Emotion-JEPA: Predictive Visual Adaptation and Audio-Modulated Fusion for Multimodal Emotion Recognition

Anonymous authors
Paper under double-blind review

Abstract

Multimodal emotion recognition (MER) requires combining visual, acoustic, and textual cues from short, noisy, and often ambiguous emotional expressions. While large pretrained multimodal models provide strong general-purpose representations, their direct use for MER can be limited by a mismatch between generic pretraining data and fine-grained affective behavior, as well as by fusion mechanisms that do not explicitly account for modality reliability. We study a two-stage framework for MER that isolates two factors: affective visual representation adaptation and reliability-aware multimodal fusion. In the first stage, we adapt a visual encoder to the emotion domain using predictive self-supervised learning on unlabeled emotion videos, without using pseudo-labels or additional manual annotations. In the second stage, we train a supervised multimodal classifier with Audio-Modulated Hybrid Fusion (AMHF), where audio cues guide cross-modal interaction through audio spectral gating, adaptive cross-modal routing, temporal memory, uncertainty estimation, and progressive fusion. On the MER2024-SEMI benchmark, visual predictive adaptation improves performance by +7.92 weighted-average F1 (WAF) over the same model without domain adaptation. Under matched encoders, parameter budgets, and training protocols, AMHF improves performance by +7.25 WAF over a capacity-matched cross-attention fusion baseline. Component-level ablations further show that each AMHF stage contributes to the final performance. These results suggest that, for MER under limited supervision, domain-aligned representation learning and reliability-aware fusion can be as important as increasing model scale.

1 Introduction

Multimodal emotion recognition (MER) aims to infer affective states from visual, acoustic, and linguistic signals. In practice, this problem remains difficult because emotional expressions are often subtle, short-lived, and unevenly expressed across modalities. Facial cues may be weak, occluded, or affected by pose and lighting, acoustic cues may vary with speaker characteristics, background noise, recording quality, and prosodic expressiveness, and textual transcripts may be sparse, incomplete, or only weakly related to the expressed emotion. These challenges are especially pronounced in settings with limited labeled data, where a model must learn to combine heterogeneous and noisy cues without over-relying on any single modality.

Large pretrained multimodal models provide strong general-purpose representations and have shown impressive transfer across video, audio, and language tasks (Alayrac et al., 2022; Lin et al., 2024; Liu et al., 2023). However, their direct use for affective understanding is not always sufficient. First, visual encoders trained on broad video corpora may emphasize scene-level semantics or action dynamics, while MER often depends on fine-grained facial motion and subtle temporal changes. Second, standard fusion mechanisms, including cross-attention-based designs (Vaswani et al., 2017; Tsai et al., 2019), usually combine modalities in a largely symmetric manner. Such mechanisms do not explicitly model the fact that modality reliability can vary from sample to sample: audio may contain strong prosodic evidence when the face is partially occluded, while visual cues may be more informative when speech is neutral or transcripts are uninformative. As a result, improvements in MER can be difficult to attribute: they may come from larger pretrained backbones, from better domain alignment, or from the inductive bias of the fusion mechanism itself.

This paper studies these two factors separately. We ask whether predictive visual adaptation on unlabeled emotion-domain videos improves affective representations, and whether audio-conditioned reliability-aware fusion improves multimodal decision making under matched encoders and training protocols. Our approach is motivated by the predictive learning view of V-JEPA-style video representation learning, where a model learns by predicting masked latent spatiotemporal features rather than reconstructing pixels (Assran et al., 2025). We use the term *predictive visual adaptation* in this restricted sense: the visual encoder is further adapted on emotion-domain videos through latent feature prediction. We do not claim that the resulting MER model performs planning, nor that predictive objectives are universally preferable to generative or contrastive alternatives. Instead, we evaluate whether this form of in-domain predictive adaptation is useful for fine-grained affect recognition.

We introduce Emotion-JEPA, a two-stage framework for multimodal emotion recognition. In the first stage, the visual encoder is adapted using unlabeled emotion videos with a masked predictive objective. This stage uses only video and does not rely on pseudo-labels, self-training, or additional manual annotations. To the best of our knowledge, this is among the first studies to examine V-JEPA-style predictive visual adaptation for multimodal emotion recognition. Its goal is to align the visual representation with affect-relevant facial dynamics before supervised multimodal training. In the second stage, the adapted visual encoder is combined with pretrained audio and text encoders through Audio-Modulated Hybrid Fusion (AMHF). AMHF uses audio features as a reliability signal to guide multimodal interaction through audio spectral gating, adaptive cross-modal routing, temporal memory, uncertainty estimation, and progressive fusion. This design reflects the empirical observation that acoustic prosody is often a strong and stable cue in MER, while visual and textual evidence can be more variable across samples.

Our emphasis is not on introducing a new large-scale pretrained model, but on a controlled analysis of representation adaptation and fusion design. We compare the proposed model against ablations and capacity-matched fusion baselines under the same modality encoders, parameter budgets, and optimization settings. This allows us to separate the effect of visual domain adaptation from the effect of the multimodal fusion mechanism. On MER2024-SEMI, predictive visual adaptation improves performance by +7.92 WAF over the same model without domain adaptation. Under matched encoders and training protocols, AMHF improves performance by +7.25 WAF over a capacity-matched cross-attention fusion baseline. Additional unimodal, component-level, and diagnostic analyses show that the final performance depends on both affect-aligned visual representations and reliability-aware multimodal fusion.

Contributions:

- **An early study of V-JEPA-style predictive visual adaptation for MER:** We adapt a V-JEPA-style visual encoder to emotion-domain videos using unlabeled face-centric clips and a masked latent prediction objective. This provides an early study of predictive video representation adaptation for multimodal emotion recognition and improves affect-sensitive visual representations without pseudo-labeling or additional manual supervision.

- **Audio-Modulated Hybrid Fusion for reliability-aware multimodal learning:** We introduce AMHF, a fusion module in which audio features guide multimodal interaction through audio spectral gating, adaptive cross-modal routing, temporal memory, uncertainty estimation, and progressive fusion. Under matched encoders and training protocols, AMHF improves over a capacity-matched cross-attention baseline.

- **Controlled empirical analysis of adaptation and fusion:** Through unimodal evaluation, framework-level ablations, AMHF component ablations, comparisons with multimodal foundation models, and transfer evaluation on MER2025-SEMI, we analyze how visual domain adaptation and audio-conditioned fusion contribute to MER performance.

Together, these results suggest that robust MER under limited supervision depends not only on the scale of pretrained models, but also on whether representations are aligned with affective dynamics and whether fusion mechanisms can account for modality-specific reliability.

2 Related Work

2.1 Benchmarks and Limited Supervision in Multimodal Emotion Recognition

Recent progress in multimodal emotion recognition (MER) has been shaped by both benchmark design and model development. Early datasets such as IEMOCAP (Busso et al., 2008) enabled controlled studies of audiovisual and linguistic emotion recognition, while MERBench (Lian et al., 2026) highlighted the need for standardized evaluation across feature extractors, fusion modules, splits, and training protocols. The MER challenge series further expands this setting: MER2024 introduced semi-supervised learning, noise robustness, and open-vocabulary emotion recognition (Lian et al., 2024b), and MER2025 extends the benchmark toward LMM-based affective computing, fine-grained recognition, descriptive emotion reasoning, and personality-related prediction (Lian et al., 2025b).

Limited labeled data remains a central difficulty in this setting. Recent MER2024-SEMI systems often use pseudo-labeling, self-training, modality dropout, voting, or iterative refinement over unlabeled data (Qi et al., 2024; Ge et al., 2024; Shi & Gao, 2024). These strategies can improve benchmark performance by converting unlabeled examples into additional supervisory signal, but they also make it harder to isolate gains from model architecture or representation learning. In contrast, Emotion-JEPA uses unlabeled clips only for predictive visual adaptation and does not assign pseudo-labels to unlabeled samples, allowing us to study representation adaptation and reliability-aware fusion under a controlled supervision protocol.

2.2 Large Multimodal Models for Affective Understanding

Large multimodal models (LMMs) such as CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), Video-LLaVA (Liu et al., 2023), Video-LLaMA (Zhang et al., 2023), and Qwen-VL/Qwen-Omni (Bai et al., 2023; Xu et al., 2025) provide strong general-purpose multimodal representations. Their success has motivated affective computing methods based on prompting, instruction tuning, and multimodal reasoning. For example, EmotionLLaMA (Cheng et al., 2024) adapts instruction-tuned multimodal models for emotion recognition and reasoning, while AffectGPT and MER-Caption extend this direction with descriptive emotion annotations and multimodal emotion understanding benchmarks (Lian et al., 2025a).

Recent work also explores open-vocabulary MER, where models predict flexible affective descriptions rather than choosing from a fixed label set (Lian et al., 2024a). These developments show the growing role of LMMs in affective computing, but they also raise a complementary question: how should modality-specific evidence be adapted and fused in compact discriminative MER systems? Emotion-JEPA addresses this question by studying affective visual adaptation and audio-conditioned fusion under controlled encoder and training settings, rather than relying primarily on instruction tuning, generative descriptions, or model scale.

2.3 Self-Supervised Video Learning and Affective Adaptation

Self-supervised learning has become a standard approach for learning transferable visual representations, including contrastive methods (Chen et al., 2020; He et al., 2020), non-contrastive methods (Grill et al., 2020), and masked prediction models (He et al., 2022). In video understanding, predictive and masked objectives are especially relevant because they encourage models to capture temporal structure across frames.

The Joint-Embedding Predictive Architecture (JEPA) (Assran et al., 2023) predicts representations in latent space rather than reconstructing pixels. V-JEPA extends this idea to video through masked latent prediction (Bardes et al., 2024), and V-JEPA2 further scales predictive video representation learning (Assran et al., 2025). These models provide strong generic video features, but direct transfer may be suboptimal for MER because affective signals are often subtle, face-centric, and short-lived. Emotion recognition frequently depends on brief facial movements, gaze shifts, and low-intensity expression changes that may be underrepresented in large action- or scene-oriented video corpora. Recent work has also begun to explore V-JEPA for facial expression recognition using pretrained video encoders and shallow classifiers (Eing et al., 2026). Our work differs by studying in-domain predictive visual adaptation for multimodal emotion recognition, where the adapted visual representation is evaluated together with audio and text.

