AFFECTVERSE: EMOTIONAL WORLD MODELS FOR MULTIMODAL AFFECTIVE COMPUTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans infer emotions by integrating observed multimodal cues with expectations about how affective states may unfold. Existing multimodal large language models (MLLMs), however, often treat emotion recognition as static fusion over complete audiovisual-text inputs, leaving affective dynamics implicit. We propose AffectVerse, a Qwen2.5-Omni-based model equipped with a lightweight Emotion World Module (EWM), an action-free representation-level module for short-horizon latent affective prediction. EWM contains three modules: Cross-Modal Temporal Imagination predicts future video/audio representations from past tokens with multi-step rollout; MAMA Belief Aggregation compresses imagined tokens into modality-aware belief tokens; and Belief Injection inserts these belief tokens into the LLM for affective reasoning. AffectVerse uses future prediction as a past-conditioned self-supervised signal: the belief state is trained to encode transition cues that predict subsequent affective change, requiring no future frames or audio at inference. Across nine benchmarks, AffectVerse improves the full-modality average by 3.40 points over its Qwen2.5-Omni-7B baseline under identical data, prompts, and training—2.57 points over the strongest prior MLLM—while controlled ablations show additive gains from temporal imagination, cross-modal rollout, and belief aggregation. These results suggest predictive belief-state modeling is a practical alternative for affective computing.

1 INTRODUCTION

Consider a job interview in which a candidate's voice begins to quiver almost imperceptibly and their smile tightens into a mask-like expression. A human observer does not wait for the full utterance to conclude—they update a belief state over how the candidate's affect may evolve. This form of *predictive emotional inference* is well documented in affective science Friston (2010); Barrett (2017), yet current MLLMs often treat emotion recognition as static pattern matching over a complete clip.

This mismatch highlights a gap. Existing MLLM-based methods Lian et al. (2025a); Ye et al. (2023); Xie et al. (2024) project multimodal features into a language model's embedding space and rely on the LLM's general reasoning ability to classify emotions. While effective, this paradigm often reduces affective dynamics—such as escalation, recovery, or cross-modal conflict—to a static snapshot. Future prediction complements past-context modeling: it provides a past-conditioned objective that encourages the representation to encode cues predictive of subsequent affective change.

Our key insight is that emotion understanding should not be formulated merely as static recognition over a complete input, but as predictive belief-state modeling under partial observations. World-model-style predictive representation learning has been effective in reinforcement learning Ha & Schmidhuber (2018); Hafner et al. (2020) and robotics, but affective inference requires an action-free formulation without rewards, policies, or control planning. AffectVerse instantiates a representation-level variant for affective inference: it predicts latent future audiovisual representations from observed prefixes and uses the learned transition regularities to update the LLM's belief state.

We present **AffectVerse**, whose EWM instantiates the three main blocks in Figure 2. Each block is tied to a distinct principle from affective science and motivated by converging evidence that emotion is not a fixed perceptual category, but a temporally unfolding process shaped by predictive inference, appraisal dynamics, bodily expression, multisensory evidence, and contextual interpretation Friston (2010); Barrett (2017); Scherer (2009); Ernst & Banks (2002) :

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

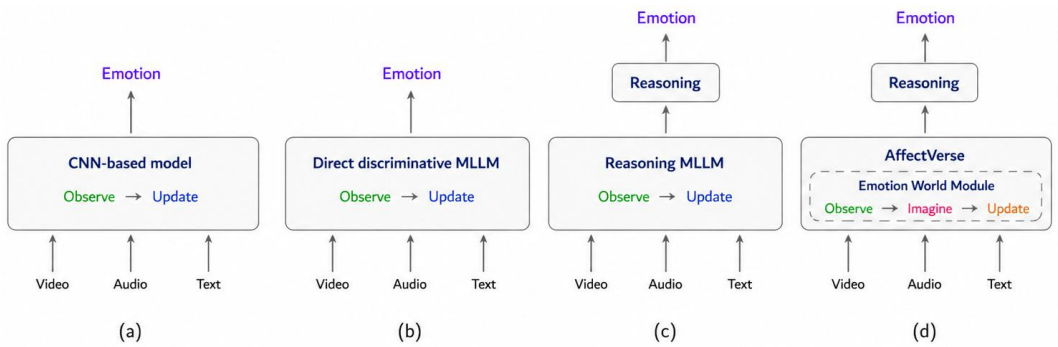

Figure 1: **Motivation and positioning of AffectVerse.** Existing CNN-based models, direct discriminative MLLMs, and reasoning MLLMs mainly follow an Observe–Update pattern. AffectVerse introduces an Emotion World Module that inserts an intermediate Imagine stage, enabling the model to predict latent affective dynamics before updating the LLM's emotional context.

1. **Cross-Modal Temporal Imagination.** Grounded in predictive processing Friston (2010); Clark (2013); Barrett (2017) and multisensory integration Ernst & Banks (2002), EWM predicts future video/audio representations from past tokens with cross-modal context, learnable future queries, multi-step rollout, and modality dropout.
2. **MAMA Belief Aggregation.** Inspired by Scherer's Component Process Model Scherer (2009), MAMA aggregates imagined video/audio tokens into compact belief tokens using type embeddings, dynamic belief queries, and boundary-state residual injection.
3. **Belief Injection.** The belief tokens are projected back to the LLM hidden space and inserted at two positions in the token sequence, allowing Qwen2.5-Omni to reason over both observed evidence and imagined affective continuations.

2 Related Work

2.1 Multimodal Emotion Recognition

Multimodal emotion recognition (MER) identifies human emotional states by jointly leveraging visual, acoustic, and linguistic cues Poria et al. (2019); Zadeh et al. (2018). The field has progressed from modality-specific encoders with hand-crafted fusion Zadeh et al. (2016) to attention-based fusion strategies and, more recently, MLLM-based approaches. Methods such as AffectGPT Lian et al. (2025a), mPLUG-Owl Ye et al. (2023), AffectGPT-R1 Lian et al. (2025c), and Emotion-LLaMA v2 Peng et al. (2026) leverage LLMs as unified multimodal reasoners, while recent fusion-oriented work explores modality complementarity Huang et al. (2026). However, these methods primarily focus on feature alignment or fusion, without explicitly learning how emotional states evolve over time. AffectVerse instead equips the MLLM with an action-free predictive affective module that predicts latent affective transitions and constructs evolving belief states from partial observations.

2.2 World Models

World models, originating from model-based reinforcement learning, learn internal representations of environment dynamics for future state prediction and planning. From compressed latent representations Ha & Schmidhuber (2018) to Dreamer V2/V3 Hafner et al. (2020; 2023), imagination-based training has shown strong potential in learning predictive dynamics for video prediction, autonomous driving, and embodied AI. The core appeal lies in predictive representation learning: instead of only reacting to current observations, a model learns latent transition structure that supports more robust downstream reasoning. EWM brings this predictive representation learning to affect: an action-free module—no rewards, policies, or control planning involved—that predicts short-horizon latent audiovisual continuations and converts them into belief tokens for emotion interpretation.

3 Method

As illustrated in Figure 1, the method is motivated by the gap between static multimodal recognition and temporally unfolding affect. Figure 2 details the three modules: Cross-Modal Temporal Imagination predicts future audiovisual representations, MAMA Belief Aggregation distills them into compact belief tokens, and Belief Injection inserts the tokens into the LLM sequence.

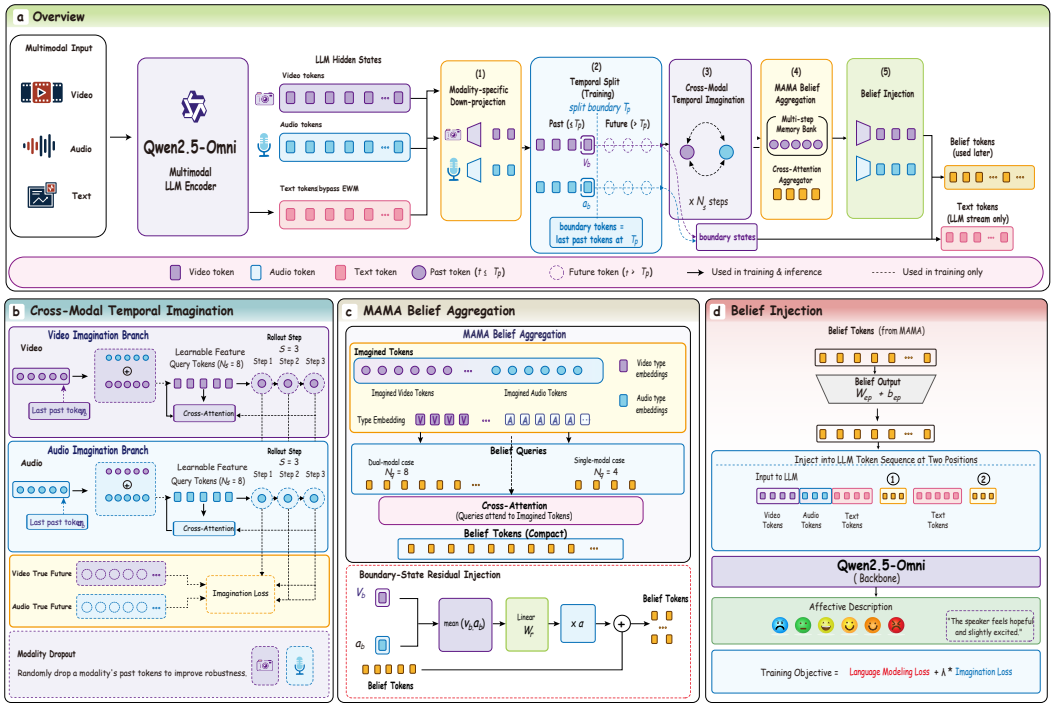

Figure 2: **Overall framework of AffectVerse.** AffectVerse extracts audiovisual hidden states from Qwen2.5-Omni, temporally splits them at T_p, imagines future latent tokens with cross-modal multi-step rollout, aggregates imagined tokens through MAMA with boundary tokens $\mathbf{v}_b/\mathbf{a}_b$, and interleaves the resulting belief tokens into the LLM sequence for affective generation.

