DualCap: Dual-Space Understanding with Multi-Model Feedback for Video Emotion Captioning

Anonymous ACL submission

Abstract

Generating emotionally aligned language remains a key challenge for large language models. We present **DualCap**, a multimodal framework that formulates affective understanding as a generation task rather than discrete classification. DualCap performs dual-space reasoning by integrating surface-level multimodal cues with psychologically grounded Valence–Arousal–Dominance (VAD) representations embedded in hyperbolic space. To ensure emotional and linguistic coherence, we introduce a multi-model feedback mechanism where multiple LLMs collaboratively evaluate and refine captions through aggregated affective and dimensional feedback, analogous to affect-oriented reinforcement learning from human feedback (Affective-RLHF). Experiments on DFEW and MAFW show that DualCap achieves strong recognition performance while substantially improving the expressiveness, interpretability, and emotional fidelity of generated language, demonstrating the value of combining cognitive emotion modeling with feedback-driven generation for emotionally intelligent LLMs.

1 Introduction

Understanding and generating emotionally aligned language is fundamental to affect-aware human–computer interaction (HCI) systems (Cheng et al., 2017a,b). In real-world settings, emotional expression is inherently multimodal—combining visual (Li et al., 2024b; Le Ngwe et al., 2024; Wang et al., 2020), acoustic (Fan et al., 2021; Hsu et al., 2021; Kondratenko et al., 2022), and linguistic cues (Devlin et al., 2019; Lei et al., 2023; Hung and Alias, 2023). Integrating such heterogeneous signals enables applications in social robotics (Chen et al., 2019; Heredia et al., 2022; Chen et al., 2021), mental health (Cao and Wan, 2020; Hutchison and Gerstein, 2017), education (Imani and Montazer, 2019), and driver safety.

The challenge lies in fusing asynchronous and sometimes conflicting cues (Wang et al., 2023b; Han et al., 2024; Uddin et al., 2022). Conventional models predict discrete categories (Verma et al., 2020; Zhang et al., 2021; Jia et al., 2021), which overlook emotion intensity, co-occurrence, and dynamics. Human affect is typically continuous and mixed, making single-label classification insufficient.

Recent work shifts from fixed labels to natural language descriptions. SECap (Xu et al., 2024) projects speech emotion features into LLaMA representations via a Q-Former, but mainly exploits acoustic cues. Similar models (Yang et al., 2023; Zhang et al., 2023a) underuse visual and dimensional affective signals essential for holistic reasoning.

Multimodal large language models (MLLMs) (Zhao et al.; Cheng et al., 2024) enable joint reasoning over images, speech, and text. However, most depend on shallow feature fusion, limiting interpretability and introducing emotional bias. Capturing subtle, evolving affective states demands deeper psychological grounding.

Temporal architectures such as MMA-DFER (Chumachenko et al., 2024) enhance dynamic emotion analysis on DFEW (Jiang et al., 2020) and MAFW (Liu et al., 2022a), yet remain tied to observable cues (e.g., facial motion, vocal pitch) rather than intrinsic affect representations.

We address these limitations with a dimensionally informed framework that unifies emotion recognition and language generation. Using continuous Valence–Arousal–Dominance (VAD) representations (Zhao et al., 2022; Sahu, 2019; Ye et al., 2023), DualCap embeds psychological principles into deep multimodal learning for interpretable, cognitively consistent affective modeling.

DualCap performs dual-space reasoning—integrating surface-level multimodal cues with hyperbolic VAD embeddings—and

refines generated captions via multi-model feedback, where several LLMs collaboratively evaluate and improve affective alignment, analogous to affect-oriented RLHF.

Our contributions are:

- **Affective Language Generation:** Reformulates multimodal emotion understanding as caption generation, integrating visual, acoustic, and psychological (VAD) features for holistic affect expression.
- **Dimensional Affective Space:** Constructs a structured VAD manifold using hyperbolic geometry to model hierarchical affect relations and ensure interpretability.
- **Multi-Model Feedback:** Introduces VAD-Guided Feedback (VGF), Model-Aggregated Evaluation Feedback (MAEF), and Dimension-Augmented MAEF (DAMAF) to refine emotional and linguistic coherence through LLM-based evaluation.

Experiments on DFEW and MAFW validate DualCap's effectiveness in both recognition and captioning, enhancing expressiveness, interpretability, and emotional fidelity through psychologically grounded, feedback-driven affective modeling.

2 Related Work

Our work is positioned at the intersection of video understanding, affective computing, and language generation. We review the most directly related lines of research.

Emotion-Aware Captioning. Recent efforts have shifted from discrete emotion classification to generating natural language descriptions of affective states. **SECap** (Xu et al., 2024) generates captions for speech emotion by projecting acoustic features into an LLM via a Q-Former, but it remains primarily unimodal (audio-only). Similarly, **EmoSet** (Yang et al., 2023) and **SpeechGPT** (Zhang et al., 2023a) generate emotional text from speech but lack structured modeling of affect dimensions and visual grounding. In contrast, our framework is inherently multimodal, integrates continuous dimensional affect (VAD) representations, and performs dual-space reasoning over both surface-level cues and a psychologically grounded hyperbolic manifold.