We therefore use predictive self-supervised learning as an in-domain adaptation mechanism rather than as a new pretraining objective. The visual encoder is adapted on unlabeled face-centric emotion videos before supervised multimodal training, allowing us to test whether predictive adaptation reduces the domain gap between generic video pretraining and affective visual recognition.

2.4 Reliability-Aware Multimodal Fusion

Fusion remains a central problem in MER because each modality can be informative, ambiguous, or misleading depending on the sample. Classical approaches include early and late fusion, tensor-based fusion (Zadeh et al., 2017), attention-based fusion (Liang et al., 2018; Tsai et al., 2019), and gated fusion mechanisms (Arevalo et al., 2017). More recent methods emphasize sample-adaptive reliability modeling, including audio-guided fusion (Shi & Gao, 2024) and modality-specific dynamic emotion experts (Fang et al., 2025). These approaches move beyond static fusion by estimating which modality or interaction pathway should be emphasized for each example.

Emotion-JEPA follows this reliability-aware direction but uses audio as an explicit conditioning signal. AMHF combines audio spectral gating, adaptive cross-modal routing, temporal memory, uncertainty estimation, and progressive fusion to regulate how visual and textual evidence contribute to the final prediction. This design uses acoustic prosody as a stabilizing cue when visual or textual evidence is weak, sparse, or noisy.

Positioning. Emotion-JEPA complements prior work on MER benchmarks, LMM-based affective understanding, semi-supervised learning, and adaptive fusion by isolating the effects of predictive visual adaptation and audio-conditioned reliability-aware fusion under a controlled labeled protocol.

3 Methodology

Overview: Emotion-JEPA is a two-stage framework for multimodal emotion recognition. Stage 1 adapts a pretrained visual video encoder to the affective domain using unlabeled emotion-domain videos and a masked latent prediction objective. This stage is visual-only: it does not use audio, text, pseudo-labels, or manual annotations. Stage 2 trains a supervised multimodal classifier that combines the adapted visual representation with audio and text through Audio-Modulated Hybrid Fusion (AMHF). AMHF integrates audio spectral gating, adaptive cross-modal routing, temporal memory, uncertainty estimation, and progressive fusion to regulate cross-modal evidence. The resulting framework separates the representation question from the fusion question: whether predictive adaptation improves affect-sensitive visual features, and whether audio-conditioned fusion improves decision making when modality reliability varies across samples.

3.1 Problem Setup and Notation

We formulate multimodal emotion recognition as supervised classification over $C = 6$ emotion categories in MER2024-SEMI (Lian et al., 2024b). The labeled dataset is

$$\mathcal{D} = \{(x_i^v, x_i^a, x_i^t, y_i)\}_{i=1}^N,$$

where x_i^v denotes a video clip, x_i^a its synchronized audio segment, x_i^t the corresponding transcript tokens, and $y_i \in \{1, \ldots, C\}$ the emotion label. An unlabeled video set

$$\mathcal{U}_v = \{x_j^v\}_{j=1}^M$$

is used only during Stage 1 for visual predictive adaptation.

Each modality $m \in \{v, a, t\}$ is processed by an encoder f_m and projection head P_m:

$$h_m = P_m(f_m(x^m)), \qquad h_m \in \mathbb{R}^d, \quad d = 512. \tag{1}$$

All fusion operations in Stage 2 are performed in this shared embedding space.

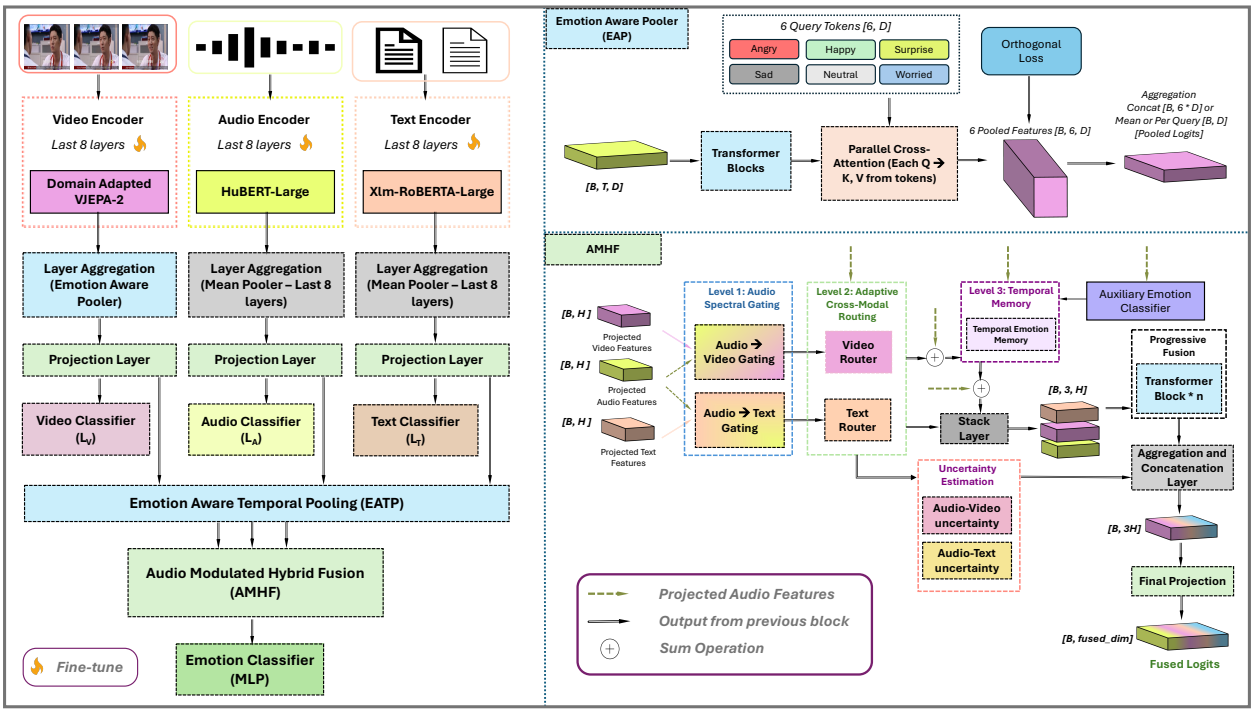


Figure 1: **Overview of Emotion-JEPA.** Stage 1 adapts the V-JEPA2 visual encoder on unlabeled emotion-domain videos using masked latent prediction, without audio, text, or pseudo-labels. Stage 2 performs supervised MER using projected visual, audio, and text representations. AMHF combines audio spectral gating, adaptive cross-modal routing, temporal memory, uncertainty estimation, and progressive fusion for classification.

Stage 1 updates only the visual branch using $\mathcal{U}_v$. Stage 2 performs supervised multimodal learning on $\mathcal{D}$, where projection heads and the fusion module are fully trainable, while only the upper layers of the pretrained modality encoders are fine-tuned. Earlier encoder layers are kept frozen to preserve pretrained representations and reduce overfitting under limited labeled supervision. We restrict predictive adaptation to the visual encoder because the visual branch is most sensitive to the domain gap between generic video pretraining and face-centric affect recognition, while the audio and text branches already benefit from large-scale speech and language pretraining.

3.2 Stage 1: Affective Visual Domain Adaptation

The goal of Stage 1 is to adapt the visual encoder to emotion-domain video before supervised multimodal training. Generic video encoders capture broad motion and scene dynamics, but MER often depends on subtle visual cues such as brief facial movements, gaze changes, and low-intensity expression transitions. We therefore continue training the visual encoder on unlabeled face-centric emotion videos using a masked latent prediction objective.

Adaptation is performed on approximately 113k unlabeled clips from MER2024. Each clip is sampled at 4 FPS and resized or cropped to 256×256. We apply multi-scale spatiotemporal masking over a large portion of the video volume, encouraging the model to infer masked latent targets from visible temporal context. This objective encourages the encoder to represent short-range facial dynamics and stable facial structure without requiring emotion labels.

Following a JEPA-style formulation (Assran et al., 2025), each clip is divided into visible context regions and masked target regions. Let f_θ denote the student encoder, f_ξ the exponential-moving-average target

encoder, and g_θ the predictor. The predictive adaptation objective is

$$\mathcal{L}_{\text{pred}} = \left\| g_\theta\big(f_\theta(x_{\text{ctx}})\big) - \text{sg}(f_\xi(x_{\text{tgt}})) \right\|_2^2, \tag{2}$$

where x_{ctx} denotes the visible context, x_{tgt} denotes the masked target regions, and $\text{sg}(\cdot)$ stops gradients through the EMA target branch. The target encoder parameters ξ are updated as an exponential moving average of the student parameters θ. During adaptation, only the upper transformer blocks of the visual encoder and the predictor are updated and the lower visual layers remain frozen. This preserves general visual features while allowing higher-level representations to specialize toward affective dynamics.

After adaptation, the resulting visual encoder f_v^* is used in Stage 2. Additional details on the trainable visual subset, masking policy, EMA update, optimization settings, and compute budget are provided in Appendix B. Tables 8 and 9 summarize the Stage 1 parameter and masking configurations.

3.3 Stage 2: Supervised Multimodal Learning

Stage 2 trains the multimodal emotion classifier using the labeled set $\mathcal{D}$. The adapted visual encoder f_v^* processes the video stream, while pretrained HuBERT-Large and XLM-RoBERTa-Large encoders process the audio and text streams, respectively. Each modality representation is projected into the shared embedding space defined in Eq. 1, yielding (h_v, h_a, h_t).

Video and audio are aligned at the clip level before fusion. Text is treated as an utterance-level semantic cue rather than a temporally aligned stream, because transcripts are often short and do not provide reliable token-level timing. Accordingly, temporal pooling is applied only to the video and audio representations in the final configuration, while text is used as a global contextual feature. Full details on modality encoders, projection heads, Stage 2 optimization, and training throughput are provided in Appendix C.

