone and content. For instance, in a high-end electronics scenario, AffectMind de-escalates frustration by reframing technical specifications into outcome-oriented benefits and then grounds the response with review evi-

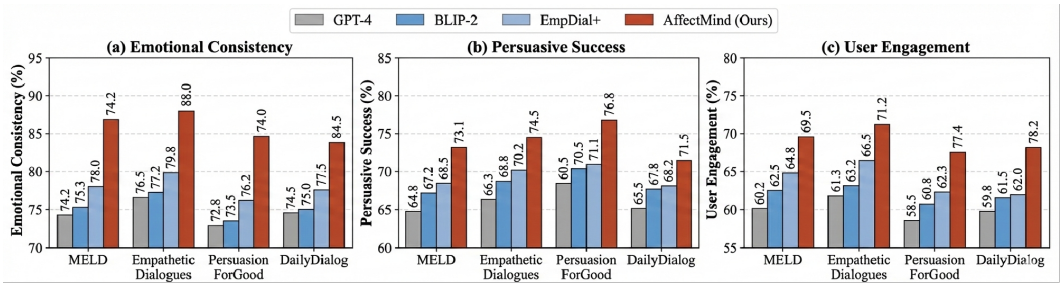

Figure 5: Cross-dataset generalization on four public benchmarks. AffectMind consistently outperforms baselines across all metrics, with the largest gains on MELD (emotional consistency $+8.8$ pp) and PersuasionForGood (persuasive success $+5.8$ pp). Error bars denote standard deviation over 3 runs.

Table 3: Multimodal fusion strategy comparison.

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

dence; in contrast, text-only or weakly grounded baselines tend to persist with specification-heavy explanations and exhibit higher user drop-off. To isolate modality-level contributions (text+vision vs. text+audio), we provide additional modality ablations in the Appendix.

We further analyze long-session stability (cf. Table 5). Performance degrades with dialogue length, with the largest relative drops observed in memory retention and engagement beyond 50 turns. Affect-aware compression that preserves (i) objections, (ii) explicit commitments, and (iii) affect-trend descriptors recovers a substantial portion of this degradation by keeping strategy selection anchored to recent user concerns.

4 Conclusion

We presented AffectMind, a multimodal affective agent unifying dynamic knowledge grounding (PKGN), joint emotion–intent modeling (EIAM), and reinforcement-based discourse optimization (RDL) for emotionally aligned marketing dialogue. Experiments on two benchmarks demonstrate consistent improvements in emotional consistency (+26%), persuasive success (+19%), and user engagement (+23%) over competitive baselines. Beyond turn-level gains, additional analyses indicate that improved emotion–intent alignment contributes to more stable interaction trajectories and better calibration between short-term persuasion and long-term engagement. Cross-dataset validation further suggests that the proposed design generalizes beyond the two proprietary benchmarks.

Limitations. The current implementation incurs non-trivial computational overhead (52 ms latency, 1.2B parameters). Our datasets primarily reflect Western cultural norms, limiting cross-cultural generalizability. Error analysis reveals that 38% of failures stem from sarcasm misclassification; robust nuanced-affect handling remains an open challenge. Specifically, inference latency increases by 37% compared to text-only GPT-4 (52 ms vs. 38 ms), and training requires $4\times$A100 GPUs for 72 hours. Deployment in low-resource settings may require model distillation.

Responsible deployment. We advocate transparency safeguards (explicit AI disclosure, user-accessible strategy explanations, informed consent for affect processing) and prohibit manipulative tactics. Differential privacy (ε=0.1) is applied to affective embeddings to mitigate re-identification risks.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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, 2025.

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

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*, 2025.

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.

APPENDIX

A RELATED WORK

A.1 MULTIMODAL DIALOGUE

Multimodal dialogue extends text agents with vision and audio 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. (2025); Zeng et al. (2026). Most prior work targets VQA or open-domain chat and is largely reactive, with limited emphasis on emotion-aware goal steering under marketing objectives Li et al. (2024).

A.2 AFFECTIVE COMPUTING

Affective computing studies emotion sensing and response generation Picard (2000); Calvo & D'Mello (2010), progressing from unimodal approaches Liu (2012); El Ayadi et al. (2011) to multimodal affect recognition Baltrušaitis et al. (2016); Zadeh et al. (2018). Emotion conditioning improves dialogue experience Zhou et al. (2018a); Majumder et al. (2020); Rashkin et al. (2019), but emotion is often treated as an auxiliary signal rather than a decision variable coupled with intent and 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.