3.1 Multimodal LLM Backbone

We build upon Qwen2.5-Omni Xu et al. (2025), a multimodal LLM that natively processes video, audio, and text in a unified transformer architecture. The model's *thinker* module tokenizes each modality—video tokens $\{v_1, \ldots, v_M\}$, audio tokens $\{a_1, \ldots, a_N\}$, and text tokens $\{t_1, \ldots, t_K\}$—and processes them jointly to produce hidden states:

$$\mathbf{H} = \text{Thinker}([\mathbf{v}_{1:M}; \mathbf{a}_{1:N}; \mathbf{t}_{1:K}]) \in \mathbb{R}^{L \times d}, \tag{1}$$

where L is the total sequence length (including formatting tokens) and d=3584. We apply LoRA Hu et al. (2022) (r=16, α=32) to all attention projections (Q, K, V, O) for parameter-efficient fine-tuning, preserving pretrained multimodal capabilities while adapting to emotion-specific reasoning.

3.2 Emotion World Module

The EWM is the core module of AffectVerse. It is grounded in predictive processing framework from neuroscience Friston (2010); Clark (2013).: the brain maintains generative models that continuously predict sensory inputs, and emotions are actively constructed through this predictive inference process Barrett (2017). As shown in Figure 2, EWM consists of Cross-Modal Temporal Imagination, MAMA Belief Aggregation, and Belief Injection, which respectively predict latent affective continuations, compact them into belief tokens, and insert those tokens into the LLM context.

Preparation: Bottleneck Projection and Temporal Splitting. The predictive processing framework holds that the brain must infer hidden states from incomplete sensory evidence Friston (2010). We operationalize this by establishing partial observations through temporal splitting. Given hidden states $\mathbf{H}$, we extract video tokens $\mathbf{H}_v \in \mathbb{R}^{M \times d}$ and audio tokens $\mathbf{H}_a \in \mathbb{R}^{N \times d}$, and project them into the working space via modality-specific down-projections:

$$\mathbf{Z}_v = \mathbf{W}_v^{\downarrow}\mathbf{H}_v \in \mathbb{R}^{M \times d_w}, \quad \mathbf{Z}_a = \mathbf{W}_a^{\downarrow}\mathbf{H}_a \in \mathbb{R}^{N \times d_w}. \tag{2}$$

This bottleneck reduces computation while acting as an Information Bottleneck Tishby et al. (2000).

Each modality is then split temporally into *past* and *future* portions at the split boundary $T_{p,m} = \lfloor \kappa \cdot L_m \rfloor$ (shown as T_p in Figure 2), where κ is the keep ratio. We denote the last retained video and

audio boundary tokens as $\mathbf{v}_b = \mathbf{Z}_v[T_{p,v}-1]$ and $\mathbf{a}_b = \mathbf{Z}_a[T_{p,a}-1]$, respectively. The future portion $\mathbf{Z}^{\text{fut}}$ serves as a detached self-supervised target for the imagination loss, so the prediction task teaches EWM to infer latent affective continuations from observed prefixes without requiring future tokens at inference. We employ a random keep ratio strategy $\kappa \sim \text{Uniform}[0.7, 1.0)$ during training. During inference, κ=1.0 and all observed audiovisual tokens are retained; EWM still provides additional belief-state context learned from the predictive objective.

Cross-Modal Temporal Imagination. A key principle of both world models and predictive processing is *incremental prediction*: the brain does not predict the entire future at once but extends predictions step-by-step Friston (2010). In the affective domain, this corresponds to the temporal unfolding of emotional appraisal Scherer (2009). Furthermore, the brain leverages complementary cross-modal information to improve prediction accuracy.

Each modality has an independent imagination module (unshared parameters), but during prediction each modality attends to both its own and the other modality's past, distinguished by learnable tag embeddings $\mathbf{e}_{\text{self}}, \mathbf{e}_{\text{cross}} \in \mathbb{R}^{d_w}$:

$$\mathbf{C}_v^{(1)} = [(\mathbf{Z}_v^{\text{past}} + \mathbf{e}_{\text{self}});\ (\mathbf{Z}_a^{\text{past}} + \mathbf{e}_{\text{cross}})]. \tag{3}$$

Symmetrically, audio imagination uses $\mathbf{C}_a^{(1)} = [(\mathbf{Z}_a^{\text{past}} + \mathbf{e}_{\text{self}}); (\mathbf{Z}_v^{\text{past}} + \mathbf{e}_{\text{cross}})]$. When only one modality is available, the context reduces to the self-modal past alone. We do not enforce explicit timestamp-level alignment between audio and video tokens in EWM. Instead, we rely on Qwen2.5-Omni's native multimodal encoding to provide implicit cross-modal temporal correspondence, and perform temporal splitting within each modality stream at the same keep ratio κ.

We perform S=3 rollout steps using N_s=8 learnable future query tokens per step. At each step s:

$$\mathbf{F}^{(s)} = \text{CrossAttn}(\mathbf{f}_{1:N_s} + \mathbf{e}_{\text{step}}^{(s)},\ \mathbf{C}_m^{(s)}), \quad \mathbf{C}_m^{(s+1)} = [\mathbf{C}_m^{(1)};\ \hat{\mathbf{Y}}_m^{(s)}], \tag{4}$$

where $\hat{\mathbf{Y}}_m^{(s)} = \mathbf{W}_{\text{out}}\mathbf{F}^{(s)}$ is the imagined output appended to the context for the next step, progressively extending the imagination horizon. Here, $\mathbf{C}_m^{(s)}$ denotes the rollout context for target modality $m \in \{v, a\}$ at step s; using m keeps the equation modality-agnostic, with $\mathbf{C}_v^{(s)}$ and $\mathbf{C}_a^{(s)}$ instantiated by the video and audio imagination branches respectively. All rollout steps share cross-attention parameters, with step embeddings $\mathbf{e}_{\text{step}}^{(s)}$ providing temporal distinction, this encourages a general temporal prediction function. Each step's output is supervised against real future tokens:

$$\mathcal{L}_{\text{imagine}}^{(s)} = \text{MSE}(\hat{\mathbf{Y}}_m^{(s)}, \bar{\mathbf{Z}}_m^{\text{fut}}) + \tfrac{1}{2}(1 - \cos(\hat{\mathbf{Y}}_m^{(s)}, \bar{\mathbf{Z}}_m^{\text{fut}})), \tag{5}$$

where $\bar{\mathbf{Z}}_m^{\text{fut}}$ is obtained by adaptive pooling the real future tokens to match N_s. Concretely, for future sequence $\mathbf{Z}_m^{\text{fut}} \in \mathbb{R}^{L_m^{\text{fut}} \times d_w}$, we apply one-dimensional adaptive average pooling along the temporal axis to produce exactly N_s bins: $\bar{\mathbf{Z}}_m^{\text{fut}} = \text{AdaptiveAvgPool1D}_{N_s}(\mathbf{Z}_m^{\text{fut}})$. This preserves the coarse temporal ordering of the future segment while matching the number of learnable future queries. The combined MSE (magnitude alignment) and cosine (directional alignment) losses prevent collapse to mean representations while maintaining semantic consistency. Details are in Appendix D.

MAMA Belief Aggregation. Scherer's Component Process Model (CPM) Scherer (2009) posits that the appraisal system independently evaluates different stimulus aspects before integration. MAMA operationalizes this "appraise independently, integrate later" principle: the imagination modules (Stage 2) independently produce modality-specific multi-step predictions, each rollout step is supervised by the imagination loss, and MAMA integrates the multi-step imagined tokens into a unified belief state. The rollout outputs are assembled into a multi-step memory bank with *type embeddings* distinguishing modality identity:

$$\mathbf{M} = [\hat{\mathbf{Y}}_v^{(1:S)};\ \hat{\mathbf{Y}}_a^{(1:S)}] + \mathbf{E}_{\text{type}}(\text{type_ids}) \in \mathbb{R}^{N_M \times d_w}, \tag{6}$$

where $N_M = |\mathcal{S}|SN_s$ for available modality set $\mathcal{S}$. A key design choice is *dynamic belief capacity*: the number of belief query tokens adapts to available modalities (N_q=8 for dual-modality, N_q=4 for single-modality), with *separate* aggregation modules for each regime. This ensures richer inputs receive proportionally larger representation. Belief queries attend to the memory via cross-attention:

$$\mathbf{b}^{(\text{out})} = \text{CrossAttn}(\mathbf{b}_{1:N_q},\ \mathbf{M}), \tag{7}$$

augmented by a *residual boundary injection* that preserves context at the temporal split point: $\mathbf{b}^{(\text{final})} = \mathbf{b}^{(\text{out})} + \alpha \cdot \mathbf{W}_r \bar{\mathbf{z}}_{\text{bnd}}$, where $\bar{\mathbf{z}}_{\text{bnd}} = \text{mean}(\mathbf{v}_b, \mathbf{a}_b)$ is computed from the last past video/audio tokens at T_p (using available modalities only) and α is a learnable scalar initialized to 0.1.

Belief Injection After belief aggregation, the compact belief state is projected back to the LLM's hidden dimension and distributed across the token sequence. The belief tokens are up-projected via $\mathbf{B}_{\text{inject}} = \text{LayerNorm}(\mathbf{W}_{ep}\mathbf{b}^{(\text{final})} + \mathbf{b}_{ep}) \in \mathbb{R}^{N_q \times d}$. Rather than inserting all tokens at a single position, we split $\mathbf{B}_{\text{inject}}$ into two halves injected at: (1) immediately after the last audiovisual token, enriching the text encoding with affective context from the imagined future; and (2) immediately before the answer region, providing a direct "emotional prior" for response generation. This interleaved strategy ensures that the LLM can attend to belief information at multiple stages of reasoning.

During training, audiovisual tokens are truncated according to κ by applying a token-level keep mask to the embedded sequence: the removed video/audio positions are dropped before the LM loss is computed, and the attention mask and labels are rebuilt after padding the shortened batch. This partial-observation regularization encourages the LLM to attend to the injected belief state instead of treating it as unused extra context. Belief token labels are set to -100 (ignored in loss computation) and their attention mask is set to 1. During inference (κ=1.0), no truncation occurs—all audiovisual tokens are retained alongside the belief tokens, so AffectVerse never discards available evidence at test time. Implementation details are in Appendix D.