3.4 Emotion-Aware Pooling

Before fusion, encoder outputs are converted into clip-level embeddings. Instead of relying only on average pooling, we use learned emotion queries to aggregate modality evidence in a class-aware manner refered as emotion-aware pooling(EAP). The same EAP operation is used for the visual Emotion-Aware Pooler (vEAP), which aggregates V-JEPA2 visual tokens before projection, and the post-projection audiovisual pooling path.

Given an input sequence $S \in \mathbb{R}^{L \times d_s}$ and learned emotion queries $Q \in \mathbb{R}^{C \times d_s}$, we compute

$$O = \text{MHA}(Q, S, S), \qquad \alpha = \text{softmax}(\psi(O)), \qquad \text{EAP}(S; Q) = \sum_{c=1}^{C} \alpha_c O_c, \tag{3}$$

where MHA denotes multi-head cross-attention, ψ maps each query output to a scalar, and the softmax is taken over the C emotion-query outputs. The queries are learned end-to-end from the MER objective and are not conditioned on the ground-truth label during inference. When enabled, we add an orthogonality regularizer $\|QQ^\top - I_C\|_F^2$ to encourage diverse emotion queries.

For the visual branch, let $X_v = f_v^*(x^v)$ be the sequence of visual tokens from the adapted V-JEPA2 encoder. vEAP aggregates these tokens before projection:

$$\bar{x}_v = \text{EAP}(X_v; Q_v), \qquad h_v = P_v(\bar{x}_v). \tag{4}$$

This replaces generic visual average pooling with emotion-query-based aggregation over visual tokens.

For the post-projection pooling path, we apply the same EAP operation to the projected video and audio representations, while text remains an utterance-level feature. We refer to this path as Emotion-Aware Temporal Pooling (EATP), following the implementation name used in the ablations. In the final model, EATP provides an additional emotion-query aggregation layer over projected audiovisual features rather than a strict token-level temporal alignment module. For notational simplicity, we continue to denote the resulting modality representations as (h_v, h_a, h_t) in the fusion module.

3.5 Audio-Modulated Hybrid Fusion

AMHF combines modality evidence while accounting for sample-dependent reliability. In MER, audio may provide strong prosodic evidence when visual cues are weak, visual information may dominate when facial expressions are clear, and text may help only when transcripts contain emotion-bearing content. AMHF therefore uses audio as a conditioning signal for cross-modal fusion. As shown in Figure 1, AMHF consists of five components: audio spectral gating, adaptive cross-modal routing, temporal memory, uncertainty estimation, and progressive fusion.

Audio spectral gating: Given projected embeddings $h_v, h_a, h_t \in \mathbb{R}^d$, AMHF first uses the audio representation to modulate the video and text representations. For each non-audio modality $x \in \{v, t\}$, h_x is split into G channel groups $\{h_x^{(g)}\}_{g=1}^{G}$. The audio embedding predicts group-level gates

$$c = \sigma(Ah_a), \qquad A \in \mathbb{R}^{G\times d},$$

and the non-audio features are modulated as

$$\hat{h}_x^{(g)} = c_g h_x^{(g)}, \qquad x \in \{v, t\}, \quad g = 1, \ldots, G. \tag{5}$$

Although the gates are applied to feature groups, we refer to this block as audio spectral gating because the gating signal is predicted from the audio representation.

Adaptive cross-modal routing: After gating, AMHF routes each audio-modality pair through a sparse mixture of experts. For $x \in \{v, t\}$, the routed representation is computed from the concatenated pair $[h_a \| \hat{h}_x]$. The audio-conditioned router produces expert weights

$$\pi = \mathrm{softmax}(Rh_a), \qquad R \in \mathbb{R}^{E\times d},$$

and selects the top-k experts:

$$r_x = \sum_{j\in \text{Top-}k(\pi)} \pi_j E_j\big([h_a \| \hat{h}_x]\big), \qquad x \in \{v, t\}, \tag{6}$$

where each expert $E_j : \mathbb{R}^{2d} \to \mathbb{R}^d$ is an MLP. A lightweight load-balancing regularizer is used to reduce expert collapse. This routing step allows the model to use different interaction functions for different audio-conditioned multimodal states.

Temporal memory: The routed features attend to a small set of learnable memory prototypes that represent recurring affective patterns. This pathway provides temporal stabilization when the current sample contains weak or ambiguous modality evidence. Let a_{mem} denote the memory-refined audio representation obtained after attending to these prototypes. A lightweight transformer refinement layer then integrates the memory-enhanced audio stream with the routed audio-video and audio-text streams.

Uncertainty estimation: AMHF estimates scalar uncertainty scores for the routed audio-video and audio-text interactions:

$$u_{av} = \sigma(\phi_{av}(r_v)), \qquad u_{at} = \sigma(\phi_{at}(r_t)), \tag{7}$$

where ϕ_{av} and ϕ_{at} are lightweight MLPs. Lower uncertainty corresponds to higher confidence. The normalized confidence weights are

$$\omega_{av} = \frac{1-u_{av}}{(1-u_{av})+(1-u_{at})+\epsilon}, \qquad \omega_{at} = \frac{1-u_{at}}{(1-u_{av})+(1-u_{at})+\epsilon}. \tag{8}$$

The uncertainty-weighted AMHF representation is

$$z_{\text{AMHF}} = P_f\left([a_{\text{mem}} \| \omega_{av} r_v \| \omega_{at} r_t]\right), \tag{9}$$

where $P_f : \mathbb{R}^{3d} \to \mathbb{R}^d$.

Progressive fusion: We preserve direct modality evidence through a residual projection of the original modality embeddings:

$$z_{\text{res}} = P_{\text{res}}([h_v \| h_a \| h_t]), \qquad z_{\text{fused}} = z_{\text{AMHF}} + \gamma z_{\text{res}}. \tag{10}$$

Unless otherwise specified, we use $\gamma = 1$ in all controlled comparisons. We refer to the staged composition of gating, routing, memory refinement, uncertainty weighting, and residual aggregation as progressive fusion. The final embedding $z_{\text{fused}} \in \mathbb{R}^d$ is passed to a linear classifier for emotion prediction. AMHF hyperparameters, including the number of gate groups, experts, top-k setting, memory slots, and refinement layers, are listed in Appendix C.3.

3.6 Training Objective

Stage 2 is trained with supervised classification losses for the fused prediction and modality-specific auxiliary heads. The auxiliary heads encourage each modality branch to retain discriminative emotion evidence during joint multimodal training and are later used for post-hoc modality contribution analysis. The total objective is

$$\mathcal{L} = \mathcal{L}_{\text{fused}} + \lambda_{\text{mod}} \frac{\mathcal{L}_v + \mathcal{L}_a + \mathcal{L}_t}{3} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}} + \lambda_{\text{bal}} \mathcal{L}_{\text{bal}}, \tag{11}$$

where $\mathcal{L}_{\text{fused}}$ is the cross-entropy loss for the final fused prediction, and $\mathcal{L}_v$, $\mathcal{L}_a$, and $\mathcal{L}_t$ are cross-entropy losses from the video, audio, and text auxiliary classifier heads. We use $\lambda_{\text{mod}} = 0.3$ in the saved best configuration. The terms $\mathcal{L}_{\text{orth}}$ and $\mathcal{L}_{\text{bal}}$ encourage diversity among emotion-query embeddings and reduce expert collapse in the routing module, respectively. The final prediction is obtained from the fused representation:

$$\hat{y} = \text{softmax}(W z_{\text{fused}} + b).$$

4 Experiments and Results

This section evaluates Emotion-JEPA on the MER2024-SEMI benchmark and analyzes the contribution of its two main design choices: predictive visual adaptation and Audio-Modulated Hybrid Fusion (AMHF). We first describe the evaluation protocol, then compare Emotion-JEPA with large multimodal models and representative MER2024-SEMI systems. We then examine modality contributions, fusion behavior, component ablations, per-class performance, and training dynamics.

4.1 Evaluation Setup

Dataset: We evaluate on MER2024-SEMI (Lian et al., 2024b), a multimodal emotion recognition benchmark containing short conversational clips with video, audio, and textual transcripts. Each clip is annotated with one of six emotion categories: *Angry*, *Happy*, *Neutral*, *Sad*, *Surprise*, and *Worried*. The labeled portion contains 5,030 clips. Following the official protocol, we split this labeled set into 4,275 training clips and 755 validation clips using an 85/15 split. Final evaluation is performed on the MER2024-SEMI test set, which contains 1,169 clips.

MER2024 also provides over 110k unlabeled clips. In Emotion-JEPA, these clips are used only for Stage 1 predictive visual adaptation. They are not used for pseudo-labeling, self-training, voting, or ensembling. This distinction is important because our goal is to isolate the effect of in-domain representation adaptation and fusion design rather than to expand the supervised training set. Additional details on class distribution, preprocessing, and modality-specific data characteristics are provided in Appendix A.

Metrics: Following the MER2024 evaluation protocol, we use Weighted Average F1-score (WAF) as the primary metric because the dataset is class-imbalanced. We also report overall accuracy and Macro-F1 to provide complementary views of classification performance.

Implementation: Emotion-JEPA uses three pretrained modality encoders. The video stream is processed by a V-JEPA2 ViT-Giant encoder adapted in Stage 1 using predictive masked video learning. The audio

stream is encoded with HuBERT-Large, and the text stream is encoded with XLM-RoBERTa-Large (Conneau et al., 2020). Each modality is projected into a shared 512-dimensional embedding space before fusion. During supervised multimodal training, projection heads and AMHF are fully trainable, while only the upper layers of the pretrained modality encoders are fine-tuned.

The full model contains approximately 1.9B parameters, of which approximately 0.6B are updated during Stage 2 training. Details on optimization, hardware, throughput, and reproducibility are provided in Appendices C and F.