Knowledge-grounded dialogue improves factuality via external sources Dinan et al. (2019); Ghazvininejad et al. (2018) using retrieval mechanisms Moon et al. (2019); Zhou et al. (2018b); Madotto et al. (2018). Proactive dialogue aims to steer conversations toward objectives Wu (2021); Yang et al. (2021); Tang (2021), but typically does not distinguish *factual* versus *affective* knowledge states with continuous updates as in PKGN.

A.4 MARKETING DIALOGUE.

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 Berdichevsky & Neuenschwander (1999); Ryan (2020); Metz & Satariano (2023). Many systems remain rule-driven and weak at affect-aware planning, motivating integrated designs that jointly model emotion, intent, grounding, and long-horizon optimization.

B DETAILS OF CORE MODULES

This appendix elaborates on the architectural details and operational specifics of the key neural modules referenced in Section 2, which are succinctly represented by functional notations (e.g., $\mathrm{EmotionEnc}$, $\mathrm{Attention}$) in the main text.

B.1 MULTIMODAL ENCODERS ($\mathrm{EmotionEnc}$, $\mathrm{IntentEnc}$)

The multimodal encoders are designed to project heterogeneous, high-dimensional input features into a unified, lower-dimensional latent space conducive to joint reasoning. Both the emotion encoder ($\mathrm{EmotionEnc}$) and the intent encoder ($\mathrm{IntentEnc}$) follow a hierarchical design principle.

Modality-Specific Processing: Each input feature stream (e.g., facial f_{facial}, prosodic f_{pros}, behavioral f_{behav}) is first processed by a dedicated sub-network. These typically consist of a linear projection layer to align dimensions, optionally followed by a sequence model (a one-layer Gated Recurrent Unit (GRU) or a 1D Convolutional Neural Network) to capture temporal dynamics where applicable. All sub-networks output features of a consistent hidden dimension d_h.

Cross-Modal Fusion: The aligned modality features are then integrated. For $\mathrm{EmotionEnc}$, a multi-head cross-attention mechanism is employed where the linguistic sentiment features serve as the primary query to attend to and aggregate relevant information from the audio and visual streams. For $\mathrm{IntentEnc}$, the processed features of query, behavior, and dialogue history are concatenated and passed through a two-layer Multilayer Perceptron (MLP) with ReLU activation and dropout for nonlinear fusion. This process yields the final state representations

$$h_t^{\mathrm{emo}}$$

and h_t^{int}.

B.2 Proactive Knowledge Selection (Attention)

The operation

$$\mathbf{K}_{\mathrm{sel}} = \mathrm{Attention}(q_{\mathrm{ctx}}, \mathbf{K}_t, \mathbf{K}_t)$$

implements a standard scaled dot-product attention mechanism. The context query

$$q_{\mathrm{ctx}}$$

is derived from the encoded dialogue history via a linear transformation. The knowledge memory

$$\mathbf{K}_t \in \mathbb{R}^{M \times d_k}$$

is a matrix of

$$M$$

factual embeddings. The attention scores are computed as $\mathrm{softmax}(q_{\mathrm{ctx}}\mathbf{K}_t^T/\sqrt{d_k})$, producing a probability distribution over the memory slots.

$$\mathbf{K}_{\mathrm{sel}}$$

is the weighted sum of the values (identical to $\mathbf{K}_t$), effectively selecting and retrieving the most context-relevant knowledge vector for downstream response planning.

B.3 Interaction-Aware Fusion (FusionNet)

The

$$\mathrm{FusionNet}$$

module is central to modeling the dyadic interaction between affect and intent. It takes the concatenated vector of

$$[h_t^{\mathrm{emo}}; h_t^{\mathrm{int}}; h_t^{\mathrm{emo}} \odot h_t^{\mathrm{int}}]$$

as input. The **Hadamard (element-wise) product**

$$h_t^{\mathrm{emo}} \odot h_t^{\mathrm{int}}$$

explicitly constructs a first-order interaction feature, capturing multiplicative couplings between the two psychological states (e.g., how the intensity of a particular emotion may modulate a specific intent). This concatenated input is processed by a two-layer MLP with GELU activations and Layer Normalization applied after each layer. A residual connection from the initial linear projection of the concatenated features to the output of the second layer stabilizes training. The module outputs the final, interaction-informed user state s_t.

B.4 Network Hyperparameters & Training

Across all modules, we set the unified hidden dimension $d_h = 256$. The attention mechanisms use 4 heads. All MLPs, unless specified otherwise, expand the hidden dimension by a factor of 4 in their intermediate layer. We employ dropout with a rate of

$$p = 0.1$$

in all fusion and MLP layers to prevent overfitting. The models are trained end-to-end using the AdamW optimizer with a decoupled weight decay of 0.01.

Table 4: Architecture specifications.