3.3 Training Objectives

Detached future audiovisual hidden states supervise EWM to learn affective transitions, while belief-token-augmented partial sequences train the LLM to use predictive belief representations.

AffectVerse is trained end-to-end with two complementary objectives. Let $\mathbf{H}_{\text{augmented}}$ denote the LLM hidden-token sequence after interleaved belief injection. Given the original hidden sequence $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_L]$, we split the projected belief tokens as $\mathbf{B}_{\text{inject}} = [\mathbf{B}^{(1)}_{\text{inject}}; \mathbf{B}^{(2)}_{\text{inject}}]$ and insert them at the AV–Text boundary p_{AV} and the Text–Answer boundary p_{ans}. The **language modeling loss** is autoregressive cross-entropy on assistant response tokens:

$$\mathbf{H}_{\text{augmented}} = [\mathbf{H}_{1:p_{\text{AV}}}; \mathbf{B}^{(1)}_{\text{inject}}; \mathbf{H}_{p_{\text{AV}}+1:p_{\text{ans}}-1}; \mathbf{B}^{(2)}_{\text{inject}}; \mathbf{H}_{p_{\text{ans}}:L}]. \tag{8}$$

$$\mathcal{L}_{\text{lm}} = -\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \log p(y_i | y_{<i}, \mathbf{H}_{\text{augmented}}). \tag{9}$$

The imagination loss(Eq. 5) is averaged over all S rollout steps and all available modalities $\mathcal{S}$. The total loss balances task performance and temporal prediction:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{lm}} + \lambda_{\text{img}} \mathcal{L}_{\text{imagine}}, \quad \lambda_{\text{img}} = 1.0. \tag{10}$$

The equal weighting $\lambda_{\text{img}} = 1.0$ treats imagination as a co-primary objective alongside language modeling—the imagination loss drives the EWM to produce meaningful belief states, while the language modeling loss ensures task-specific performance. We ablate this choice in Table 6.

4 Experiments

We evaluate AffectVerse on nine benchmarks: MER2023/MER2024 Lian et al. (2024), MELD Poria et al. (2019), IEMOCAP Busso et al. (2008), CMU-MOSI Zadeh et al. (2016), CMU-MOSEI Zadeh et al. (2018), CH-SIMS Yu et al. (2020), CH-SIMS v2 Liu et al. (2022), and OV-MERD+ Lian et al. (2025b;a). We report weighted F1 for emotion classification and binary accuracy for sentiment analysis, following MER-UniBench Lian et al. (2025a); details are in Appendix A.

4.1 Main Results

Table 1 presents the main results across nine benchmarks under three modality settings: audio+text (A+T), video+text (V+T), and audio+video+text (A+V+T). "AffectVerse" denotes our proposed model. Table 1 shows consistent gains for AffectVerse across modality settings. In full A+V+T, AffectVerse reaches **77.34 %** average and ranks first on 8/9 benchmarks, exceeding Qwen2.5-Omni (74.76%) and AffectGPT (74.77%) by 2.57 points. Under the controlled same-backbone protocol—identical backbone, data, prompts, and training configuration—the contribution of EWM

Table 1: **Main results under different modality settings.** A/V/T denote audio, video, and text. We report F1 for emotion classification and binary accuracy for sentiment analysis; Mean is averaged over nine benchmarks. **Bold** indicates the best result, and underline indicates the second-best result.

Model	Modality			Basic				Sentiment				Fine-grained	Mean
	A	V	T	MER2023	MER2024	MELD	IEMOCAP	MOSI	MOSEI	SIMS	SIMS v2	OV-MERD+	
OneLLM Han et al. (2024)	✓	×	✓	25.52	17.21	28.32	33.44	64.01	54.09	63.39	61.98	22.25	41.14
SECap Xu et al. (2024)	✓	×	✓	40.95	52.46	25.56	36.92	55.76	54.18	59.51	57.41	36.97	46.64
PandaGPT Su et al. (2023)	✓	×	✓	33.57	39.04	31.91	36.55	66.06	61.33	62.93	58.88	31.33	46.84
Qwen-Audio Chu et al. (2023)	✓	×	✓	41.85	31.61	49.09	35.47	70.09	46.90	70.73	65.26	32.36	49.26
Qwen2.5-Omni **?**	✓	×	✓	70.13	74.17	54.64	58.60	85.06	79.22	82.79	83.53	61.61	72.19
SALMONN Tang et al. (2024)	✓	×	✓	55.53	45.38	45.62	46.84	81.00	67.03	68.69	65.93	45.00	57.89
AffectGPT Lian et al. (2025a)	✓	×	✓	72.94	73.41	56.63	55.68	83.46	80.74	82.99	83.75	59.98	72.18
AffectVerse	✓	×	✓	**75.27**	**76.59**	**58.32**	**64.32**	**85.38**	**81.25**	**83.53**	**84.42**	**62.87**	**74.66**
Otter Li et al. (2025a)	×	✓	✓	16.41	14.65	22.57	29.08	52.89	50.44	57.56	53.12	16.63	34.82
Video-LLaVA Lin et al. (2023)	×	✓	✓	36.93	30.25	30.73	38.95	56.37	61.64	53.28	57.45	34.00	44.40
PandaGPT Su et al. (2023)	×	✓	✓	39.13	47.16	38.33	47.21	58.50	64.25	62.07	65.25	35.07	50.77
Video-ChatGPT Maaz et al. (2024)	×	✓	✓	44.86	46.80	37.33	56.83	54.42	63.12	64.82	65.80	39.80	52.64
VideoChat2 Li et al. (2024a)	×	✓	✓	33.67	54.50	36.64	48.70	66.84	54.32	69.49	70.66	39.21	52.67
LLaMA-VID Li et al. (2024b)	×	✓	✓	50.72	57.60	42.75	46.02	61.78	63.89	69.35	67.48	45.01	56.07
VideoChat Li et al. (2025b)	×	✓	✓	48.73	57.30	41.11	48.38	65.13	63.61	69.52	72.14	44.52	56.71
Chat-UniVi Jin et al. (2024)	×	✓	✓	57.62	65.67	45.61	52.37	54.53	63.18	68.15	66.36	48.00	57.94
mPLUG-Owl Ye et al. (2023)	×	✓	✓	56.86	59.89	49.11	55.54	72.40	72.91	72.13	75.00	48.18	62.45
Qwen2.5-Omni **?**	×	✓	✓	71.14	71.97	54.48	58.27	79.02	77.18	84.79	84.13	61.49	71.39
AffectGPT Lian et al. (2025a)	×	✓	✓	74.58	75.29	**57.63**	**62.19**	82.39	**81.57**	**87.20**	**86.29**	61.65	**74.31**
AffectVerse	×	✓	✓	**75.28**	**76.35**	56.38	60.78	**82.58**	80.78	84.76	85.00	**64.84**	74.08
PandaGPT Su et al. (2023)	✓	✓	✓	40.21	51.89	37.88	44.04	61.92	67.61	68.38	67.23	37.12	52.92
Emotion-LLaMA Cheng et al. (2024)	✓	✓	✓	59.38	73.62	46.76	55.47	66.13	67.66	78.32	77.23	52.97	64.17
Qwen2.5-Omni **?**	✓	✓	✓	77.43	77.35	55.29	60.51	83.42	81.54	85.70	85.53	63.36	74.76
AffectGPT Lian et al. (2025a)	✓	✓	✓	78.54	78.80	55.65	60.54	81.30	80.90	**88.49**	86.18	62.52	74.77
AffectVerse	✓	✓	✓	**80.92**	**82.39**	**58.47**	**63.87**	**84.45**	**86.60**	86.60	**86.30**	**66.50**	**77.34**

Table 2: **Modality/temporal robustness ablation on eight benchmarks.** Three settings: Full (A+V+T), No Audio (V+T), No Video (A+T), evaluated across keep ratios. Left: sentiment benchmarks (Acc-2, %); Right: classification benchmarks (WA-F1, %). CMU-MOSI is fully measured.

Dataset	Setting	1.0	0.7	0.5	0.3	0.1	Avg.
CMU-MOSI	Full	84.5	82.7	82.7	82.7	**84.3**	83.4
	No Audio	82.6	81.1	81.0	81.4	81.1	81.4
	No Video	**85.4**	**86.0**	**84.9**	**86.0**	84.2	85.3
CMU-MOSEI	Full	**86.6**	85.8	85.4	85.2	85.0	85.6
	No Audio	80.8	80.1	79.7	79.3	79.1	79.8
	No Video	81.3	80.6	80.1	79.8	79.5	80.3
SIMS	Full	**86.6**	84.2	83.5	83.0	82.5	84.0
	No Audio	84.8	83.4	82.8	82.3	81.9	83.0
	No Video	83.5	84.8	83.2	82.7	82.3	83.3
SIMS v2	Full	**86.3**	85.6	85.1	84.7	84.4	85.2
	No Audio	85.0	84.3	83.9	83.5	83.1	84.0
	No Video	84.4	84.4	84.0	83.6	83.2	83.9
Avg.	Full	86.0	84.6	84.2	83.9	84.1	84.6
	No Audio	83.3	82.2	81.9	81.6	81.3	82.1
	No Video	83.7	84.0	83.1	83.0	82.3	83.2

Dataset	Setting	1.0	0.7	0.5	0.3	0.1	Avg.
MER2023	Full	**80.9**	80.1	79.5	79.0	78.6	79.6
	No Audio	75.3	74.9	74.6	74.3	74.0	74.6
	No Video	75.3	76.0	74.9	74.6	74.2	75.0
MER2024	Full	82.4	**82.9**	82.0	81.5	81.1	82.0
	No Audio	76.4	75.9	75.5	75.2	74.9	75.6
	No Video	76.6	76.9	76.2	75.9	75.5	76.2
MELD	Full	**58.5**	57.9	57.5	57.2	56.8	57.6
	No Audio	56.4	56.0	55.7	55.4	55.1	55.7
	No Video	58.3	57.9	57.5	57.2	56.9	57.6
IEMOCAP	Full	63.9	63.8	63.4	63.1	62.7	63.4
	No Audio	60.8	60.9	60.5	60.1	59.8	60.4
	No Video	**64.3**	60.3	59.5	58.9	58.3	60.3
Avg.	Full	71.4	71.2	70.6	70.2	69.8	70.6
	No Audio	67.2	66.9	66.6	66.3	66.0	66.6
	No Video	68.6	67.8	67.0	66.7	66.2	67.3

is **+3.40 points** (73.66→77.06, Table 4), isolating the module's effect from backbone strength. The largest gains appear on CMU-MOSEI (+5.06 over Qwen2.5-Omni; +5.70 over AffectGPT) and MER2024 (+3.59% over AffectGPT), indicating improved modeling of temporally evolving affect. In dual-modality settings, A+T (74.66%) is stronger than V+T (74.08%), consistent with prosody carrying stronger emotional signals than visual appearance Barrett (2017); even in V+T, AffectVerse gains +2.69 over its same-backbone counterpart, and the full-modality setting—where cross-modal imagination has both streams to bridge—remains AffectVerse's largest margin. On fine-grained OV-MERD+, AffectVerse improves to 66.50% (+3.98% over AffectGPT), suggesting that belief-state injection is particularly helpful for subtle categories. We defer temporal truncation and missing-modality analysis to the ablation section.