4.2 Comparison with Large Multimodal Models

We first compare Emotion-JEPA with open-source large multimodal models, including Apollo, Qwen-VL, and Qwen-Omni. These models provide strong general-purpose multimodal representations and are evaluated under modality configurations supported by each model. For Qwen-Omni, we include low-rank adaptation (LoRA) fine-tuned variants because the model supports audio-conditioned multimodal input. Table 1 shows that Emotion-JEPA achieves 85.72% WAF, outperforming the strongest Qwen-Omni configuration by more than 6 WAF points. The best Qwen-Omni result is obtained with audio and text, while adding video does not improve performance. This suggests that direct use of general-purpose multimodal models may not fully exploit the fine-grained visual dynamics needed for MER, even when parameter-efficient fine-tuning is used. In contrast, Emotion-JEPA benefits from affective visual adaptation and a task-specific fusion module while using fewer total parameters than the 7B LMM baselines.

Table 1: Comparison with open-source large multimodal models on MER2024-SEMI.

Model	Params	Modality	Acc (%)	WAF (%)	Macro-F1 (%)
Apollo 3B	3B	V,T	51.92	51.92	44.84
Apollo 7B	7B	V,T	46.54	45.47	38.19
Qwen-VL 3B	3B	V,T	55.21	55.06	49.57
Qwen-VL 7B	7B	V,T	55.38	55.53	49.33
Qwen-Omni 7B (LoRA)	7B	A,T	79.21	79.16	77.63
Qwen-Omni 7B (LoRA)	7B	A,V,T	78.87	78.56	73.80
Emotion-JEPA (ours)	< 2**B**	**A,V,T**	**85.89**	**85.72**	**84.25**

Prompt templates, modality-specific instructions, and LoRA fine-tuning details for the LMM baselines are provided in Appendix D.

4.3 Comparison with Representative MER2024-SEMI Methods

Table 2 compares Emotion-JEPA with representative MER2024-SEMI systems. Several high-performing systems use pseudo-labeling, self-training, voting, or iterative refinement over the large unlabeled subset. These strategies are effective because they convert unlabeled examples into additional supervisory signal. Emotion-JEPA uses the same unlabeled pool differently: unlabeled clips are used only to adapt the V-JEPA2 visual encoder through masked latent prediction, and no pseudo-labels are assigned to unlabeled examples.

Under this protocol, Emotion-JEPA achieves 85.72% WAF. Although this is below the strongest pseudo-labeling-based systems and slightly below the MERTools baseline reported for MER2024-SEMI, it remains competitive without expanding the labeled training set through pseudo-labeling or ensembling. This result is notable because it suggests that a V-JEPA2-style predictive video representation, originally developed for physical-world video understanding and prediction, can be adapted effectively to affective video understanding. Rather than relying on generative reconstruction or additional pseudo-labeled supervision, the visual branch is adapted through latent prediction on emotion-domain videos and then combined with audio and text through reliability-aware fusion. We therefore view this comparison not as a leaderboard claim, but as evidence that in-domain predictive visual adaptation and audio-conditioned fusion offer a complementary path to improving MER under limited supervision.

Table 2: Comparison with representative MER2024-SEMI systems.

Method	Key Technique	WAF (%)	Pseudo-labeling
MERTools Baseline (Lian et al., 2024b)	HuBERT + CLIP + cross-attention fusion	86.73	No
EmoVCLIP (Qi et al., 2024)	V-L prompting + self-training	90.51	Yes
Shi & Gao (Shi & Gao, 2024)	Semi-supervised training	89.83	Yes
Ge et al. (Ge et al., 2024)	Audio-text fusion + voting	88.25	Yes
Emotion-JEPA (ours)	Predictive adaptation + AMHF fusion	85.72	No

4.4 Transfer to MER2025-SEMI

To evaluate transfer beyond MER2024-SEMI, we further test Emotion-JEPA on MER2025-SEMI, which follows the same six-way categorical MER setting. We consider direct transfer from the MER2024-trained checkpoint and target-domain adaptation by fine-tuning the same checkpoint on the MER2025 training set.

Table 3: MER2025-SEMI transfer and target-domain adaptation results

Method / setting	MER2025 training	WAF (%)	Acc. (%)
Official MER2025 baseline (Lian et al., 2025b)	Yes	78.63	78.77
Emotion-JEPA, direct transfer	No	74.47	74.98
Emotion-JEPA, fine-tuned	Yes	**79.15**	**79.07**

Table 3 reports results on the MER2025-SEMI test set of 2,026 clips. The MER2024-trained model obtains 74.47% WAF without MER2025 training, showing reasonable transfer to the newer MER2025-SEMI benchmark. After fine-tuning on MER2025, Emotion-JEPA improves to 79.15% WAF, slightly exceeding the official multimodal baseline. We use this experiment as a transfer and adaptation analysis rather than a state-of-the-art comparison, since many MER2025 challenge systems use pseudo-labeling, label refinement, ensembling, or additional auxiliary cues.

4.5 Modality Contributions and Fusion Behavior

Table 4 reports modality-specific auxiliary-head performance alongside the full multimodal prediction. These auxiliary heads are trained jointly with the fused classifier through the modality loss in Eq. 11, they are not separately trained unimodal models. Audio provides the strongest individual signal, reaching 75.74% WAF, while the video and text auxiliary heads reach 44.37% and 18.42% WAF, respectively. The weak text performance is consistent with the dataset characteristics: many transcripts are short, incomplete, weakly emotional, or not reliably aligned with the moment of peak affect, as further discussed in Appendix A.3.

Despite this modality imbalance, the full multimodal model achieves 85.72% WAF, improving by 9.98 WAF over the audio auxiliary head. This indicates that visual and textual cues still provide complementary information when their contribution is regulated by an effective fusion mechanism.

Table 4: Auxiliary-head unimodal and full multimodal performance on MER2024-SEMI. Unimodal rows are obtained from modality-specific auxiliary classifier heads trained jointly with the fused classifier, not from separately trained unimodal models.

Setting	Acc (%)	WAF (%)	Macro-F1 (%)
Video auxiliary head	47.82	44.37	35.70
Audio auxiliary head	76.13	75.74	74.74
Text auxiliary head	23.61	18.42	17.57
Full multimodal model	**85.89**	**85.72**	**84.25**

We further compare AMHF with a standard cross-attention fusion baseline under the same encoders and training protocol. Replacing AMHF with cross-attention reduces WAF from 85.72% to 78.47%, suggesting

that symmetric fusion is less effective when modality reliability varies across samples. AMHF improves performance by using audio to guide how visual and textual evidence are incorporated, especially when text is sparse or visual cues are ambiguous.

4.6 Ablation Studies

To quantify the source of the gains, we conduct two sets of ablations on MER2024-SEMI: one over the main framework components and another over the internal AMHF stages. For controlled comparison, both tables use the conservative archived-best AMHF result as the full-model reference.

Framework-level ablations: Table 5 summarizes the contribution of the main framework components. Freezing all encoders causes the largest degradation, reducing WAF from 85.72% to 72.07%, which indicates that pretrained representations require task-specific adaptation for MER. Removing Stage 1 predictive visual adaptation reduces WAF by 7.92 points, supporting the role of in-domain visual adaptation. Replacing AMHF with cross-attention reduces WAF by 7.25 points, showing that the proposed fusion design contributes beyond the choice of encoders. Face-centric preprocessing and emotion-aware pooling also provide measurable gains.

Table 5: Framework-level ablation on MER2024-SEMI.

Configuration	WAF (%)	Δ	$\Delta_{\%}$
Full Model	**85.72**	-	-
No vEAP	85.45	-0.27	-0.32%
No vEAP + No EATP	82.46	-3.26	-3.80%
No Face Detection	80.84	-4.88	-5.69%
Cross-attention fusion	78.47	-7.25	-8.46%
No Domain Adaptation	77.80	-7.92	-9.24%
Frozen Encoders	72.07	-13.65	-15.92%

AMHF component ablations: Table 6 analyzes the internal design of AMHF. Removing audio spectral gating reduces WAF by 3.28 points, suggesting that audio-conditioned gating helps regulate noisy visual and textual features. Removing adaptive cross-modal routing causes a larger drop of 5.56 WAF, indicating that a fixed fusion pathway is less effective for heterogeneous multimodal interactions. Temporal memory and uncertainty estimation each reduce WAF by 5.33 points when removed, showing that temporal stabilization and reliability estimation both contribute to robust prediction. The largest drop occurs without progressive fusion, which reduces WAF by 8.24 points, suggesting that staged integration of modality evidence is central to AMHF.

Table 6: AMHF component ablation on MER2024-SEMI.

Configuration	WAF (%)	Δ	$\Delta_{\%}$
Full AMHF	**85.72**	-	-
No Audio Spectral Gating	82.44	-3.28	-3.83%
No Adaptive Cross-Modal Routing	80.16	-5.56	-6.49%
No Temporal Memory	80.39	-5.33	-6.22%
No Uncertainty Estimation	80.39	-5.33	-6.22%
No Progressive Fusion	77.48	-8.24	-9.61%

4.7 Per-Class Behavior and Error Patterns

Figure 2 shows the normalized confusion matrix on MER2024-SEMI. Emotion-JEPA performs strongly on *Happy*, *Sad*, and *Angry*. Most remaining errors occur among *Neutral*, *Worried*, and *Surprise*, where af-

fective differences are often subtle and may depend on short temporal expressions or prosodic cues. The model does not collapse toward the dominant *Neutral* class, suggesting that affect-adapted visual features and audio-modulated fusion help preserve discrimination among low-arousal categories. Additional diagnostics, including per-class metrics, Top-K analysis, qualitative examples, and Stage 2 training dynamics, are provided in Appendices E and C.6.

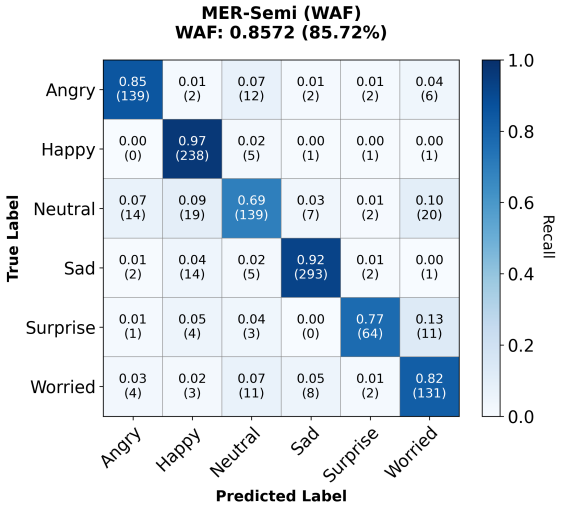


Figure 2: Normalized confusion matrix on MER2024-SEMI.