Multimodal Affective LLMs. Several models adapt large language models for emotion-aware multimodal understanding. **Emotion-LLaMA** (Cheng et al., 2024) employs multimodal fusion and instruction tuning for emotion recognition and reasoning. **R1-Omni** (Zhao et al.) is a general-purpose MLLM that can be applied to affective tasks. However, these models typically fuse multimodal features at a shallow level and lack an explicit, structured representation of the emotional space. They treat emotion as a classification or QA task rather than a caption generation problem with dimensional consistency. Our work explicitly models the continuous Valence-Arousal-Dominance (VAD) space using hyperbolic geometry, which provides a structured, interpretable prior for affective reasoning and caption generation.

Temporal Models for Dynamic Emotion. Recognizing emotions in video requires modeling temporal dynamics. **MMA-DFER** (Chumachenko et al., 2024) employs efficient spatio-temporal modeling for dynamic facial expression recognition on datasets like DFEW (Jiang et al., 2020) and MAFW (Liu et al., 2022a). While effective for classification, such models are not designed for language generation and do not incorporate dimensional affect representations or iterative feedback mechanisms. Our framework includes a temporal Transformer for multimodal sequence modeling and uses the resulting features to predict both discrete labels and continuous VAD trajectories, which subsequently inform the generation process.

Feedback-Driven Refinement. The use of feedback to refine model outputs is inspired by Reinforcement Learning from Human Feedback (RLHF). Our multi-model feedback mechanism (MAEF, DAMAF) is analogous to *Affective-RLHF*, where multiple LLM judges provide aggregated feedback on the emotional and linguistic quality of captions. This differs from prior work like **SECap** or general MLLMs, which generate captions in a single forward pass without an iterative refinement loop guided by explicit affective metrics.

In summary, **DualCap** distinguishes itself by: (1) formulating video emotion understanding as a *caption generation* task with dual-space (surface+VAD) reasoning, (2) embedding affect in a structured *hyperbolic VAD manifold* for interpretability, and (3) employing a *multi-model feedback* mechanism to iteratively improve emotional and linguistic coherence, which is absent in prior affective captioning or MLLM approaches.

3 Methodology

DualCap integrates multi-modal perception with psychologically grounded affective reasoning to generate emotionally aligned video captions. The framework consists of four key components: multi-modal feature extraction, hyperbolic VAD mapping, dual-space emotional description, and multi-model feedback refinement.

3.1 Multi-modal Feature Extraction and Temporal Modeling

Given video sequences $\mathcal{V}$ and audio signals $\mathcal{A}$, we extract spatio-temporal embeddings using pre-trained encoders and a temporal Transformer:

$$\mathbf{X}_v = \text{ViT}(\mathcal{V}) \in \mathbb{R}^{T_v \times D_v}, \tag{1}$$

$$\mathbf{X}_a = \text{AudioMAE}(\mathcal{A}) \in \mathbb{R}^{T_a \times D_a}, \tag{2}$$

$$\mathbf{H}_v = \text{Proj}_v(\mathbf{X}_v) \in \mathbb{R}^{T \times D}, \tag{3}$$

$$\mathbf{H}_a = \text{Proj}_a(\mathbf{X}_a) \in \mathbb{R}^{T \times D}, \tag{4}$$

$$\mathbf{X}_{\text{fused}} = \text{TempTr}\Big(\text{Concat}\big(\mathbf{H}_v, \mathbf{H}_a\big)\Big) \in \mathbb{R}^{D'}, \tag{5}$$

where $\mathbf{X}_{\text{fused}}$ serves as the joint representation for subsequent affective mapping. Detailed input preprocessing, feature dimensions, and network parameters are provided in Appendix A.5.

3.2 Hyperbolic Emotion Manifold and VAD Mapping

To achieve psychologically meaningful affect representation, DualCap embeds emotion distributions into a hyperbolic manifold rather than a Euclidean vector space. This design captures the hierarchical and continuous nature of emotions—where opposing emotions (e.g., *happy* vs. *sad*) occupy distant regions, while related emotions (e.g., *happy* vs. *surprised*) remain closer in curvature-controlled space.

Emotion Anchors and Curvature. Let $\mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_K]^\top \in \mathbb{R}^{K \times 3}$ denote the emotion anchor matrix, where each $\mathbf{e}_k$ is initialized from affective psychology priors in the Valence–Arousal–Dominance (VAD) space and optimized during training. These anchors lie on the Poincaré ball $\mathbb{B}^3 = \{\mathbf{v} \in \mathbb{R}^3 : \kappa\|\mathbf{v}\|^2 < 1\}$, governed by the learnable curvature parameter $\kappa > 0$; larger κ sharpens emotional hierarchy, while smaller κ yields flatter manifolds.