4.2 Ablation Studies

Modality/Temporal Robustness. Table 2 shows graceful degradation under temporal truncation and missing modalities. Full-modality performance remains stable as κ decreases; on CMU-MOSI, accuracy stays in [82.65%, 84.45%] from κ=1.0 to 0.1, and the no-video setting peaks at κ=0.7 (86.01%), higher than κ=1.0 (85.38%). MER2024 exhibits the same trend (82.85% vs. 82.39%), indicating that moderate truncation acts as an implicit regularizer for cross-modal learning. At κ=1.0, removing audio causes larger drops than removing video on most datasets (e.g., MER2024: -6.04%

Table 3: **Component ablation.** Each row adds one component to the previous configuration. Δ shows the gain over the LoRA baseline.

Configuration	MOSI	MOSEI	MER2024	MELD	SIMS	IEMOCAP	ΔAvg.
(a) Qwen2.5-Omni + LoRA	81.58	81.54	77.35	55.29	85.70	60.51	—
(b) + Temporal Splitting	83.78	82.38	78.92	56.14	85.89	61.07	+0.87
(c) + Cross-Modal Imagination	84.05	84.20	80.46	57.25	86.12	62.15	+1.87
(d) + MAMA (single-position inject)	84.21	85.52	81.58	57.82	86.30	62.98	+2.44
(e) + Interleaved Injection (2 pos.)	84.38	86.25	82.15	58.21	86.48	63.52	+2.89
(f) + Residual Boundary Injection	84.45	86.60	82.39	58.47	86.60	63.87	**+3.07**

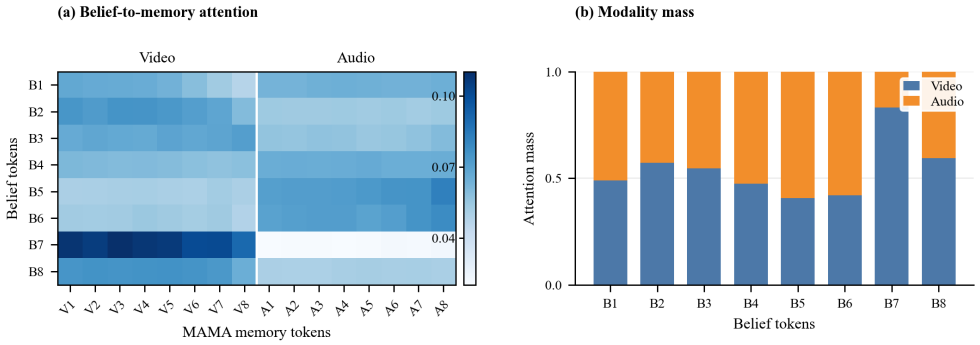

Figure 3: **Visualization of MAMA belief aggregation.** MAMA belief tokens attend non-uniformly to visual and acoustic memories: some integrate both modalities, while others specialize toward visual (e.g., $B7$) or acoustic evidence (e.g., $B5$–$B6$). This suggests learned modality-aware specialization from type-aware aggregation, rather than manually assigned token roles.

vs. −5.80%; CMU-MOSEI: −5.82% vs. −5.35%), consistent with the dominant role of prosody. At the same time, A+V+T still wins on most benchmarks, confirming cross-modal complementarity. Under missing-modality evaluation, degradation remains limited (MELD no-video: 58.32% vs. 58.47%), supporting the claim that belief-state modeling improves robustness to partial observations along modality and temporal axes. The relative advantage over the same-backbone baseline grows from +3.40 to +5.52 as observations shrink; Appendices E and F provide additional controlled experiments and full reproducibility details.

4.3 Mechanism Analysis

Component Ablation. Table 3 isolates the contribution of each proposed component by incrementally building from a LoRA-tuned Qwen2.5-Omni baseline to the full AffectVerse. Table 3 yields three direct conclusions. First, temporal prediction alone is effective: row (b) improves the baseline by +0.87 average and improves all six benchmarks. Second, cross-modal imagination is the largest single jump (row (c), +1.00% over row (b)), with strong gains on CMU-MOSEI (+1.82%) and MER2024 (+1.54%). Third, belief construction and injection are both necessary: MAMA, interleaved injection, and residual boundary injection add consistent gains from rows (d)–(f).

EWM Module Ablation. Table 4 ablates the belief construction inside EWM. Random belief tokens and pooling improve the LoRA baseline, but both remain below the EWM. This indicates that EWM's predictive belief construction contributes beyond added token capacity or simple temporal pooling.

Table 4: **EWM module ablation.** Δ is measured against the LoRA baseline.

Belief construction	MOSI	MOSEI	MER24	MELD	SIMS	IEMO	Avg.	Δ
None	81.58	81.54	77.35	55.29	85.70	60.51	73.66	0.00
Random tokens	82.10	82.20	77.80	55.70	85.85	60.90	74.09	+0.43
Pooling	83.05	83.40	79.10	56.55	86.05	61.70	74.97	+1.31
EWM	**84.45**	**86.60**	**82.39**	**58.47**	**86.60**	**63.87**	**77.06**	**+3.40**

MAMA Attention Visualization. To understand what MAMA learns internally, Figure 3 visualizes the cross-attention from the learned belief queries to the final imagined memory tokens on CMU-MOSI validation examples. The left panel shows belief-to-memory attention, where $B1$–$B8$ denote dual-modality belief tokens and $V1$–$V8$ / $A1$–$A8$ denote video and audio memory tokens. The right panel aggregates each row into modality-level attention mass. Figure 3 suggests that MAMA does more than average multimodal features. Learned belief queries form complementary appraisal slots, either integrating both modalities or focusing on modality-specific evidence. This supports MAMA's goal of building a modality-aware belief state.

Table 5: **Cross-modal vs. self-modal imagination.** Comparison of different imagination context modes. Avg Cos: average imagination cosine similarity.

Imagination Mode	Avg Cos ↑	MOSI	MOSEI	MER2024	MELD	SIMS	IEMOCAP	Avg.
Self-modal only	0.762	83.78	82.38	78.92	56.14	85.89	61.07	74.70
Cross-modal only	0.741	83.52	83.15	79.68	56.82	85.95	61.85	75.16
Full cross-modal	**0.798**	**84.45**	**86.60**	**82.39**	**58.47**	**86.60**	**63.87**	**77.06**

Imagination Depth: Rollout Steps. Figure 4 evaluates the effect of rollout depth S and relates it to imagination fidelity. Performance improves from $S{=}1$ to $S{=}3$ (+1.25% average), then declines for $S{\geq}4$ (-0.09% at $S{=}4$, -0.31% at $S{=}5$). This indicates that shallow rollouts underfit temporal dynamics, while overly deep rollouts accumulate prediction error. Imagination quality decreases monotonically with depth (cosine 0.812→0.781; MSE 0.041→0.055), while downstream performance peaks at $S{=}3$. Therefore, the best classifier is not the model with the highest one-step fidelity; it is the model with the best trade-off between temporal coverage and prediction noise. The asymmetric drift, with a larger relative change in MSE than cosine, also justifies combining magnitude and directional losses.

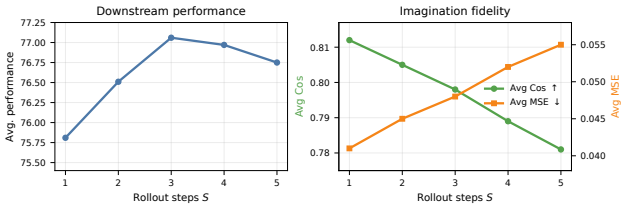

Figure 4: **Rollout depth trade-off.**

Belief Token Capacity. Figure 5 studies the impact of the base belief token count N_q^0, where dual-modality uses $2N_q^0$ tokens, probing how much capacity is needed to encode emotional appraisal. The relation is inverted-U. Too few tokens underfit belief capacity ($N_q^0{=}1$, -1.79% avg), especially on CMU-MOSEI and IEMOCAP. Too many tokens dilute signal and increase sequence burden ($N_q^0{=}16$, -0.68% avg). The optimum is $N_q^0{=}4$, corresponding to 8 tokens under dual-modality.

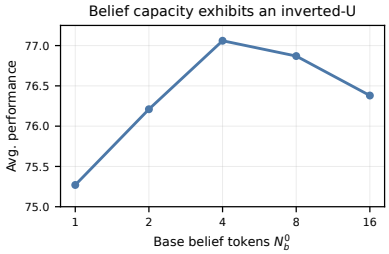

Figure 5: **Effect of belief token count.**

Cross-Modal vs. Self-Modal Imagination. The cross-modal imagination design is a key differentiator of AffectVerse: rather than predicting each modality's future in isolation, each modality attends to both its own and the other modality's past. Table 5 isolates this contribution by comparing three context modes.