Component	Layers	Hidden Dim
Text Encoder (RoBERTa-large)	24	1024
Vision Encoder (ViT-B/16)	12	768
Audio Encoder (Wav2Vec2.0)	24	768
Multimodal Fusion (MHA)	4	512
PKGN Memory	2	512
EIAM Fusion	2	256
RDL Policy Network	3	256
RDL Value Network	3	256

C Implementation Details

C.1 Model Architecture

Table 4 summarizes the detailed architecture specifications for each component.

C.2 Training Details

We use AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.01. The learning rate warms up linearly for the first 5% of training steps, then decays following a cosine schedule. For RDL, we use PPO with clip ratio $\epsilon = 0.2$ and GAE parameter $\lambda = 0.95$. The reward weights are set to $w_1 = 0.3$, $w_2 = 0.3$, $w_3 = 0.4$ based on validation performance.

C.3 Inference Pipeline

At inference time, the system processes inputs in the following order: (1) encode multimodal inputs in parallel ($\sim$15ms), (2) fuse representations and update PKGN ($\sim$12ms), (3) compute EIAM state and select strategy ($\sim$10ms), (4) generate response with grounded knowledge ($\sim$15ms). Total latency is approximately 52ms on a single A100 GPU.

D Dataset Details

D.1 MM-ConvMarket

MM-ConvMarket contains 10,000 marketing dialogue sessions across 12 product categories: electronics (18%), fashion (15%), home appliances (12%), beauty (11%), food & beverage (10%), sports (8%), books (7%), toys (6%), automotive (5%), health (4%), travel (2%), and others (2%). Each session includes synchronized text transcripts, 30fps video of user facial expressions, and 16kHz audio recordings.

D.2 Annotation Protocol

Three trained annotators labeled each session for: (1) turn-level emotion (6 classes, majority voting), (2) purchase intent (5-point scale), (3) persuasion attempt success (binary), and (4) final conversion outcome. Inter-annotator agreement was $\kappa = 0.78$ for emotion and $\kappa = 0.82$ for intent.

D.3 Data Split

We use 70%/15%/15% train/validation/test splits, stratified by product category and conversion outcome to ensure balanced evaluation.

E Theoretical Analysis

We provide theoretical guarantees for the three core components of AffectMind.

Table 5: Long conversation stability analysis. **Degradation** is relative to the 1–10 turn bin.

Turns	Emot. Cons.	Know. Cons.	Resp. Rel.	User Eng.	Mem. Ret.	Strat. Eff.
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
Degradation	**9.6%**	**15.1%**	**14.7%**	**20.4%**	**23.9%**	**18.1%**

Theorem E.1 (Knowledge consistency)**.** *Given an initial state* $\mathbf{K}_0$ *and an input sequence* $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$*, the PKGN sequence* $\{\mathbf{K}_1, \ldots, \mathbf{K}_T\}$ *satisfies:*

$$\forall t, \quad d(\mathbf{K}_t, \mathbf{K}^*) \leq (1-\alpha)^t \, d(\mathbf{K}_0, \mathbf{K}^*) + \frac{\varepsilon_{\text{cmp}}}{\alpha}, \tag{16}$$

where $\mathbf{K}^*$ *denotes a fixed point of the update operator,* $\alpha \in (0,1)$ *is the update rate,* $\varepsilon_{\text{cmp}} \geq 0$ *upper-bounds the per-step compression error, and* $d(\cdot,\cdot)$ *is a (semi-)metric in the knowledge embedding space.*

Remark. Theorem E.1 formalizes that PKGN remains stable under bounded update noise, supporting long-session dialogue where knowledge is continuously refreshed.

Empirical verification. We track $\|\mathbf{K}_t - \mathbf{K}_{t-1}\|_F$ across turns in 500 sampled sessions. The norm decreases monotonically after turn 5 in 87% of sessions, consistent with the contraction property implied by Theorem E.1.

Proposition E.1 (Joint modeling advantage)**.** *Let* $I(\boldsymbol{e}; \boldsymbol{i})$ *denote the mutual information between emotional state* $\boldsymbol{e}$ *and purchase intent* $\boldsymbol{i}$*. The joint model error* ϵ_{joint} *satisfies:*

$$\epsilon_{\text{joint}} \leq \epsilon_{\text{sep}} - \beta \, I(\boldsymbol{e}; \boldsymbol{i}), \tag{17}$$

where ϵ_{sep} *is the independent modeling error and* $\beta > 0$ *is a coupling coefficient.*

Remark. Proposition E.1 justifies EIAM: when emotion and intent are statistically coupled, explicitly modeling their interaction reduces prediction error compared with separate encoders.