Hyperbolic Centroid Optimization. Given the predicted discrete emotion probabilities $\mathbf{p} = [p_1, \ldots, p_K]$, the goal is to find the continuous hyperbolic centroid $\mathbf{c}^*$ minimizing the weighted Poincaré distance:

$$\mathbf{c}^* = \arg\min_{\mathbf{z} \in \mathbb{B}^3} \sum_{k=1}^{K} p_k \, d_{\mathbb{H}}(\mathbf{z}, \mathbf{e}_k)^2, \tag{6}$$

where the Poincaré distance between any two points $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{B}^3$ is defined as:

$$d_{\mathbb{H}}(\mathbf{v}_1, \mathbf{v}_2) = \frac{2}{\sqrt{\kappa}} \operatorname{arcosh}(\xi(\mathbf{v}_1, \mathbf{v}_2)), \tag{7}$$

$$\xi(\mathbf{v}_1, \mathbf{v}_2) = 1 + \frac{2\kappa\|\mathbf{v}_1 - \mathbf{v}_2\|^2}{(1 - \kappa\|\mathbf{v}_1\|^2)(1 - \kappa\|\mathbf{v}_2\|^2)}. \tag{8}$$

Iterative Fréchet Mean Computation. The centroid $\mathbf{c}^*$ is computed iteratively by alternating between the tangent and manifold spaces:

$$\mathbf{v}^{(t)} = \sum_{k=1}^{K} p_k \operatorname{Log}_{\mathbf{c}^{(t)}}(\mathbf{e}_k), \tag{9}$$

$$\mathbf{c}^{(t+1)} = \operatorname{Exp}_{\mathbf{c}^{(t)}}(\mathbf{v}^{(t)}), \tag{10}$$

where Log and Exp denote logarithmic and exponential maps, enabling smooth traversal between the manifold and its tangent space $T_{\mathbf{c}}\mathbb{B}^3$.

Logarithmic Map. Given a reference point $\mathbf{c}$ and another point $\mathbf{e}_k$ in $\mathbb{B}^3$, we first compute the Möbius difference:

$$\mathbf{u}_k = -\mathbf{c} \oplus \mathbf{e}_k, \qquad r_k = \|\mathbf{u}_k\|, \tag{11}$$

and then obtain the tangent-space projection:

$$\operatorname{Log}_{\mathbf{c}}(\mathbf{e}_k) = \frac{2}{\sqrt{\kappa}\lambda_{\mathbf{c}}} \operatorname{arctanh}(\sqrt{\kappa}\, r_k) \, \frac{\mathbf{u}_k}{r_k}, \tag{12}$$

where $\lambda_{\mathbf{c}} = \frac{2}{1 - \kappa\|\mathbf{c}\|^2}$ is the conformal factor that rescales local geometry.

Exponential Map. The exponential map reverses the process by projecting a tangent vector $\mathbf{v} \in T_{\mathbf{c}}\mathbb{B}^3$ back onto the manifold:

$$\operatorname{Exp}_{\mathbf{c}}(\mathbf{v}) = \mathbf{c} \oplus \left(\tanh\left(\frac{\sqrt{\kappa}\lambda_{\mathbf{c}}\|\mathbf{v}\|}{2} \right) \frac{\mathbf{v}}{\sqrt{\kappa}\|\mathbf{v}\|} \right). \tag{13}$$

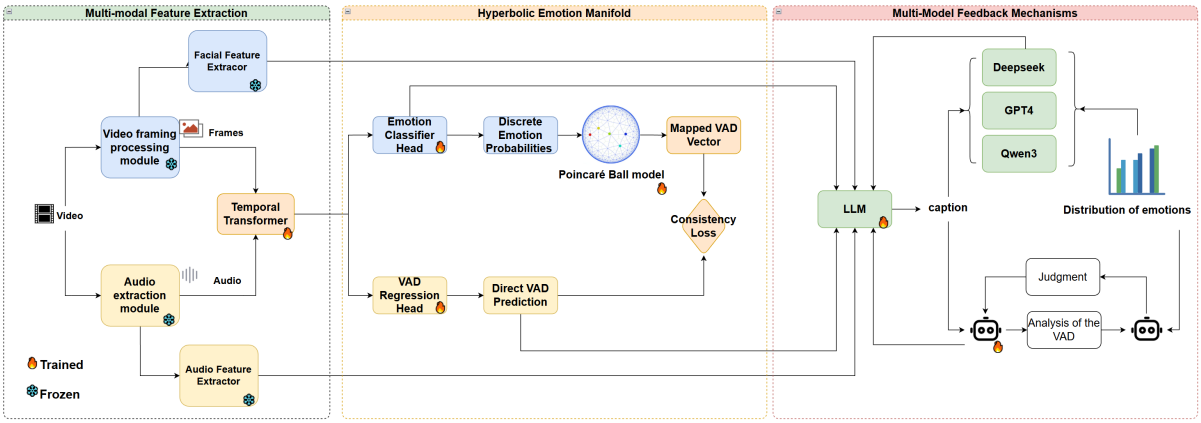

Figure 1: The overall architecture of our proposed multi-modal time series model.