The results show that full cross-modal imagination outperforms self-modal only by +2.36% average points, with gains on CMU-MOSEI (+4.22%) and MER2024 (+3.47%) where audiovisual complementarity is pronounced. Cross-modal *only* mode (attending exclusively to the other modality) slightly beats self-modal only (+0.46% avg), indicating that the other modality's temporal history carries predictive value. However, its lower imagination quality (Cos 0.741 vs. 0.762) reveals a trade-off: the other modality provides complementary but less direct evidence for predicting a modality's own future. The full cross-modal mode combines both signals through learnable tag embeddings ($\mathbf{e}_{\text{self}}, \mathbf{e}_{\text{cross}}$), achieving the highest imagination quality (Cos 0.798) and the best downstream performance. This is consistent with the *maximum likelihood estimation* principle in multisensory integration Ernst & Banks (2002), where channels are weighted by reliability.

Modality Dropout Ratio. Modality dropout ($p{=}0.15$) trains the EWM to form beliefs from partial sensory evidence. Figure 6 reveals the trade-off between full-modality performance and missing-modality robustness within this partial-observation setting. The trade-off is clear. Without dropout ($p{=}0$), full-modality scores are highest (e.g., IEMOCAP: 64.32%), but the missing-modality gap reaches 5.12%, indicating over-reliance on cross-modal cues that may be unavailable at test time. At $p{=}0.30$, the gap shrinks to 0.66%, but full-modality performance drops by 1.95%, suggesting that the model is trained too often in suboptimal single-modality regimes. At $p{=}0.15$, only 0.57% are sacrificed in the full setting, while the missing-modality gap is reduced by

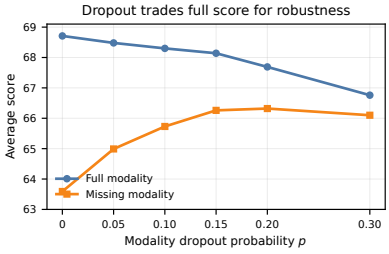

Figure 6: **Effect of modality dropout.**

Table 6: **Effect of imagination loss weight λ_{img}.** We report Avg Cos (imagination quality) and downstream performance on three representative benchmarks plus the six-benchmark average.

λ_{img}	Avg Cos ↑	MOSI	MOSEI	MER2024	MELD	SIMS	IEMOCAP	Avg.
0 (no img loss)	—	83.62	82.75	78.15	55.82	85.72	60.85	74.49
0.01	0.725	83.85	83.92	79.48	56.58	85.95	61.62	75.23
0.1	0.768	84.12	85.38	81.25	57.65	86.28	62.85	76.26
0.5	0.790	84.32	86.18	82.05	58.18	86.48	63.52	76.79
1.0	**0.798**	**84.45**	**86.60**	**82.39**	**58.47**	**86.60**	**63.87**	**77.06**
2.0	0.805	84.28	86.35	82.12	58.25	86.42	63.58	76.83
5.0	0.812	83.82	85.45	81.18	57.42	86.05	62.72	76.11

63%, placing it near the Pareto frontier for modality robustness. This pattern is consistent with findings in sensory neuroscience that moderate uncertainty promotes robust multisensory integration Ernst & Banks (2002).

Imagination Loss Weight λ_{img}. We find λ_{img}=1.0 gives the best downstream average, confirming that imagination should act as a co-primary objective rather than a weak auxiliary loss. Table 6 shows that the imagination objective is necessary. The λ_{img}=0 setting keeps the full EWM architecture and the LM loss but removes future supervision—a capacity-matched LM-only belief module, not a random-token baseline. It drops 2.57 points below λ_{img}=1.0. Increasing λ_{img} improves both imagination quality and downstream performance up to 1.0, while larger weights overemphasize regression and reduce task performance.

Summary. Taken together, the ablations show that AffectVerse's gains emerge from the *synergy* of cross-modal imagination (+2.36 over self-only), structured belief aggregation, and calibrated training strategies. The hyperparameters explored here show nontrivial trade-offs, indicating the model is well-behaved rather than brittle.

4.4 Case Study

Figure 7 shows how AffectVerse's belief evolves as evidence accumulates: the speaker calmly mentions a memorial day (ground truth: neutral), yet AffectGPT, fusing the complete clip in one pass, commits to *sadness, anger, frustration, stress* at every observation ratio, while AffectVerse refines its belief with each observation—from *stress* through sadness-related beliefs to *neutral*, separating affective language about negative events from actual negative affect. An additional high-arousal positive case is in Appendix D.7.



Ground truth: neutral

Subtitle: "June 11th is the memorial day of Shen Qian's boyfriend."

Observed	10%	30%	50%	70%	100%
AffectGPT	*sadness, anger, frustration, stress (identical at all ratios)*				
AffectVerse (ours)	stress	sadness, depression	sadness, heaviness, sorrow	sadness, nostalgia	**neutral** ✓

Figure 7: **Case study of belief-state updating.**

5 Conclusion

We presented AffectVerse, a Qwen2.5-Omni-based model that augments multimodal emotion recognition with a lightweight Emotion World Module composed of Cross-Modal Temporal Imagination, MAMA Belief Aggregation, and Belief Injection. Rather than treating an affective clip as a static input, AffectVerse predicts latent audiovisual continuations from observed evidence and compresses the resulting transition information into belief tokens for the LLM. Across nine benchmarks, AffectVerse improves the full-modality average by 3.40 points over the same Qwen2.5-Omni backbone and 2.57 points over the strongest prior MLLM, with consistent gains on 8/9 benchmarks; controlled ablations show that temporal imagination, cross-modal rollout, MAMA aggregation, and modality-dropout training each contribute to the final performance.

Limitations. AffectVerse is evaluated with fixed rollout and keep-ratio hyperparameters, and its robustness analysis focuses on partial observations rather than broader domain shifts; extending it to stronger alignment and dialogue-level emotion tracking remains future work.

AI USE STATEMENT

In this work, we have not used generative AI tools for any of the tasks with required disclosure, including generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, or interpreting results; those tasks are not applicable to this work. Additionally, we used generative AI tools for the following tasks with recommended disclosure: creating and editing software code, and drafting parts of the manuscript. We have reviewed all AI-assisted work: AI-assisted code was reviewed and tested by the authors, and all AI-assisted text was reviewed and revised by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI. See the ICLR 2027 AI Policy for Authors for more details.

REFERENCES

Lisa Feldman Barrett. The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, 12(1):1–23, 2017.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems 37*, 2024.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2): 127–138, 2010.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Zhexian Huang, Bo Zhao, Hui Ma, Zhishu Liu, Jie Zhang, Ruixin Zhang, Shouhong Ding, and Zitong Yu. Complementarity-supervised spectral-band routing for multimodal emotion recognition. *arXiv preprint arXiv:2603.13340*, 2026. doi: 10.48550/arXiv.2603.13340.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7543–7557, 2025a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. VideoChat: Chat-centric video understanding. *Science China Information Sciences*, 68 (10):200102, 2025b. doi: 10.1007/s11432-024-4321-9.

Yanwei Li, Chengyao Wang, and Jiaya Jia. LLaMA-VID: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, 2024b.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, Jiangyan Yi, Rui Liu, Kele Xu, Bin Liu, Erik Cambria, Guoying Zhao, Björn W. Schuller, and Jianhua Tao. MER 2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, 2024.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025a.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. OV-MER: Towards open-vocabulary multimodal emotion recognition. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025b.

Zheng Lian, Fan Zhang, Yazhou Zhang, Jianhua Tao, Rui Liu, Haoyu Chen, Xiaobai Li, and Bin He. Affectgpt-r1: Leveraging reinforcement learning for open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2508.01318*, 2025c. doi: 10.48550/arXiv.2508.01318.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 247–258, 2022.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Xiaojiang Peng, Jingyi Chen, Zebang Cheng, Bao Peng, Fengyi Wu, Yifei Dong, Shuyuan Tu, Qiyu Hu, Huiting Huang, Yuxiang Lin, Jun-Yan He, Kai Wang, Zheng Lian, and Zhi-Qi Cheng. Emotion-llamav2 and mmeverse: A new framework and benchmark for multimodal emotion understanding. *arXiv preprint arXiv:2601.16449*, 2026. doi: 10.48550/arXiv.2601.16449.

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 the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

Klaus R Scherer. Emotions are emergent processes: they require a dynamic computational architecture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3459–3474, 2009.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, 2024.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. EmoVIT: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26596–26605, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Ke qin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *ArXiv*, abs/2503.20215, 2025. URL `https://api.semanticscholar.org/CorpusID:277322543`.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shixiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. SECap: Speech emotion captioning with large language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19323–19331, 2024.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2236–2246, 2018.

A Detailed Dataset Descriptions

We evaluate AffectVerse on nine multimodal emotion and sentiment benchmarks. Below we provide detailed descriptions of each dataset.

MER2023 Lian et al. (2024). The Multimodal Emotion Recognition Challenge 2023 dataset contains video clips annotated with 6 discrete emotion categories (angry, happy, neutral, sad, worried, surprise). Each sample consists of a short video segment with corresponding audio and textual transcripts. We use the official training and test splits.

MER2024 Lian et al. (2024). The Multimodal Emotion Recognition Challenge 2024 dataset is an updated and extended version of MER2023, featuring the same 6 emotion categories with additional samples and refined annotations. We use the official training set for training and the test set for evaluation.

MELD Poria et al. (2019). The Multimodal EmotionLines Dataset is collected from the TV series *Friends*, containing 7 emotion categories (anger, joy, sadness, neutral, disgust, fear, surprise). Each utterance is associated with video, audio, and textual transcript. We use the official train/dev/test splits.

IEMOCAP Busso et al. (2008). The Interactive Emotional Dyadic Motion Capture dataset consists of approximately 12 hours of conversational videos with 4 emotion categories (happy, sad, neutral, angry). We follow the standard 5-fold protocol, using the first 4 sessions for training and session 5 for testing.