4.8 Discussion

The results point to three main observations. First, in-domain predictive adaptation improves MER performance: removing Stage 1 reduces WAF by 7.92 points, indicating that generic video features do not fully capture affect-relevant facial dynamics. Second, fusion design matters under fixed encoders: replacing AMHF with cross-attention reduces WAF by 7.25 points, and AMHF component ablations show that routing, temporal memory, uncertainty estimation, and progressive fusion each contribute. Third, scale alone is insufficient for this setting, as the evaluated LMM baselines remain below Emotion-JEPA even with LoRA fine-tuning. At the same time, pseudo-labeling and self-training remain strong tools for maximizing MER2024-SEMI performance, so Emotion-JEPA is best viewed as a controlled study of representation adaptation and reliability-aware fusion rather than as a replacement for semi-supervised pipelines.

5 Conclusion and Future Work

This paper studied two factors often entangled in multimodal emotion recognition: affective representation adaptation and fusion design. Emotion-JEPA first adapts a visual encoder to emotion-domain videos through masked latent prediction, then performs supervised multimodal classification with Audio-Modulated Hybrid Fusion (AMHF). Unlabeled videos are used only for predictive visual adaptation, without pseudo-labeling or self-training. Experiments on MER2024-SEMI show that predictive visual adaptation improves affect-sensitive representations and that AMHF improves over a capacity-matched cross-attention baseline under identical encoders and training protocols. Comparisons with large multimodal models further suggest that model scale alone does not guarantee strong performance on fine-grained affect recognition.

Limitations and future work: Emotion-JEPA relies on independently pretrained visual, audio, and text encoders whose objectives are not explicitly aligned for affective reasoning. The method also benefits from face-centric preprocessing and reasonably synchronized audio-video inputs, which may limit robustness in unconstrained settings with missing faces, off-screen speakers, noisy audio, or temporal misalignment. Future work should explore joint multimodal self-supervised adaptation, robustness to missing or asynchronous modalities, and transfer to fine-grained, open-vocabulary, and descriptive emotion recognition benchmarks.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL `https://arxiv.org/abs/2308.12966`.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video, 2024. URL `https://arxiv.org/abs/2404.08471`.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359, 12 2008. doi: 10.1007/s10579-008-9076-6.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 110805–110853. Curran Associates, Inc., 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/c7f43ada17acc234f568dc66da527418-Paper-Conference.pdf`.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451, 2020.

Lennart Eing, Cristina Luna-Jiménez, Silvan Mertes, and Elisabeth André. Video joint-embedding predictive architectures for facial expression recognition. *arXiv preprint arXiv:2601.09524*, 2026.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14314–14324, June 2025.

Mengying Ge, Mingyang Li, Dongkai Tang, Pengbo Li, Kuo Liu, Shuhao Deng, Songbai Pu, Long Liu, Yang Song, and Tao Zhang. Early joint learning of emotion information makes multimodal model understand you better. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 54–61, 2024.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, et al. Ov-mer: Towards open-vocabulary multimodal emotion recognition. *arXiv preprint arXiv:2410.01495*, 2024a.

Zheng Lian, Haiyang Sun, Licai Sun, Zhuofan Wen, Siyuan Zhang, Shun Chen, Hao Gu, Jinming Zhao, Ziyang Ma, Xie Chen, et al. 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*, pp. 41–48, 2024b.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025a.

Zheng Lian, Rui Liu, Kele Xu, Bin Liu, Xuefei Liu, Yazhou Zhang, Xin Liu, Yong Li, Zebang Cheng, Haolin Zuo, et al. Mer 2025: When affective computing meets large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13837–13842, 2025b.

Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp. 472–476, 2018.

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. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Anbin Qi, Zhongliang Liu, Xinyong Zhou, Jinba Xiao, Fengrun Zhang, Qi Gan, Ming Tao, Gaozheng Zhang, and Lu Zhang. Multimodal emotion recognition with vision-language prompting and modality dropout. In *Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing*, pp. 49–53, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Pujin Shi and Fei Gao. Audio-guided fusion techniques for multimodal emotion analysis. *arXiv preprint arXiv:2409.05007*, 2024.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1656. URL https://aclanthology.org/P19-1656/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

Amir Zadeh, Minghai Chen, Soujanya Poria, E. Cambria, and Louis philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing*, 2017. URL https://api.semanticscholar.org/CorpusID:950292.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 543–553, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.49. URL https://aclanthology.org/2023.emnlp-demo.49/.

Appendix

A Dataset Description

A.1 Labeled and Unlabeled Splits

We use the official MER2024 dataset. The labeled portion contains 5,030 clips, each with synchronized video, audio, and text transcripts. Following the competition protocol, we use 85% of the labeled set for training and 15% for validation, stratified by emotion class. Final evaluation is performed on the MER2024-SEMI test set with 1,169 labeled clips. The additional unlabeled set contains approximately 113k clips and is used only for Stage 1 self-supervised visual adaptation.

Each clip is 4-6 seconds long and is annotated with one of six emotions: *Angry*, *Happy*, *Neutral*, *Sad*, *Surprise*, and *Worried*. We do not use pseudo-labeling or self-training on the unlabeled or test sets.

Table 7 summarizes the class distribution for the 5,030-sample labeled set and the MER-SEMI test set.

Table 7: Class distribution for the MER2024 labeled set (5,030) and MER2024-SEMI test set (1,169).

Emotion	**MER2024 Labeled (5,030)**		**MER2024-SEMI Test (1,169)**	
	Count	**Pct.**	**Count**	**Pct.**
Sad	1,267	25.18%	317	27.12%
Happy	1,042	20.72%	246	21.04%
Neutral	901	17.92%	201	17.19%
Angry	703	13.98%	163	13.94%
Worried	670	13.32%	159	13.60%
Surprise	447	8.89%	83	7.10%
Total	5,030	100%	1,169	100%

A.2 Per-Modality Preprocessing

Video: Clips are decoded at 4 FPS and resized to 256×256. We detect faces with MediaPipe using a confidence threshold of 0.5 and expand the bounding box by 40% to include head pose and shoulder context. If no face is detected, a centered crop is used. During training, we apply light augmentation: random resized cropping with scale in $[0.7, 1.0]$, aspect-ratio jitter in $[0.9, 1.1]$, and horizontal flipping.

Audio: Audio is resampled to 16 kHz mono and truncated or padded to match the clip duration. We optionally apply cepstral mean and variance normalization (CMVN) to reduce channel and loudness variability in heterogeneous recordings.

Text: We use the official Chinese and English transcripts. Text is normalized and tokenized with XLM-RoBERTa-Large. We do not enforce forced alignment, text is treated as an utterance-level summary and bypasses temporal alignment in Stage 2.

A.3 Text Modality Characteristics and Limitations

Although MER2024 provides Chinese and English transcripts for all clips, the text modality has several dataset-specific limitations for fine-grained emotion recognition.

Short and weakly emotional utterances: As shown in Figure 3, most transcripts are short. The median length is 16 words for Chinese and 13 words for English in the training split, and 9 words for Chinese and 5 words for English in MER2024-SEMI. Many utterances contain only a few semantically neutral words, and spoken content is often task-oriented or contextually vague rather than emotionally expressive. This limits the discriminative power of the text encoder.

Missing, partial, or weakly aligned transcriptions: Some clips contain truncated text, transcription errors, missing segments, or ambiguous speaker boundaries. In addition, transcripts are not force-aligned with audio or video and often do not correspond to the moment of peak facial or vocal expression. For this reason, we treat text as a global contextual cue rather than a temporally aligned signal in Stage 2.

Impact on fusion: These properties help explain the weak text auxiliary-head performance reported in Table 4. Text is therefore incorporated as a supporting modality, while AMHF can reduce reliance on text when acoustic and visual cues provide stronger evidence.

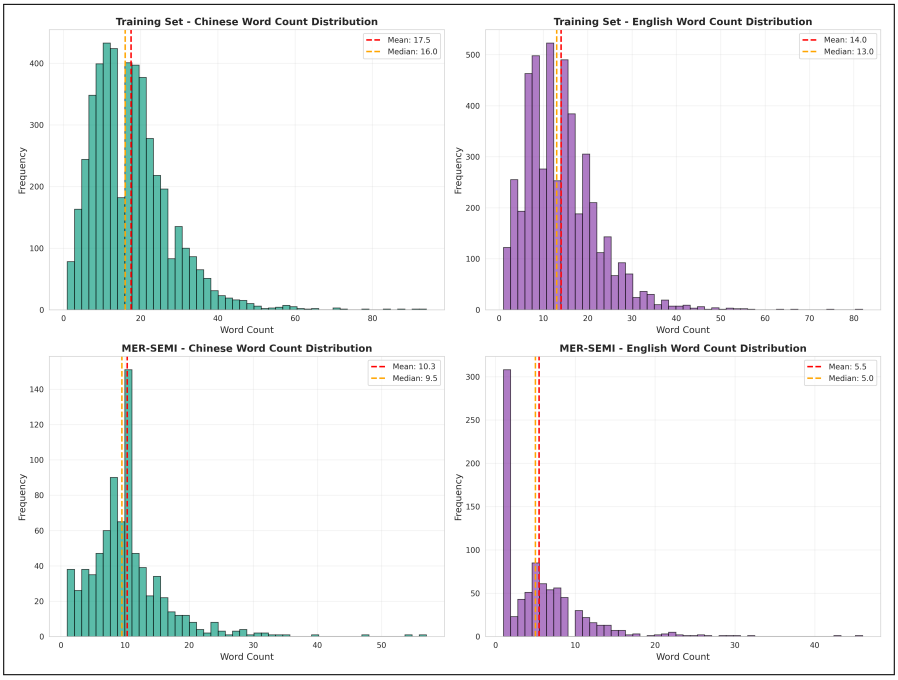


Figure 3: Word-count distributions for Chinese and English transcripts in the MER2024 training and MER2024-SEMI test splits. Most transcripts are short, which contributes to weak unimodal text discriminability.