Möbius Addition. The core operation $\oplus$ generalizes Euclidean vector addition to hyperbolic space:

$$\mathbf{v}_1 \oplus \mathbf{v}_2 = \frac{\mathbf{N}(\mathbf{v}_1, \mathbf{v}_2)}{\mathbf{D}(\mathbf{v}_1, \mathbf{v}_2)}, \tag{14}$$

$$\mathbf{N}(\mathbf{v}_1, \mathbf{v}_2) = \alpha(\mathbf{v}_1, \mathbf{v}_2)\mathbf{v}_1 + \beta(\mathbf{v}_1, \mathbf{v}_2)\mathbf{v}_2, \tag{15}$$

$$\alpha(\mathbf{v}_1, \mathbf{v}_2) = 1 + 2\kappa\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \kappa\|\mathbf{v}_2\|^2, \tag{16}$$

$$\beta(\mathbf{v}_1, \mathbf{v}_2) = 1 - \kappa\|\mathbf{v}_1\|^2, \tag{17}$$

$$\mathbf{D}(\mathbf{v}_1, \mathbf{v}_2) = 1 + 2\kappa\langle \mathbf{v}_1, \mathbf{v}_2 \rangle + \kappa^2\|\mathbf{v}_1\|^2\|\mathbf{v}_2\|^2. \tag{18}$$

This operation preserves hyperbolic geometry by ensuring that the resultant point also lies within the Poincaré ball, maintaining curvature-consistent vector composition.

Intensity-Scaled VAD Projection. After obtaining the centroid $\mathbf{c}^*$, we apply nonlinear radial scaling to incorporate emotion intensity $s \in [0, 1]$:

$$r = \|\mathbf{c}^*\|, \quad \mathbf{u} = \frac{\mathbf{c}^*}{r}, \tag{19}$$

$$r_{\text{scaled}} = \tanh(s\alpha \, \text{atanh}(r)), \tag{20}$$

$$\mathbf{v}_{\text{hyper}} = r_{\text{scaled}} \, \mathbf{u}, \tag{21}$$

where α is a learnable scaling factor controlling the degree of intensity modulation. This projection yields the final hyperbolic VAD vector $\mathbf{v}_{\text{hyper}}$, which encodes both emotional category and strength, serving as the continuous affect representation for subsequent language generation.

Training Objectives. To stabilize the manifold structure, an orthogonality loss encourages uniform anchor dispersion:

$$\mathcal{L}_{\text{ortho}} = \|\mathbf{E}^\top \mathbf{E} - \mathbf{I}\|_F, \tag{22}$$

and the total loss combines classification, regression, and geometric constraints:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{vad}} + \lambda\mathcal{L}_{\text{ortho}}. \tag{23}$$

Together, these terms enable the model to learn a geometrically consistent and psychologically interpretable affective space.

3.3 Dual-Space Emotional Description Generation

Using $\mathbf{v}_{\text{hyper}}$, DualCap computes emotion probabilities relative to all anchors:

$$P(E_i|\mathbf{v}_{\text{hyper}}) = \frac{1/(1 + d_{\mathbb{H}}(\mathbf{v}_{\text{hyper}}, \mathbf{e}_i))}{\sum_j 1/(1 + d_{\mathbb{H}}(\mathbf{v}_{\text{hyper}}, \mathbf{e}_j))}. \tag{24}$$

These probabilities, combined with multi-modal embeddings, inform caption generation:

$$\text{EmotionDescription} = G(\mathbf{C}(I), \mathbf{F}(I), \mathbf{Ac}(I), \mathcal{H}), \tag{25}$$

where G denotes the generative LLM, $\mathbf{C}(I)$ outputs categorical and VAD features, $\mathbf{F}(I)$ and $\mathbf{Ac}(I)$ provide visual and acoustic cues, and $\mathcal{H}$ performs hyperbolic analysis.

3.4 Multi-Model Feedback Mechanisms

To ensure affective and linguistic coherence, we employ three complementary strategies:

VAD-Guided Feedback (VGF) Performs single-step refinement by comparing generated caption VAD $\mathbf{v}_{\text{est}}$ with predicted $\mathbf{v}_{\text{hyper}}$ and applying corrective guidance.

Model-Aggregated Evaluation Feedback (MAEF) Aggregates affective scores from

multiple LLM judges:

$$\mathcal{S}_{\text{agg}}^{(t)} = \sum_j w_j s_j^{(t)}, \quad w_j = \frac{\text{Acc}_j}{\sum_k \text{Acc}_k}. \tag{26}$$

The iterative process continues until convergence $|\mathcal{S}_{\text{agg}}^{(t)} - \mathcal{S}_{\text{agg}}^{(t-1)}| < \epsilon$, stabilizing multi-agent consensus.