CMU-MOSI Zadeh et al. (2016). The CMU Multimodal Opinion-level Sentiment Intensity dataset contains 2,199 opinion utterance-video segments from YouTube movie reviews, annotated with continuous sentiment scores in $[-3, +3]$. We report binary accuracy (Acc-2) following standard practice.

CMU-MOSEI Zadeh et al. (2018). The CMU Multimodal Opinion Sentiment and Emotion Intensity dataset is a large-scale extension of CMU-MOSI, containing 23,453 annotated video segments from over 1,000 YouTube speakers. Sentiment annotations are provided as continuous scores. We report binary accuracy on the standard test split (4,659 samples).

SIMS. The CH-SIMS (Chinese Multimodal Sentiment Intensity) dataset contains 2,281 video clips from Chinese movies, TV series, and variety shows, annotated with multimodal sentiment intensity scores. It provides both unimodal and multimodal annotations. We report binary accuracy on the official test split.

SIMS v2. SIMS v2 is an extended version of CH-SIMS with improved annotation quality and additional samples. It follows the same annotation protocol as SIMS but with a larger and more diverse collection. We report binary accuracy.

OV-MERD+. The Open-Vocabulary Multimodal Emotion Recognition Dataset Plus (OV-MERD+) is a fine-grained emotion benchmark that goes beyond basic categorical labels, requiring models to recognize a broader and more nuanced set of emotional states. We report weighted-average F1-score on the official test split.

All datasets provide aligned video, audio, and text (subtitle/transcript) for each sample. Each input is formatted as a multi-turn conversation following the Qwen2.5-Omni chat template, with the system prompt "You are an emotion recognition assistant," the user message containing video, audio, subtitle context, and the question, and the assistant response containing the emotion label or sentiment description.

A.1 Evaluation Protocol

We follow the MER-UniBench protocol associated with AffectGPT Lian et al. (2025a) for dataset coverage, modality settings, and primary metrics. The main table includes externally reported results from this benchmark family as reference points, because rerunning all MLLM baselines under identical training recipes is computationally prohibitive. To avoid overclaiming from heterogeneous training histories, we use these external numbers only for broad benchmark positioning. The method-

specific evidence is based on Qwen2.5-Omni and AffectVerse runs in our codebase, together with the controlled ablations that keep the backbone, data format, prompts, and optimization setup fixed.

A.2 Evaluation Metrics

For discrete emotion classification (MER2024, IEMOCAP, MELD), we report **weighted average F1-score (WAF)** as the primary metric, along with overall accuracy. For continuous sentiment regression (CMU-MOSEI, CMU-MOSI), we report **Mean Absolute Error (MAE)**, **Pearson Correlation**, and **binary accuracy (Acc-2)** following standard practice.

A.3 Implementation Details

AffectVerse is built upon Qwen2.5-Omni-7B **?** with LoRA Hu et al. (2022) adaptation ($r = 16$, $\alpha = 32$, dropout $= 0.05$) applied to all attention projection matrices (Q, K, V, O). The Emotion World Module operates in a bottleneck working dimension $d_w = 512$, with $N_s = 8$ future query tokens per rollout step, $S = 3$ rollout steps using 2-layer cross-attention with 8 heads, and $N_q = 4$ base belief query tokens (8 for dual-modality). The keep ratio κ is sampled from Uniform$[0.7, 1.0)$ during training and fixed at 1.0 during inference. Modality dropout probability is 0.15. The imagination loss weight is $\lambda_{\text{img}} = 1.0$. Videos are sampled at 3 FPS and resized to 224×224. The maximum sequence length is 1024 tokens. The total EWM parameter overhead is ∼25M.

Training uses the AdamW optimizer with a cosine-annealed learning rate of 2×10^{-4}, weight decay 0.01, and linear warm-up for the first 3% of steps. The effective batch size is 32 (4 per GPU × 8 gradient accumulation). We train for 3 epochs on the merged training set of all nine benchmarks. All experiments are conducted on 8 NVIDIA A100-80G GPUs; a full training run takes approximately 18 hours. The LoRA modules and EWM are jointly optimized; the backbone parameters remain frozen.

B Detailed architecture specifications

Table 7 summarizes the key architectural hyperparameters of AffectVerse.

C Algorithm

Algorithm 1 provides a pseudocode summary of the AffectVerse training procedure.

Algorithm 1 AffectVerse Training Step

Require: Multimodal input (V, A, T), labels y

1: Tokenize and prepare inputs: input_ids, attn_mask, labels
2: **Full-observation Thinker Forward:** $\mathbf{H} \leftarrow$ Thinker(input_ids) {Used to form detached predictive targets}
3: **Extract modality tokens:** $\mathbf{H}_v, \mathbf{H}_a \leftarrow$ ExtractAV($\mathbf{H}$)
4: **Sample keep ratio:** $\kappa \sim$ Uniform$[\kappa_{\min}, 1.0)$
5: **Modality dropout:** randomly drop one modality with $p = 0.15$
6: **Bottleneck projection:** $\mathbf{Z}_v \leftarrow \mathbf{W}_v^{\downarrow}\mathbf{H}_v$, $\mathbf{Z}_a \leftarrow \mathbf{W}_a^{\downarrow}\mathbf{H}_a$
7: **Temporal split:** $\mathbf{Z}_m^{\text{past}}, \mathbf{Z}_m^{\text{fut}}, \mathbf{z}_m^{\text{bnd}}, T_{p,m} \leftarrow \text{Split}(\mathbf{Z}_m, \kappa)$ for $m \in \{v, a\}$
8: **Cross-modal imagination:** $\hat{\mathbf{Y}}_v^{(1:S)}, \hat{\mathbf{Y}}_a^{(1:S)} \leftarrow$ Imagine($\mathbf{Z}^{\text{past}}$) {Multi-step rollout}
9: **Imagination loss:** $\mathcal{L}_{\text{img}} \leftarrow$ MSE+Cos($\hat{\mathbf{Y}}^{(1:S)}$, stopgrad($\mathbf{Z}^{\text{fut}}$))
10: **MAMA:** $\mathbf{b} \leftarrow$ BeliefAggregation($[\hat{\mathbf{Y}}_v^{(1:S)}; \hat{\mathbf{Y}}_a^{(1:S)}]$, type_ids) {Dynamic N_q}
11: **Residual boundary:** $\mathbf{b} \leftarrow \mathbf{b} + \alpha \cdot \mathbf{W}_r \bar{\mathbf{z}}_{\text{bnd}}$
12: **Up-project:** $\mathbf{B}_{\text{inject}} \leftarrow \text{LN}(\mathbf{W}_{ep}\mathbf{b} + \mathbf{b}_{ep})$
13: **Partial-observation inject:** $\mathbf{H}_{\text{aug}} \leftarrow \text{Inject}(\mathbf{H}, \mathbf{B}_{\text{inject}}, \kappa)$ {Training only; κ=1.0 at inference}
14: $\mathcal{L}_{\text{lm}} \leftarrow \text{CrossEntropy}(\text{LMHead}(\mathbf{H}_{\text{aug}}), \text{labels})$
15: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{lm}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}}$
16: Backpropagate $\mathcal{L}_{\text{total}}$ and update parameters

Table 7: Architecture hyperparameters of AffectVerse.

Component	Configuration
Backbone (Qwen2.5-Omni-7B)	
Hidden dimension d	3584
LoRA rank r / scaling α	16 / 32
LoRA dropout	0.05
LoRA target modules	Q, K, V, O projections
Max sequence length	1024
Emotion World Module	
Working dimension d_w	512
Future query tokens N_s (per step)	8
Imagination cross-attention layers	2
Attention heads	8
Rollout steps S	3
Belief query tokens N_q (single / dual)	4 / 8
Belief aggregation layers	1
Type embedding classes	6
Modality tag embeddings	2 (self / cross)
Modality dropout	0.15
Residual scale α (initial)	0.1
FFN activation	GELU
Temporal Splitting	
Keep ratio κ (training)	$\sim$ Uniform$[0.7, 1.0)$
Keep ratio κ (inference)	1.0 (no truncation)
Injection strategy	Interleaved (2 positions)
Training	
Imagination loss	MSE + 0.5 $\times$ Cosine
λ_{img}	1.0
Total EWM parameters	$\sim$25M

D Detailed Method Formulations

This appendix provides extended formulations and theoretical discussions for the Emotion World Module, complementing the concise presentation in §3.2.

D.1 Bottleneck Projection and Temporal Splitting

Information Bottleneck Perspective. The down-projection $\mathbf{W}_m^{\downarrow} \in \mathbb{R}^{d_w \times d}$ (d=3584 $\rightarrow$ d_w=512) serves a dual purpose: (1) it reduces computational cost for all subsequent operations, and (2) it acts as an Information Bottleneck Tishby et al. (2000) that forces the model to compress multimodal representations, retaining only the most emotionally relevant information.

Temporal Splitting Details. For each modality, we split the projected tokens along the temporal axis:

$$\mathbf{Z}_v^{\text{past}} = \mathbf{Z}_v[:T_{p,v}], \quad \mathbf{Z}_v^{\text{fut}} = \mathbf{Z}_v[T_{p,v}:], \quad \mathbf{v}_b = \mathbf{Z}_v[T_{p,v}{-}1], \quad T_{p,v} = \lfloor \kappa \cdot M \rfloor. \tag{11}$$

The same splitting applies to audio tokens with boundary $T_{p,a}$ and boundary token $\mathbf{a}_b = \mathbf{Z}_a[T_{p,a}{-}1]$; Figure 2 denotes the modality-specific boundary generically as T_p. The future portion $\mathbf{Z}^{\text{fut}}$ serves as a detached self-supervised target—the EWM must learn to imagine these representations from the past alone, but the target itself is used only for training-time predictive supervision.

Random Keep Ratio. We employ a stochastic strategy:

$$\kappa \sim \begin{cases} \text{Uniform}[\kappa_{\min}, 1.0) & \text{during training} \\ 1.0 & \text{during inference} \end{cases}, \tag{12}$$

where $\kappa_{\min} = 0.7$. During inference, $\kappa = 1.0$ means no tokens are truncated—the LLM retains all audiovisual tokens while additionally receiving the belief state as an auxiliary predictive summary.