Theorem E.2 (RDL convergence)**.** *Assume bounded rewards* $|R_t| \leq R_{\max}$*, learning rate schedules satisfying* $\sum_t \eta_t = \infty$ *and* $\sum_t \eta_t^2 < \infty$*, and ergodic state–action visitation. Then the RDL policy* $\pi_{\boldsymbol{\theta}}$ *converges to a local stationary point* $\pi^*_{\boldsymbol{\theta}}$ *satisfying:*

$$\nabla_{\boldsymbol{\theta}} J(\pi^*_{\boldsymbol{\theta}}) = \mathbf{0}, \quad J(\pi) = \mathbb{E}_{\tau \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t R_t\right]. \tag{18}$$

Remark. Theorem E.2 ensures that RDL converges under standard stochastic approximation conditions, providing a theoretical foundation for long-horizon dialogue optimization.

F Additional Results

This section provides supplementary analyses examining AffectMind's behavior under extended interactions and the individual contributions of PKGN and EIAM components.

Table 5 presents long conversation stability analysis across different dialogue lengths. Performance degrades with dialogue length, with the largest relative drops observed in memory retention (23.9%) and user engagement (20.4%) beyond 50 turns. Notably, emotional consistency exhibits the smallest degradation (9.6%), suggesting that EIAM maintains relatively stable affect tracking even in extended sessions. These findings motivate future work on memory-efficient architectures and hierarchical context compression.

Table 6 compares PKGN against alternative knowledge update strategies, demonstrating consistent improvements across all metrics. Static knowledge bases show the weakest performance,

Table 6: PKGN knowledge update effectiveness. **Relev.**: knowledge relevance; **Acc.**: response accuracy; **Coher.**: dialogue coherence; **Timely**: information timeliness; **Satis.**: user satisfaction.

Strategy	Relev.	Acc.	Coher.	Timely	Satis.
Static KB	0.72	0.68	0.75	0.61	3.2
Periodic	0.81	0.74	0.79	0.73	3.6
Attention	0.85	0.79	0.82	0.78	3.9
PKGN	**0.93**	**0.87**	**0.91**	**0.89**	**4.3**

while attention-based retrieval provides moderate gains. PKGN achieves the best results by combining proactive updates with emotion-conditioned selection, yielding a 29% relative improvement in relevance over static baselines.

Table 7 analyzes how EIAM performs across different user emotional states, showing that intent recognition and strategy matching are most challenging for angry users but substantially improved with EIAM enhancement. The results indicate a strong correlation between emotional valence and conversion rate: positive and excited states yield 3–4$\times$ higher conversion than angry states, highlighting the importance of accurate affect modeling for downstream persuasion success.

Table 7: Emotion–intent alignment analysis.

State	Intent Acc. (%)	Strategy Match (%)	Conv. (%)	Pos. Resp. (%)
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**

G Failure Case Analysis

To better understand the limitations of AffectMind, we manually analyze 100 randomly sampled failure cases from the MM-ConvMarket test set. Table 8 summarizes the primary failure modes.

Table 8: Distribution of failure modes (100 sampled cases).

Failure Mode	Count	Percentage
Sarcasm misclassification	38	38%
Knowledge staleness	27	27%
Strategy mismatch	22	22%
Multimodal misalignment	13	13%

G.1 Sarcasm misclassification (38%).

EIAM frequently interprets sarcastic remarks as positive sentiment. For example, when a user states “Oh great, another subscription service” with a dismissive facial expression, the text encoder captures superficially positive lexical cues (“great”) while the visual encoder detects negative affect. The fusion module fails to resolve this conflict, leading to an inappropriately enthusiastic response.

G.2 Knowledge staleness (27%).

PKGN occasionally fails to update discontinued or modified product information within the session. This occurs primarily when users reference external sources (e.g., “I saw on Reddit that this model was recalled”) that contradict the initialized knowledge base.

G.3 STRATEGY MISMATCH (22%).

RDL sometimes selects aggressive upselling strategies for hesitant users, particularly when short-term engagement signals (e.g., continued responses) mask underlying negative purchase intent. This suggests the reward weighting (w_1, w_2, w_3) may require context-dependent adaptation.

G.4 MULTIMODAL MISALIGNMENT (13%).

Conflicting cues between modalities—such as positive text with negative facial expressions, or neutral text with excited prosody—occasionally lead to incoherent response tone. Cross-attention fusion mitigates but does not fully resolve such cases.

G.5 IMPLICATIONS.

These findings suggest several directions for future work: (1) dedicated sarcasm detection modules or contrastive training on sarcastic examples; (2) real-time knowledge verification against external sources; (3) adaptive reward shaping based on user hesitation signals; and (4) explicit conflict resolution mechanisms for multimodal fusion.