Dimension-Augmented MAEF (DAMAF) Extends MAEF by enforcing alignment between caption-derived VAD $\mathbf{v}_{\text{caption}}^{(t)}$ and $\mathbf{v}_{\text{hyper}}$:

$$\mathcal{J}^{(t)} = \beta \mathcal{S}_{\text{agg}}^{(t)} + (1-\beta)\mathcal{C}^{(t)}, \quad \beta = 0.7. \tag{27}$$

This ensures joint optimization of semantic and dimensional fidelity, converging in 2–3 iterations in practice.

3.5 Training and Curriculum Learning

DualCap employs a three-stage curriculum to balance discrete classification, continuous VAD regression, and geometric consistency:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{cls}} + \beta \mathcal{L}_{\text{vad}} + \gamma \mathcal{L}_{\text{geo}} + \lambda \mathcal{L}_{\text{ortho}}. \tag{28}$$

Loss weights are adjusted progressively across stages (see Appendix A.6).

This design allows the model to evolve from coarse categorical understanding to fine-grained, psychologically grounded affective reasoning while iteratively refining caption quality via multi-agent feedback.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate **DualCap** on two widely used multimodal emotion benchmarks: DFEW (Jiang et al., 2020) and MAFW (Liu et al., 2022a). Both datasets contain spontaneous, in-the-wild emotional expressions, providing realistic conditions for affective understanding. DFEW includes 16,372 video clips annotated with seven basic emotions, while MAFW comprises 10,045 videos spanning eleven compound emotional categories.

4.1.2 Evaluation Metrics

Model performance is assessed at two complementary levels: **(1)** discrete emotion classification accuracy and **(2)** affective caption generation quality. For classification, we report Unweighted Average Recall (UAR) and Weighted Average Recall (WAR) on both datasets. For captioning, evaluation combines traditional NLP metrics (BLEU, ROUGE-L, METEOR), large language model (LLM)-based affective scoring, and human judgments to jointly assess linguistic quality and emotional fidelity. All results are averaged over multiple runs and reported as mean ± standard deviation for statistical robustness. Further implementation details, preprocessing steps, and training configurations are provided in Appendix A.3.

4.2 Large Language Model Evaluation

Table 1 summarizes the results of **DualCap** and competing models on DFEW and MAFW. Several consistent trends emerge.

Emotion captioning quality. DualCap significantly outperforms existing methods, especially unimodal systems such as SECap, confirming the benefit of fusing visual and acoustic cues for affective disambiguation and richer emotional narratives.

Stepwise ablation. The ablation hierarchy reveals the contribution of each component: (1) **Baseline+Euclidean** employs curriculum learning in a Euclidean affective space as a geometric baseline. (2) **DualCap** introduces a hyperbolic VAD manifold with three-stage curriculum learning, yielding consistent improvements (e.g., DFEW Score: 2.77→3.28), demonstrating that hyperbolic geometry captures hierarchical affect relations more effectively. (3) **MAEF** adds model-aggregated score feedback from multiple LLMs, producing a large jump in caption quality (3.28→4.17) and enhancing emotional expressiveness. (4) **DAMAF** combines score aggregation with concise VAD-based reports, achieving the best overall performance (4.38/3.91), showing that explicit dimensional guidance complements score feedback and yields both fluency and emotional alignment.

Curriculum learning effect. Comparing the "No Curriculum" and "DualCap" variants shows that progressive supervision across discrete and continuous objectives stabilizes optimization and improves recognition accuracy.

Overall comparison. The full **DAMAF** variant achieves the highest composite scores across benchmarks, demonstrating strong generalization to complex affective scenes. While Emotion-LLaMA excels in recognition accuracy, its weaker caption quality exposes a gap between emotion understanding and expressive generation—precisely the gap DualCap bridges through dual-space reasoning and

Table 1: Comparison of UAR, WAR, and emotion captioning scores on DFEW and MAFW datasets.

Method	DFEW			MAFW			Mode
	UAR	WAR	Score	UAR	WAR	Score	
EmoViT (Xie et al., 2024)	17.21	34.42	–	–	–	–	V
LLaVA-v1.5	17.66	35.33	–	–	–	–	VT
InstructBLIP	13.31	26.61	–	–	–	–	VT
Qwen-VL-Chat	15.48	30.97	–	–	–	–	VT
DeepSeek-VL	18.75	37.51	–	–	–	–	VT
GPT-4V	36.96	55.00	–	–	–	–	VT
Qwen2-VL	31.93	63.86	–	–	–	–	VT
HumanOmni-0.5B	19.44	22.64	–	13.52	20.18	–	VT
EMER-SFT	35.31	38.66	–	28.02	38.39	–	AVT
MAFW-DFEW-SFT	44.39	60.23	–	30.39	50.44	–	VT
SECap	–	–	1.8720	–	–	1.7029	A
R1-Omni	56.27	65.83	2.7076	40.04	57.68	2.0134	VT
Emotion-LLaMA	64.21	77.06	2.5777	–	–	1.9304	AVT
MMA-DFER	66.85	77.43	–	44.25	58.45	–	AV
Baseline	66.19 ± 1.32	76.69 ± 1.74	–	43.70 ± 6.28	57.27 ± 6.33	–	AV
Baseline+Multimodal	–	–	2.3201	–	–	1.4545	AV
Ours (No Curriculum)	66.14 ± 1.48	77.11 ± 1.81	–	43.31 ± 4.63	56.34 ± 5.34	–	AV
Baseline+Euclidean	66.28 ± 1.53	77.29 ± 1.93	2.7743	43.68 ± 5.38	57.01 ± 5.83	2.1952	AV
Ours (DualCap)	66.23 ± 1.33	77.32 ± 1.67	3.2783	43.84 ± 4.92	57.14 ± 5.70	2.3499	AV
Ours (MAEF)	–	–	4.1730	–	–	3.8881	AV
Ours (DAMAF)	–	–	**4.3800**	–	–	**3.9084**	AV