Modality Dropout. We randomly drop an entire modality with probability $p_{\text{drop}} = 0.15$ during training (7.5% per modality). When a modality is dropped, the EWM switches to single-modality mode, training the model to form emotional beliefs from partial sensory evidence.

D.2 Cross-Modal Imagination Architecture

Cross-Modal Context. For video imagination, the context is:

$$\mathbf{C}_v^{(1)} = [(\mathbf{Z}_v^{\text{past}} + \mathbf{e}_{\text{self}});\ (\mathbf{Z}_a^{\text{past}} + \mathbf{e}_{\text{cross}})], \tag{13}$$

where $\mathbf{e}_{\text{self}}, \mathbf{e}_{\text{cross}} \in \mathbb{R}^{d_w}$ are learnable tag embeddings. Symmetrically, $\mathbf{C}_a^{(1)} = [(\mathbf{Z}_a^{\text{past}} + \mathbf{e}_{\text{self}});\ (\mathbf{Z}_v^{\text{past}} + \mathbf{e}_{\text{cross}})]$. When only one modality is available, the context reduces to self-modal past alone.

Multi-Step Rollout Details. The complete rollout procedure for each modality m:

$$\mathbf{Q}^{(s)} = \mathbf{f}_{1:N_s} + \mathbf{e}_{\text{step}}^{(s)}, \tag{14}$$

$$\mathbf{F}^{(s)} = \text{CrossAttn}(\mathbf{Q}^{(s)},\ \mathbf{C}_m^{(s)}), \quad s = 1, \ldots, S, \tag{15}$$

$$\hat{\mathbf{Y}}_m^{(s)} = \mathbf{W}_{\text{out}}\mathbf{F}^{(s)}, \quad \mathbf{C}_m^{(s+1)} = [\mathbf{C}_m^{(1)};\ \hat{\mathbf{Y}}_m^{(s)}], \tag{16}$$

where each cross-attention block consists of 2 layers with Pre-LN, 8 attention heads, and GELU-activated FFN. All rollout steps share the same cross-attention parameters—step embeddings $\mathbf{e}_{\text{step}}^{(s)}$ provide the only temporal distinction, which reduces model size while encouraging a general temporal prediction function.

D.3 MAMA: Memory and Belief Details

Type Embedding Table. MAMA uses the multi-step imagined outputs from each available modality as memory tokens. The type embedding table $\mathbf{E}_{\text{type}} \in \mathbb{R}^{6 \times d_w}$ maps type IDs to dense vectors (6 types reserved for extensibility):

$$\text{type_id}(\mathbf{m}_i) \in \{0 : \text{video imagination},\ 1 : \text{audio imagination}\}. \tag{17}$$

Dynamic Belief Capacity. The number of belief tokens adapts to available modalities:

$$N_q = \begin{cases} 2N_q^0 = 8 & \text{dual-modality} \\ N_q^0 = 4 & \text{single-modality} \end{cases}. \tag{18}$$

The dual- and single-modality settings use *separate* belief aggregation modules (independent parameters), allowing each to specialize in its respective information regime.

Belief Aggregation. Learnable belief queries $\mathbf{b}_{1:N_q} \in \mathbb{R}^{N_q \times d_w}$ attend to all memory tokens through a single cross-attention layer with Pre-LN, 8 heads, and GELU FFN:

$$\mathbf{b}^{(\text{out})} = \text{CrossAttn}(\mathbf{b}_{1:N_q},\ \mathbf{M}) \in \mathbb{R}^{N_q \times d_w}. \tag{19}$$

D.4 Interleaved Injection Details

Up-Projection. $\mathbf{B}_{\text{inject}} = \text{LayerNorm}(\mathbf{W}_{ep}\mathbf{b}^{(\text{final})} + \mathbf{b}_{ep}) \in \mathbb{R}^{N_q \times d}$, where $\mathbf{W}_{ep} \in \mathbb{R}^{d \times d_w}$ and $\mathbf{b}_{ep} \in \mathbb{R}^d$.

Injection Positions. Belief tokens are split into two halves:

$$\mathbf{B}_{\text{inject}} = [\ \underbrace{\mathbf{b}_1, \ldots, \mathbf{b}_{N_q/2}}_{\text{Position 1: AV–Text boundary}}\ ;\ \underbrace{\mathbf{b}_{N_q/2+1}, \ldots, \mathbf{b}_{N_q}}_{\text{Position 2: Text–Answer boundary}}\ ]. \tag{20}$$

Token-Level Truncation. During training, a keep mask is built over the embedded token sequence. For each contiguous video/audio region, only the first κ fraction is retained and the remaining suffix is removed before computing the LM loss. The shortened examples are then padded to a common length, with attention masks and labels rebuilt to match the new sequence. This trains the LLM to use belief tokens under partial observations; their labels are set to -100 (ignored in loss), while their attention mask is set to 1. During inference (κ=1.0), no audiovisual token is removed, and the belief tokens enrich rather than replace the full multimodal context.

D.5 Training Details

Imagination Loss Averaging. The imagination loss averages over all rollout steps and available modalities:

$$\mathcal{L}_{\text{imagine}} = \frac{1}{|\mathcal{S}|} \sum_{m \in \mathcal{S}} \frac{1}{S} \sum_{s=1}^{S} \mathcal{L}_{\text{imagine},m}^{(s)}, \tag{21}$$

where $\mathcal{S}$ is the set of available modalities after modality dropout. The MSE component enforces magnitude alignment while the cosine component enforces directional (semantic) alignment, preventing collapse to mean representations—consistent with strategies in self-supervised learning Grill et al. (2020); Caron et al. (2021).

D.6 Per-Step Imagination Quality Breakdown

Table 8 provides a detailed breakdown of imagination quality at each individual rollout step. Since all rollout steps share the same cross-attention parameters (with only step embeddings providing temporal distinction), and each step's output is supervised against the same adaptive-pooled future target, we report the cosine similarity between each step's output and the ground-truth future representation.

Table 8: **Per-step imagination quality breakdown.** Cosine similarity between each rollout step's imagined output and the real future representation, measured on the validation set. "—" indicates that the step does not exist for the given S configuration. Avg Cos and Avg MSE are averaged across all active steps and both modalities.

S	Step 1	Step 2	Step 3	Step 4	Step 5	Avg Cos ↑	Avg MSE ↓	Avg Perf ↑
1	0.812	—	—	—	—	0.812	0.041	75.81
2	0.811	0.799	—	—	—	0.805	0.045	76.51
3	**0.810**	**0.798**	**0.787**	—	—	**0.798**	**0.048**	**77.06**
4	0.809	0.796	0.784	0.768	—	0.789	0.052	76.97
5	0.808	0.795	0.783	0.766	0.751	0.781	0.055	76.75

Several patterns emerge from the per-step breakdown: (1) Step 1 cosine similarity is stable across all S configurations (0.808–0.812), indicating that the shared-parameter imagination module maintains similar first-step quality regardless of how many subsequent steps are appended. (2) Each additional step shows a modest quality decrease of $\sim$0.012–0.017 in cosine similarity, suggesting that error accumulation is controlled but non-negligible. (3) The quality gap between the first and last step widens with S: from 0.012 (S=2) to 0.057 (S=5), which helps explain why excessive rollout steps ($S \geq 4$) lead to diminishing returns in downstream performance despite providing more imagination tokens. This per-step analysis provides evidence supporting the choice of S=3 as the best trade-off.

D.7 Additional Case Study

In this case, AffectVerse correctly integrates the textual, acoustic, and visual cues to infer a positive emotional state. In contrast, AffectGPT mistakes the high pitch and fast speech for anxiety, while Qwen-omni produces a negative interpretation. This example shows that AffectVerse is better at distinguishing excited positive affect from negative emotional states, which is critical for accurate emotion recognition in high-arousal scenarios.

E Additional Controlled Experiments and Analyses

This appendix reports additional controlled experiments addressing the fairness of the headline comparison, the contribution of predictive supervision beyond added capacity, the affective semantics of imagined representations, cross-dataset generalization, and computational cost. Unless otherwise noted, numbers are six-benchmark means (CMU-MOSI, CMU-MOSEI, MER2024, MELD, SIMS, IEMOCAP) under the A+V+T setting with the same backbone, data, prompts, and training configuration as the main paper.

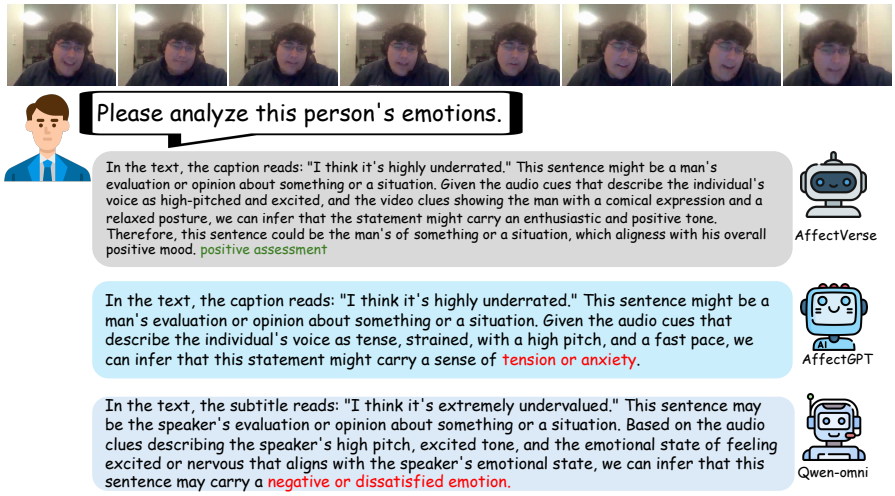

Figure 8: **Case study of multimodal emotion reasoning.** AffectVerse correctly infers a positive assessment, while AffectGPT and Qwen-omni misinterpret excited speech cues as negative emotions.

E.1 Same-Backbone and Cross-Backbone Gains

The external results in the main comparison table are intended for benchmark positioning. The contribution of EWM should instead be measured under paired comparisons using the same backbone, data, prompts, and training configuration. Table 9 reports the paired results: EWM improves the LoRA-tuned Qwen2.5-Omni-7B baseline by $+3.40$, and yields consistent gains when integrated into the released AffectGPT with its original multimodal encoders ($+3.10$) and into MiniCPM-o 2.6 ($+3.66$). Backbone strength affects absolute performance but does not explain the relative gain from EWM.