B Additional Details for Stage 1: Self-Supervised Visual Adaptation

B.1 Model Overview and Trainable Subset

Stage 1 adapts the V-JEPA2 ViT-Giant backbone to the MER2024 video domain using masked predictive learning. In our configuration, the V-JEPA2 encoder contains 1,012,173,952 parameters (≈1.01B), and the predictor network contains 22,381,312 parameters (≈22M), for a total of approximately 1.03B parameters.

During domain adaptation, we partially unfreeze the encoder and train the predictor:

- the first 32 of 40 encoder layers remain frozen;
- the top 8 encoder layers are unfrozen;
- the predictor network is fully trainable;
- the EMA target encoder is updated without backpropagation.

Table 8 summarizes the parameter and compute breakdown for Stage 1. We use × to indicate non-trainable components and ✓ to indicate trainable ones.

Table 8: V-JEPA2 components active during Stage 1 domain adaptation. × indicates non-trainable components and ✓ indicates trainable ones. FLOPs are forward-pass estimates.

Component	**Params**	**Trainable**	**% of Total**	**FLOPs/sample** (Estimate)
Frozen encoder blocks (layers 1-32)	810M	×	78%	∼140 GF
Trainable encoder blocks (layers 33-40)	202M	✓	20%	∼35 GF
Predictor network	22M	✓	2%	∼10 GF
Total (encoder + predictor)	**1.03B**	-	**100%**	∼185 GF
Trainable subset	**224M**	-	**22%**	∼45 GF

Note: FLOP values in Table 8 are forward-pass estimates for individual components and are not used as training-time compute indicators. Practical cost is reported via throughput and GPU-hours in Appendix B.5.

B.2 Masking Strategy and Sampling

We use a hierarchical masking policy with 14 spatiotemporal blocks per clip. The masks cover approximately 60-70% of the spatiotemporal volume and are biased to preserve facial regions, as summarized in Table 9. The masking hyperparameters are tuned for MER2024's short 4-6 second clips sampled at 4 FPS.

Table 9: Multi-scale masking configuration used during Stage 1 visual adaptation.

Block Type	**#**	**Spatial**	**Temporal**	**Aspect**	**Purpose**
Large temporal	6	6-10%	85-100%	1.2-2.0	Long-range motion
Medium temporal	4	8-12%	85-100%	0.5-0.9	Mid-range dynamics
Medium spatial	3	18-25%	70-100%	0.8-1.2	Structure + motion
Large spatial	1	35-45%	80-100%	0.75-1.5	Broad spatial context
Total	**14**	**60-70%**	**80-100%**	-	Multi-scale representation

B.3 EMA Target Network

We use the standard JEPA student-teacher setup, where a momentum-updated target encoder provides the prediction targets:

$$\theta_{\text{tgt}} \leftarrow m\theta_{\text{tgt}} + (1-m)\theta_{\text{stu}}, \qquad m = 0.9995.$$

Only the student encoder participates in backpropagation and the EMA target encoder stabilizes learning and prevents collapse.

B.4 Training Configuration

Stage 1 uses the following settings:

- Batch size: 24 per GPU, 96 effective across 4 GPUs.
- Optimizer: AdamW with weight decay 0.04.
- Learning rate: 2×10^{-4} with cosine decay and 2-epoch warmup.
- Precision: BF16.
- Augmentations: random resized cropping with scale in $[0.7, 1.0]$, aspect-ratio jitter in $[0.9, 1.1]$, and horizontal flipping.
- Duration: 6 epochs, continuing from epochs 41-46 of a pretrained V-JEPA2 checkpoint with iterations per epoch (IPE) of 800.

B.5 Compute and Throughput

Stage 1 visual adaptation was conducted on 4×A100 40GB GPUs. Since only the top 8 encoder blocks and the predictor are trainable, this stage is substantially lighter than full-network pretraining. All measurements are taken from the steady-state training phase.

- Throughput: approximately 95 clips/s.
- Peak memory: approximately 28 GB per GPU.
- Total compute: approximately 164 GPU-hours for 6 adaptation epochs with IPE of 800.

The FLOP estimates in Table 8 are forward-pass estimates only. Practical training cost is better reflected by throughput and GPU-hours.

B.6 Impact on Downstream Performance

Stage 1 visual adaptation produces a substantial gain in supervised MER performance:

$$\mathrm{WAF}_{\text{with DA}} = 85.72\%, \qquad \mathrm{WAF}_{\text{without DA}} = 77.80\%.$$

This corresponds to an improvement of +7.92 WAF on MER2024-SEMI.

C Additional Details for Stage 2: Multimodal Training

C.1 Modality Encoders

Video: We use the domain-adapted V-JEPA2 ViT-Giant encoder from Stage 1. Face-centric crops are resized to 256×256 and fed as 16-frame clips. The last eight Transformer blocks and the predictor remain trainable.

Audio: HuBERT-Large (Hsu et al., 2021) processes 16 kHz waveforms with a 25 ms window and 10 ms hop. We extract hidden states from the last eight Transformer layers and apply mean pooling across time to obtain a 1024-D representation.

Text: XLM-RoBERTa-Large (Conneau et al., 2020) encodes Chinese and English transcripts. We average hidden states from the final eight layers and mean-pool across tokens, yielding a 1024-D textual embedding.

C.2 Projection and Shared Embedding Space

Each modality output is projected into a shared d=512-dimensional space:

$$\mathrm{Proj}(x) = \mathrm{GELU}\big(\mathrm{LN}(Wx)\big),$$

where W is a learnable linear mapping and LN is LayerNorm. The projected features $(\mathbf{v}, \mathbf{a}, \mathbf{t}) \in \mathbb{R}^d$ are used by modality-specific auxiliary classifiers and as inputs to the Audio-Modulated Hybrid Fusion (AMHF) module.

C.3 AMHF Hyperparameters

AMHF operates in the shared embedding space with the following configuration:

- **Audio spectral gates:** G=8 groups, each in $\mathbb{R}^{64}$.
- **Routing experts:** E=4 experts with hidden size h=512; top-k=2 experts are selected per sample.
- **Temporal memory:** M=10 memory slots per emotion class.
- **Refinement Transformer:** 3 encoder layers, 8 heads, hidden size h=512.

C.4 Training Configuration

Stage 2 uses the following settings:

- **Batch size:** 12 per GPU with 3× gradient accumulation, giving an effective batch size of 144 across 4 GPUs.
- **Optimizer:** AdamW with weight decay 0.05.
- **Learning rate:** 2×10^{-4} with warmup-cosine decay, 3 warmup epochs, and minimum LR 1×10^{-6}.
- **Layer-wise LR decay:** factor 0.95 applied to all encoders.
- **Precision:** BF16.
- **Regularization:** dropout 0.15, stochastic depth 0.2, label smoothing 0.1, mixup ($\alpha = 0.2$), and cutmix ($\alpha = 0.3$).
- **Training duration:** 80 epochs with early stopping and the best checkpoint is selected at epoch 14.

C.5 Compute and Throughput

Stage 2 multimodal training was performed on the same 4×A100 40 GB workstation used in Stage 1, hardware details are provided in Appendix F. Because only the top visual blocks, selected HuBERT and XLM-RoBERTa-Large layers, and AMHF are trainable, Stage 2 is lighter than full-network training but heavier than Stage 1 due to the multimodal forward pass.

- **Throughput:** approximately 32-35 clips/s aggregate.
- **Peak memory:** approximately 28-30 GB per GPU.
- **Training duration:** approximately 9-11 GPU-hours for a full run, with the best checkpoint selected after approximately 2.5 hours.

Given the partially frozen encoders and multimodal architecture, FLOP counts are not reported. Throughput and GPU-hours provide a more reliable measure of practical compute.

C.6 Training Dynamics

Figure 4 presents the Stage 2 training curves, including WAF, accuracy, Macro-F1, precision-recall, loss trajectories, and the learning-rate schedule. Validation metrics rise steadily during the first 10-15 epochs, and no divergence or unstable optimization behavior is observed.

D LMM Evaluation Details

All LMM baselines evaluated in this work, including Apollo, Qwen-VL, and Qwen-Omni, accept raw video input directly. Qwen-Omni additionally supports raw audio input. Therefore, we do not apply manual windowing, frame segmentation, or model-specific sampling heuristics and each model processes the full clip as provided.

D.1 Prompt Templates

To ensure a fair comparison across LMMs, we use a unified modality-aware instruction template. The A+V+T, A+T, and V+T variants share the same output format and differ only in the sensory cues made available to the model.

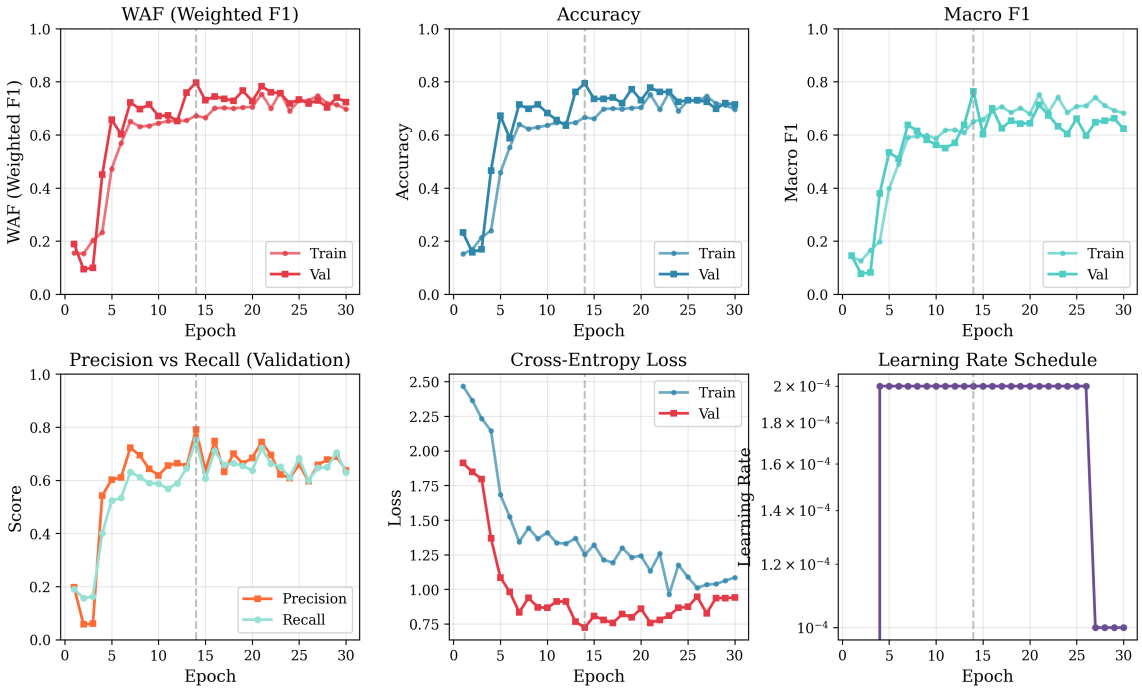


Figure 4: Extended Stage 2 training curves showing WAF, accuracy, Macro-F1, precision-recall, loss, and learning-rate schedule.