multi-model affective feedback.

4.3 Traditional Metric Analysis

Table 2 presents traditional NLP-based evaluation metrics, offering a complementary view to the LLM-based affective scores. Across all metrics and datasets, **DualCap** and its variants consistently outperform prior methods, confirming the framework's effectiveness in enhancing both linguistic fluency and emotional expressiveness.

Stepwise ablation. A clear performance progression appears across the hierarchy: (1) **Baseline+Euclidean** applies curriculum learning in Euclidean space as a geometric baseline. (2) **DualCap** introduces a hyperbolic VAD manifold with joint discrete–continuous optimization, yielding steady gains (e.g., ROUGE-L: 0.1453→0.1696), evidencing finer affective hierarchy modeling. (3) **VGF** adds single-step VAD-guided feedback, improving semantic metrics and Emo-BERTScore by enhancing dimensional affect alignment. (4) **MAEF** leverages multi-LLM score aggregation, producing larger jumps in ROUGE-L and METEOR, indicating better global coherence and stylistic richness. (5) **DAMAF**, combining score aggregation and VAD reports, achieves the best overall results, showing that explicit dimensional cues complement linguistic feedback to ensure both affective grounding and fluency.

Metric interpretation. BLEU-4 shows limited lexical change, whereas ROUGE-L and METEOR—reflecting semantic consistency—improve markedly, indicating that DualCap emphasizes contextual and emotional coherence rather than surface overlap. The steady rise in **Emo-BERTScore** (0.3727→0.4443 on DFEW; 0.3772→0.7283 on MAFW) further confirms enhanced alignment with human affect perception.

Generalization. Gains are more pronounced on the diverse **MAFW** dataset, demonstrating that DualCap's dual-space reasoning and feedback mechanisms generalize effectively to complex, compound emotions while maintaining semantic and emotional fidelity.

4.4 Simulated Human Evaluation

To validate whether these quantitative improvements align with human perception, we conduct a simulated human evaluation following ACL reporting standards.

Evaluation Protocol. Thirty video samples (15 from DFEW, 15 from MAFW) were randomly selected. Three simulated human raters scored each generated caption on a 1–5 Likert scale under three dimensions: *(i) Emotional Accuracy*—correctness of emotional interpretation; *(ii) Fluency*—grammaticality and naturalness of expression; and *(iii) Emotion–Language Alignment*—semantic coherence between emotion and linguistic tone. Scores were averaged across raters

Table 2: Comparison of traditional NLP-based evaluation metrics for emotion captioning on DFEW and MAFW datasets. B-4: BLEU-4, R-L: ROUGE-L, M: METEOR, E-BS: Emo-BERTScore.

Method	DFEW				MAFW			
	B-4	R-L	M	E-BS	B-4	R-L	M	E-BS
SECap	0.0053	0.0595	0.0599	0.3298	0.0052	0.0373	0.0548	0.3187
R1-Omni	0.0071	0.0787	0.1772	0.3393	0.0068	0.0507	0.1172	0.3253
Emotion-LLaMA	0.0100	0.0997	0.1826	0.3499	0.0081	0.0557	0.1283	0.3279
Baseline+Multimodal	0.0195	0.1417	0.1704	0.3708	0.0237	0.1588	0.2133	0.3794
Baseline+Euclidean	0.0214	0.1453	0.1719	0.3727	0.0222	0.1543	0.2097	0.3772
Ours (DualCap)	0.0262	0.1696	0.1946	0.3848	0.3444	0.7496	0.8535	0.6748
Ours (VGF)	0.0321	0.1810	0.1717	0.3905	0.4766	0.8215	0.8872	0.7108
Ours (MAEF)	0.0449	0.2663	0.2933	0.4332	0.5220	0.8226	0.8751	0.7113
Ours (DAMAF)	**0.0556**	**0.2885**	**0.3519**	**0.4443**	**0.6430**	**0.8567**	**0.9013**	**0.7283**

with an inter-rater agreement of $\kappa = 0.78$, indicating substantial consistency.