Table 9: **EWM gains across backbones** (six-benchmark mean). All pairs share the same data, prompts, and training configuration.

Backbone	Baseline	Baseline + EWM	Gain
Qwen2.5-Omni-7B	73.66	77.06	$+3.40$
AffectGPT	72.10	75.20	$+3.10$
MiniCPM-o 2.6	71.92	75.58	$+3.66$

E.2 Capacity-Matched Non-Predictive Controls

To isolate the contribution of future prediction, we compare against architecture- and capacity-matched non-predictive controls (Table 10). The LM-only ($\lambda_{\text{img}}=0$), shuffled-future, current-state-reconstruction, and orderless-future variants preserve the exact EWM architecture and differ only in the supervision signal; the parameter-matched Query-Former is a strong alternative non-predictive summarization architecture. The true temporally aligned future objective outperforms these matched controls by 1.70–2.57 points, and the strongest non-predictive control by $+1.75$. The full gain therefore cannot be explained by additional parameters, extra tokens, MAMA aggregation, or a generic auxiliary regression objective: both the future content and its temporal correspondence contribute.

E.3 What Do Imagined Representations Encode?

To test whether imagined representations encode meaningful information about future affect, we freeze the learned representations and train identical linear probes to predict the emotion of the held-out future segment (Table 11, left). We additionally evaluate future-representation retrieval (Table 11, right). Predictive beliefs substantially outperform past-state, LM-only, and reconstruction controls,

Table 10: **Capacity-matched controls** (six-benchmark mean). Controls differ from EWM only in the learning target.

Configuration	Learning target	Six-benchmark mean
Qwen2.5-Omni + LoRA	—	73.66
Random tokens	no learning	74.09
Temporal pooling	current states pooled	74.97
LM-only belief module (λ_{img}=0)	no future supervision	74.49
Shuffled future target	misordered future	74.71
Current-state reconstruction	observed states	75.18
Orderless future representation	future, no temporal order	75.36
Capacity-matched Query-Former	no future prediction	75.31
Predictive EWM	temporally aligned future	**77.06**

approaching the real future representation (the upper bound). This indicates that the imagined tokens contain decodable information about the true future affective state, rather than acting only as arbitrary additional hidden vectors.

Table 11: **Linear probing** for future-segment emotion (WAF/Acc).

Probe input	Future emotion WAF/Acc
Past boundary token	54.8
Temporal pooling	56.1
λ_{img}=0 belief	56.5
Shuffled-future belief	55.9
Current-reconstruction belief	58.3
Predictive belief	**62.9**
Real future representation	66.8

Representation	R@1	R@5
Past representation	20.9	47.3
λ_{img}=0 belief	25.6	53.4
Current reconstruction	29.8	58.1
Predictive belief	**42.7**	**71.5**

E.4 Leave-One-Dataset-Out Generalization

The main experiments follow a multi-dataset in-domain training protocol. To test whether EWM depends on the target benchmark being included during training, we conduct leave-one-dataset-out evaluation (Table 12). EWM consistently improves the held-out benchmark without in-domain training, with a mean gain of $+2.03$, indicating that the learned predictive belief structure transfers across datasets rather than fitting benchmark-specific cues.

Table 12: **Leave-one-dataset-out evaluation.** The held-out dataset is excluded from training.

Held-out dataset	Baseline	AffectVerse	Gain
MER2024	57.8	60.0	$+2.2$
MELD	43.9	45.6	$+1.7$
MOSEI	72.4	74.3	$+1.9$
SIMS	65.3	67.6	$+2.3$
Mean	59.85	61.88	$+2.03$

E.5 Graceful Degradation Under Partial Observation

We do not claim that every modality is beneficial on every dataset; in some cases removing video reduces modality-specific noise or redundancy. The relevant robustness question is whether AffectVerse degrades more gracefully than the controlled baseline under the same partial-observation condition. Tables 13 and 13 evaluate temporal-prefix and missing-modality settings: the relative advantage of AffectVerse increases as the observation becomes less complete ($+3.40$ at 100% to $+5.52$ at 10% observed), and remains positive under every missing-modality setting.

Table 13: **Temporal-prefix evaluation** (six-benchmark mean).

Observed ratio	Baseline	AffectVerse	Gain
100%	73.66	77.06	+3.40
70%	71.95	75.88	+3.93
50%	69.70	74.15	+4.45
30%	66.85	71.98	+5.13
10%	61.90	67.42	+5.52

Setting	Baseline	AffectVerse	Gain
Full	73.66	77.06	+3.40
No Audio	68.30	72.55	+4.25
No Video	68.90	72.85	+3.95

E.6 Computational Cost

We benchmark AffectGPT, the Qwen2.5-Omni thinker + LoRA baseline, and the EWM thinker on a single NVIDIA H100 80GB GPU, with each input containing eight video frames, audio, and subtitles; the thinker + LoRA and EWM thinker share the same backbone, BF16 precision, inputs, and generation configuration (Table 14). EWM adds only 25.22M parameters (≈0.36% of the 7B backbone) and ≈49 MiB of allocated tensor memory, with no observable reduction in generation speed, while improving the six-benchmark mean from 73.66 to 77.06—a favorable performance–efficiency trade-off. The small speed difference may result from GPU clock and runtime fluctuations, so we do not claim that EWM accelerates inference.

Table 14: **Efficiency comparison** on a single NVIDIA H100 80GB GPU.

	AffectGPT	Thinker + LoRA	EWM Thinker
Precision	FP16	BF16	BF16
Additional EWM parameters	—	—	25.22M
Allocated tensor memory	17.23 GiB	17.25 GiB	17.30 GiB
Peak tensor memory	17.32 GiB	17.56 GiB	17.60 GiB
Generation speed	35.28 tok/s	32.54 tok/s	33.47 tok/s

F Reproducibility and Evaluation Details

F.1 Prompt and Answer-Extraction Protocol

Our main MER-UniBench evaluation strictly follows the official AffectGPT zero-shot inference protocol. All nine datasets (MER2023, MER2024, MELD, IEMOCAPFour, CMU-MOSI, CMU-MOSEI, SIMS, SIMSv2, and OV-MERD+) are evaluated with the same open-vocabulary task instruction—"`Please recognize all possible emotional states of the character.`"—with no candidate labels provided. No prompt is separately designed or tuned for AffectVerse; the controlled Qwen2.5-Omni baseline and AffectVerse use this identical task instruction, the same audiovisual and subtitle content, and the same decoding configuration. Model-specific wrappers follow their native architectures (AffectGPT uses its released `###Human`/`###Assistant` wrapper, whereas Qwen2.5-Omni uses its native chat template); this formatting difference does not change the task instruction or the evaluation target.

All model outputs are evaluated with the official AffectGPT/MER-UniBench pipeline, including label extraction, synonym normalization, sentiment-polarity conversion, and metric computation. Discrete emotion classification uses the official label extraction, normalization, and classification metrics; sentiment analysis uses the official positive/negative/neutral mapping and binary evaluation; OV-MERD+ uses the official open-vocabulary and emotion-wheel evaluation. The baseline and AffectVerse share the same prompts, output format, and post-processing code, so observed performance differences do not result from model-specific prompt engineering or answer-extraction rules.

F.2 Definition of Past and Future

During training, past and future are defined within each temporally ordered audiovisual clip at the latent-token level. For each modality with T tokens, we sample $\kappa \sim U[0.7, 1.0)$; the first $\lfloor \kappa T \rfloor$ tokens form the observed prefix, and the remaining tokens form the detached future target after

adaptive pooling. At inference, we set κ=1: all audiovisual tokens from the given clip are retained, and no unseen future frames or audio are accessed. EWM predicts latent future states solely from the observed representations.

F.3 Full Hyperparameters

Table 15 lists the complete training and evaluation configuration.

Table 15: **Complete hyperparameters.**

Category	Hyperparameter	Value
Backbone	Base model	Qwen2.5-Omni-7B
	Fine-tuning method	LoRA
	LoRA rank	16
	LoRA alpha	32
	LoRA dropout	0.05
	LoRA target	All attention projections (Q, K, V, O)
Optimization	Optimizer	AdamW
	Learning rate	2×10^{-4}
	LR scheduler	Cosine
	Warmup ratio	0.03
	Weight decay	0.01
	Maximum gradient norm	1.0
	Training epochs	3
	Per-GPU batch size	4
	Gradient accumulation	8
	Number of GPUs	8
	Effective batch size	32
	Numerical precision	BF16
	Distributed training	DeepSpeed ZeRO-2
	Optimizer offloading	Disabled
Input	Maximum sequence length	1,024
	Maximum image pixels	262,144
	Maximum video pixels	50,176
	Maximum video length	32
EWM	Working dimension	512
	Number of attention heads	8
	Imagination layers	2
	Rollout steps	3
	Future-query tokens	8
	Belief tokens	4
	Training modality dropout	0.15
	Inference modality dropout	0
	Imagination-loss weight	1.0
	Training keep ratio	0.7
Evaluation	Inference keep ratio	1.0
	Maximum generated tokens	512
	Temperature	0.1
	Top-p	0.95
Checkpointing	Evaluation interval	200 steps
	Saving interval	200 steps
	Model-selection criterion	Validation loss
Reproducibility	Random seed	42

F.4 Reproducibility Statement

All experiments use fixed dataset splits and a random seed of 42, propagated through the LLaMA-Factory pipeline to Python, NumPy, PyTorch, the data shuffling procedure, and the data loaders,

so experiments can be consistently reproduced under the same codebase, dataset splits, software environment, and hardware configuration. To facilitate reproduction, we will publicly release the complete training and evaluation code, all training/inference/DeepSpeed configuration files, dataset preprocessing and multimodal input-construction scripts, exact commands for reproducing all main experiments and ablations, dataset split files and random-seed settings, environment and dependency specifications, and the trained LoRA adapters and EWM weights.