Unified template (abridged):

> *"You are an expert in emotion recognition. Based on the provided modality input, classify the dominant emotional state as one of:* ***happy, sad, angry, worried, surprised, neutral****. Respond with one word."*

Modality-specific cues:

- **Audio-Video-Text (A,V,T):** The model is instructed to consider facial expressions, body language, eye behavior, gestures, and tone of voice.
- **Audio-Text (A,T):** Vision cues are removed, while prosodic descriptors such as pitch, speaking rate, and pauses are retained.
- **Video-Text (V,T):** Audio cues are removed, while facial, gesture, and body-language descriptors are retained.

Full verbatim prompts: Full verbatim prompt strings used for LMM inference under different settings are provided as plain-text files in the supplementary archive under `appendix/prompts/`.

D.2 LoRA Fine-Tuning Configuration for Qwen-Omni

We fine-tune Qwen2.5-Omni 3B and 7B with parameter-efficient LoRA adapters using the `LLaMA-Factory` framework. We use the `mer-train` split with 5,030 labeled clips formatted in the multimodal `ShareGPT` conversation style. Each training example contains a user instruction requesting an emotion prediction with `<video>` and `<audio>` tags, a single-word assistant response containing the ground-truth emotion, and top-level fields specifying the corresponding video and audio paths.

Adapter parameters:

- Rank r: 16.
- Scaling factor α: 16.
- Target modules: attention and MLP projections, expanded from `"all"` to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `mlp.gate_proj`, `mlp.up_proj`, and `mlp.down_proj`.

Training hyperparameters:

- Learning rate: 1.0×10^{-5} with AdamW, cosine schedule, and warmup ratio 0.1.
- Precision: FP16.
- Per-device batch size: 1 with gradient accumulation of 4, giving an effective batch size of 8 on 2 GPUs.
- Epochs: 2.0, corresponding to approximately 2-3 effective passes depending on the internal train/validation split.

Hardware: Fine-tuning runs are executed on up to 2×NVIDIA A100 40 GB GPUs, with typical VRAM usage in the 20-40 GB range depending on model size and checkpointing configuration.

E Additional Results

E.1 Per-Class Metrics

Figure 5 reports per-class precision, recall, and F1 scores for the best Emotion-JEPA model on MER2024-SEMI.

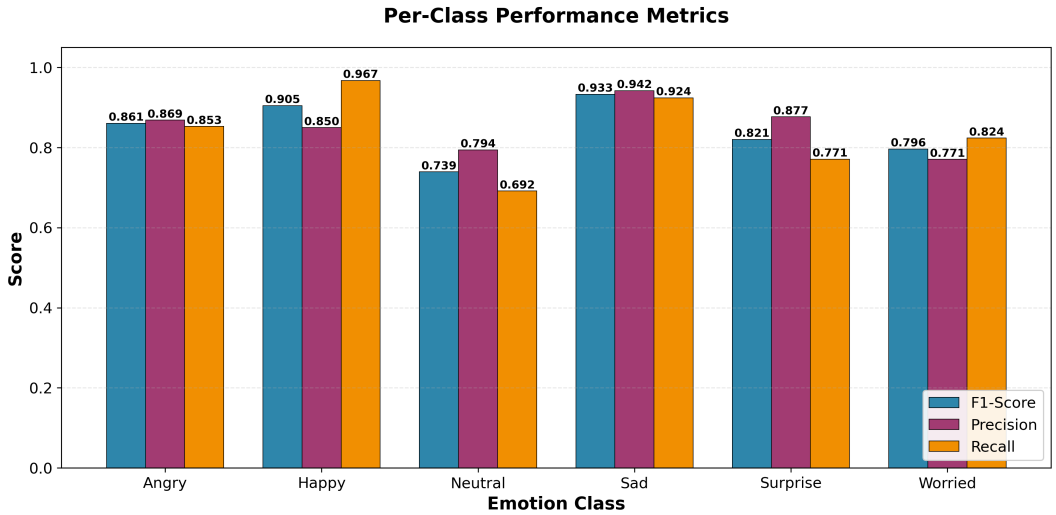


Figure 5: Per-class precision, recall, and F1 scores on MER2024-SEMI.

E.2 Oracle Top-K Analysis

In addition to standard Top-1 evaluation, we report Top-K metrics under an oracle protocol for single-label classification. A prediction is counted as correct if the ground-truth class appears among the model's Top-K

probability-ranked outputs. This does not change the task into multi-label evaluation, instead, it measures how often the correct label remains among the model's most likely alternatives.

Table 10 summarizes Oracle Top-K WAF and accuracy on MER2024-SEMI. Moving from Top-1 to Top-2 improves WAF by 9.33 points, and Top-3 improves WAF by 11.76 points, indicating that many remaining errors are near-miss ambiguities rather than arbitrary misclassifications.

Table 10: Oracle Top-K metrics on MER2024-SEMI. Top-1 corresponds to standard single-label evaluation.

Metric	**Top-1**	**Top-2**	**Top-3**
WAF (%)	85.72	95.52	97.93
Accuracy (%)	85.89	95.22	97.65

Figure 6 visualizes the per-class Top-K trends. The largest gains appear for *Neutral*, *Worried*, and *Angry*, where the ground-truth label is frequently ranked second or third.

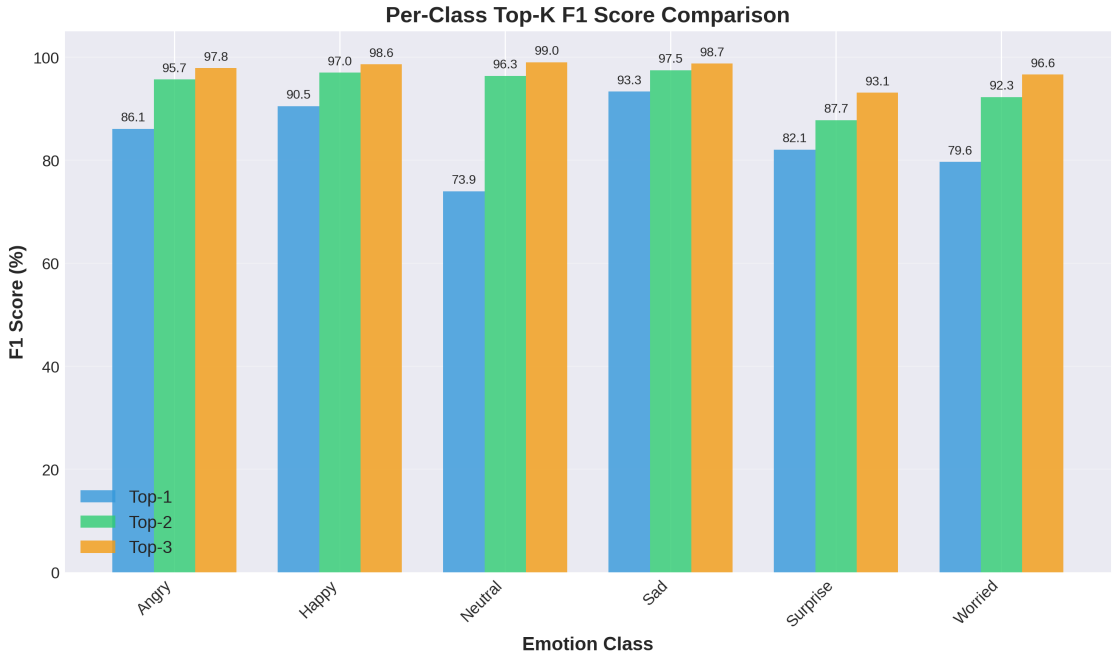


Figure 6: Per-class Top-K F1 scores on MER2024-SEMI.

E.3 Qualitative Success and Failure Cases

To illustrate model behavior, Figure 7 presents two success cases and two failure cases from MER2024-SEMI. Each example includes sampled video frames, the corresponding waveform and spectrogram, the transcript when available, and the model's probability distribution over the six emotion classes.

Successful predictions: The model performs reliably when visual affect, prosody, and textual cues are well aligned. For example, `samplenew3_00104570` (*Happy*) and `samplenew3_00032314` (*Sad*) show cases where facial expressions and audio patterns provide clear emotional evidence. These examples also illustrate robustness under modality imbalance: even when transcripts are unavailable, the model uses visual dynamics and audio cues to make confident predictions.

Failure cases: Most errors occur in ambiguous regions where facial motion, prosody, and conversational context provide weak or conflicting emotional signals. A common pattern is confusion between *Neutral*

and low-arousal states. In `samplenew3_00097648`, mild vocal tension leads to a *Worried* prediction despite largely neutral visual cues. Similarly, `samplenew3_00026688` is misclassified as *Sad*, likely due to low lighting and an introspective gaze that reduce the salience of neutral facial features.