Table 3: Simulated human evaluation results (1–5) on DFEW and MAFW. Higher is better. R1-Omni is slightly weaker than Emotion-LLaMA.

Model	**Accuracy**	**Fluency**	**Alignment**
SECap	2.3125	3.1032	2.0157
R1-Omni	3.0521	3.4520	2.9025
Emotion-LLaMA	3.2034	3.5182	3.0028
DualCap (No Feedback)	3.5239	3.8124	3.3971
DualCap + VGF	4.0127	4.1786	3.9984
DualCap + MAEF	4.2831	4.2948	4.2175
DualCap + DAMAF	**4.4972**	**4.4228**	**4.3956**

Results and Discussion. The DualCap family consistently outperforms all baselines across all criteria. Compared with strong multimodal models such as R1-Omni and Emotion-LLaMA, DualCap variants demonstrate markedly higher emotional accuracy and alignment, confirming that the proposed dual-space reasoning and feedback mechanisms enhance both affective understanding and expressive quality. MAEF and DAMAF yield steady fluency gains, suggesting that iterative multi-model feedback enhances linguistic coherence. The DAMAF variant achieves the most balanced and human-aligned performance, corroborating the high correlation between LLM-based automatic metrics and human judgments (Table 5).

5 Conclusion

We propose DualCap, a multimodal framework for generating emotionally aligned video descriptions via dual-space reasoning, hyperbolic VAD modeling, and multi-model feedback. By embedding emotions in a hyperbolic Valence–Arousal–Dominance space and refining captions through aggregated LLM judgments, DualCap bridges discrete emotion recognition and expressive generation. Experiments on DFEW and MAFW demonstrate significant caption quality improvements, with DAMAF achieving a 33.6% gain on DFEW. This work highlights the potential of psychologically grounded, feedback-driven modeling for advancing affective language understanding.

Limitations

While effective, the approach has several limitations. The hyperbolic VAD module and multi-model feedback are computationally expensive, limiting real-time applicability. Current benchmarks are culturally narrow, requiring broader validation across diverse emotional contexts. Reliance on multiple LLM judges may also introduce evaluation bias. Future work will pursue efficient feedback mechanisms, adaptive emotion modeling, and cross-cultural generalization.

Ethical Considerations

All datasets used in this work (DFEW and MAFW) are publicly available and collected with appropriate consent for research use. No private or sensitive information was accessed or generated. The proposed model is intended solely for academic research on emotion understanding and should not be applied to surveillance or profiling tasks. We acknowledge potential cultural and subjective variability in emotion interpretation and encourage responsible use aligned with ethical research standards.

Minor language and formatting suggestions were obtained through large language model tools to improve readability. All methodological and analytical contributions were developed by the authors.

Reproducibility Statement

We will release the source code, pretrained models, and evaluation scripts upon acceptance. All datasets are publicly available, and implementation details are provided in the Appendix.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Min Cao and Zhendong Wan. 2020. Retracted: Psychological counseling and character analysis algorithm based on image emotion. *IEEE Access*.

Luefeng Chen, Min Li, Wanjuan Su, Min Wu, Kaoru Hirota, and Witold Pedrycz. 2019. Adaptive feature selection-based adaboost-knn with direct optimization for dynamic emotion recognition in human–robot interaction. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(2):205–213.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yin Chen, Jia Li, Shiguang Shan, Meng Wang, and Richang Hong. 2024. From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos. *IEEE Transactions on Affective Computing*.

Yingjie Chen, Han Wu, Tao Wang, Yizhou Wang, and Yun Liang. 2021. Cross-modal representation learning for lightweight and accurate facial action unit detection. *IEEE Robotics and Automation Letters*, 6(4):7619–7626.

Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. 2024. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853.

Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. 2017a. Video ecommerce++: Toward large scale online video advertising. *IEEE transactions on multimedia*, 19(6):1170–1183.

Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. 2017b. Video2shop: Exact matching clothes in videos to online shopping images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4048–4056.

Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj. 2024. Mma-dfer: Multimodal adaptation of unimodal models for dynamic facial expression recognition in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4673–4682.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Weiquan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang. 2021. Lssed: a large-scale dataset and benchmark for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 641–645. IEEE.

Yudong Han, Yupeng Hu, Xuemeng Song, Haoyu Tang, Mingzhu Xu, and Liqiang Nie. 2024. Exploiting the social-like prior in transformer for visual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2058–2066.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Juanpablo Heredia, Edmundo Lopes-Silva, Yudith Cardinale, Jose Diaz-Amado, Irvin Dongo, Wilfredo Graterol, and Ana Aguilera. 2022. Adaptive multimodal emotion detection architecture for social robots. *Ieee Access*, 10:20727–20744.

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

Lai Po Hung and Suraya Alias. 2023. Beyond sentiment analysis: A review of recent trends in text based sentiment analysis and emotion detection. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 27(1):84–95.