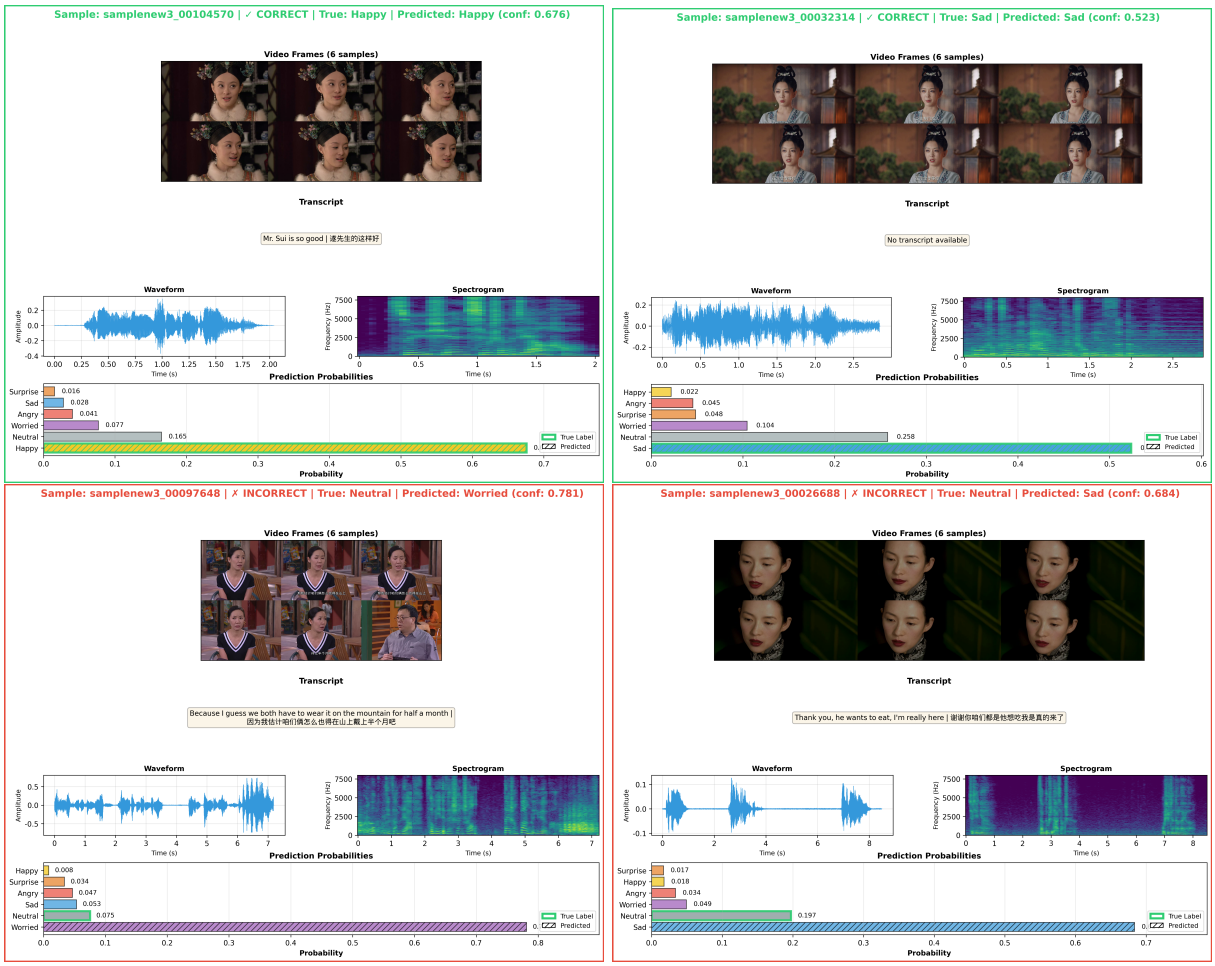


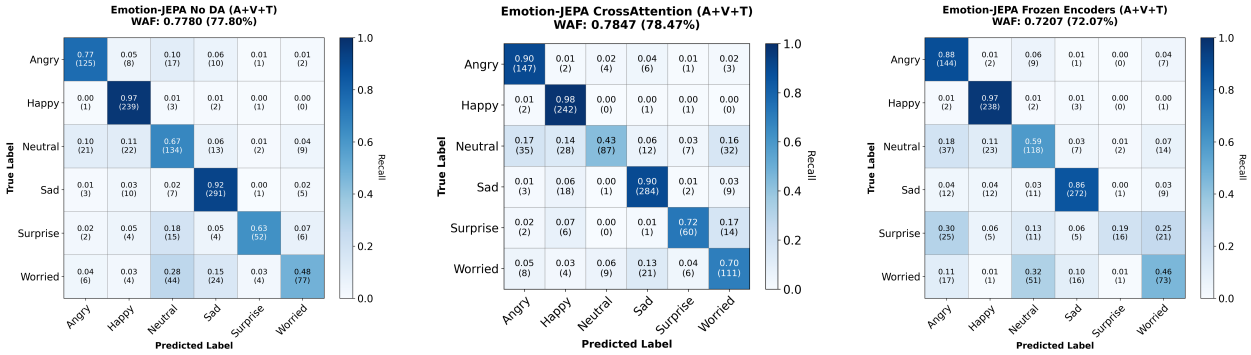
Figure 7: Qualitative examples from MER2024-SEMI. Top: success cases where Emotion-JEPA correctly identifies the target emotion using visual cues and audio prosody. Bottom: failure cases showing low-arousal ambiguity and modality conflict between facial expression and vocal tone.

Interpretation: Across failure cases, the true label often appears among the Top-K ranked predictions, consistent with the analysis in Appendix E.2. This suggests that many remaining errors reflect intrinsic ambiguity among low-arousal or closely related emotions. When prosodic signals are weak or conflicting, audio-modulated routing in AMHF becomes less decisive, producing more diffuse confidence distributions. These examples point to future directions such as stronger conflict-aware modality re-weighting.

E.4 Additional Confusion Matrices

Ablation confusion matrices: To visualize how major components affect inter-class confusions, Figure 8 shows normalized confusion matrices for three representative ablations: removing domain adaptation, replacing AMHF with cross-attention, and freezing all encoders.

LMM baselines vs. Emotion-JEPA: Figure 9 compares normalized confusion matrices for the best-performing Qwen-Omni 7B A,T baseline in zero-shot and LoRA-tuned settings against Emotion-JEPA.



(a) No domain adaptation (b) Cross-attention fusion (c) Frozen encoders

Figure 8: Normalized confusion matrices for major ablations on MER2024-SEMI.

F Compute and Reproducibility

F.1 Hardware and Runtime

Emotion-JEPA experiments were conducted on a workstation equipped with 4×NVIDIA A100-SXM4 40 GB GPUs. The software environment was:

- **OS:** Ubuntu 22.04.5 LTS.
- **GPU driver:** 550.144.03.
- **CUDA:** 12.4 system installation; 11.8 PyTorch runtime.
- **Frameworks:** Python 3.12, PyTorch 2.1, cuDNN 8.7.

Approximate runtimes are:

- **Stage 1 visual adaptation:** approximately 164 GPU-hours for 6 epochs with 24 clips/GPU, giving an effective batch size of 96.
- **Stage 2 multimodal training:** approximately 9-11 GPU-hours on 4×A100 GPUs, with the best checkpoint typically selected after approximately 2.5 hours via early stopping.

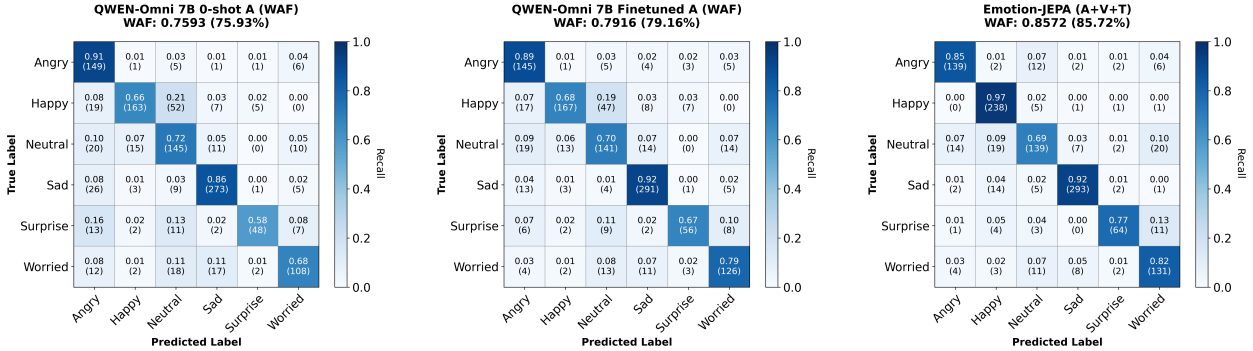
Peak GPU memory usage during Stage 2 was approximately 28-32 GB per GPU. Throughput and memory measurements were collected using `nvidia-smi` and the PyTorch profiler.

F.2 Determinism and Seeding

We fix the global random seed to 42 for all Emotion-JEPA experiments. Determinism is enabled where supported:

- `torch.use_deterministic_algorithms(True)`.
- cuDNN deterministic mode enabled and benchmark mode disabled.
- Reproducible dataloader seeds using `worker_seed = base_seed + worker_id`.

Some low-level CUDA kernels, such as atomic operations inside attention modules, may remain nondeterministic.



(a) Qwen-Omni 7B A,T zero-shot (b) Qwen-Omni 7B A,T LoRA (c) Emotion-JEPA

Figure 9: Normalized confusion matrices for LMM baselines and Emotion-JEPA on MER2024-SEMI.

F.3 Implementation Notes

Emotion-JEPA builds on the official V-JEPA2 codebase and standard PyTorch-based libraries, including `timm`, `transformers`, `librosa`, and `mediapipe` for face detection and alignment. LoRA fine-tuning of Qwen2.5-Omni models is performed with the `LLaMA-Factory` toolkit, which provides parameter-efficient adaptation and unified multimodal preprocessing. Hyperparameters for optimizer settings, learning-rate schedules, masking policies, partial-freezing rules, AMHF configuration, and LoRA fine-tuning are provided in the main paper and appendix to support reproducible implementation and evaluation.