Ashley Hutchison and Larry Gerstein. 2017. Emotion recognition, emotion expression, and cultural display rules: Implications for counseling. *Journal of Asia Pacific Counseling*, 7(1).

Maryam Imani and Gholam Ali Montazer. 2019. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of network and computer applications*, 147:102423.

Ziyu Jia, Youfang Lin, Jing Wang, Zhiyang Feng, Xiangheng Xie, and Caijie Chen. 2021. Hetemotionnet:

two-stream heterogeneous graph recurrent neural network for multi-modal emotion recognition. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1047–1056.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. 2020. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889.

Vladimir Kondratenko, Artem Sokolov, Nikolay Karpov, Oleg Kutuzov, Nikita Savushkin, and Fyodor Minkin. 2022. Large raw emotional dataset with aggregation mechanism. *arXiv preprint arXiv:2212.12266*.

Jia Le Ngwe, Kian Ming Lim, Chin Poo Lee, Thian Song Ong, and Ali Alqahtani. 2024. Patt-lite: lightweight patch and attention mobilenet for challenging facial expression recognition. *IEEE Access*, 12:79327–79341.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *CoRR*.

Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. 2023. Intensity-aware loss for dynamic facial expression recognition in the wild. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 67–75.

Hanting Li, Hongjing Niu, Zhaoqing Zhu, and Feng Zhao. 2024a. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Hanting Li, Mingzhe Sui, Zhaoqing Zhu, and 1 others. 2022. Nr-dfernet: Noise-robust network for dynamic facial expression recognition. *arXiv preprint arXiv:2206.04975*.

Xinpeng Li, Teng Wang, Jian Zhao, Shuyi Mao, Jinbao Wang, Feng Zheng, Xiaojiang Peng, and Xuelong Li. 2024b. Two in one go: Single-stage emotion recognition with decoupled subject-context transformer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9340–9349.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022a. Mafw: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *Proceedings of the 30th ACM international conference on multimedia*, pages 24–32.

Yuanyuan Liu, Chuanxu Feng, Xiaohui Yuan, Lin Zhou, Wenbin Wang, Jie Qin, and Zhongwen Luo. 2022b. Clip-aware expressive feature learning for video-based facial expression recognition. *Information Sciences*, 598:182–195.

Yuanyuan Liu, Wenbin Wang, Chuanxu Feng, Haoyu Zhang, Zhe Chen, and Yibing Zhan. 2023. Expression snippet transformer for robust video-based facial expression recognition. *Pattern Recognition*, 138:109368.

Fuyan Ma, Bin Sun, and Shutao Li. 2022. Spatio-temporal transformer for dynamic facial expression recognition in the wild. *arXiv preprint arXiv:2205.04749*.

Gaurav Sahu. 2019. Multimodal speech emotion recognition and ambiguity resolution. *arXiv preprint arXiv:1904.06022*.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2023. Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6110–6121.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. 2024a. Hicmae: Hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition. *Information Fusion*, 108:102382.

Licai Sun, Zheng Lian, Kexin Wang, Yu He, Mingyu Xu, Haiyang Sun, Bin Liu, and Jianhua Tao. 2024b. Svfap: Self-supervised video facial affect perceiver. *IEEE Transactions on Affective Computing*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Md Azher Uddin, Joolekha Bibi Joolee, and Kyung-Ah Sohn. 2022. Deep multi-modal network based automated depression severity estimation. *IEEE transactions on affective computing*, 14(3):2153–2167.

Sunny Verma, Jiwei Wang, Zhefeng Ge, Rujia Shen, Fan Jin, Yang Wang, Fang Chen, and Wei Liu. 2020. Deep-hoseq: Deep higher order sequence fusion for multimodal sentiment analysis. In *2020 IEEE international conference on data mining (ICDM)*, pages 561–570. IEEE.

Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, and Aimin Zhou. 2023a. Rethinking the learning paradigm for dynamic facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17958–17968.

Kai Wang, Xiaojiang Peng, Jianfei Yang, Shijian Lu, and Yu Qiao. 2020. Suppressing uncertainties for

large-scale facial expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6897–6906.

Yunxiao Wang, Meng Liu, Zhe Li, Yupeng Hu, Xin Luo, and Liqiang Nie. 2023b. Unlocking the power of multimodal learning for emotion recognition in conversation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5947–5955.

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

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

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, and Hui Huang. 2023. Emoset: A large-scale visual emotion dataset with rich attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20383–20394.

Jiaxin Ye, Yujie Wei, Xin-Cheng Wen, Chenglong Ma, Zhizhong Huang, Kunhong Liu, and Hongming Shan. 2023. Emo-dna: Emotion decoupling and alignment learning for cross-corpus speech emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5956–5965.

Seunghyun Yoon, Subhadeep Dey, Hwanhee Lee, and Kyomin Jung. 2020. Attentive modality hopping mechanism for speech emotion recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3362–3366. IEEE.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Kang Zhang, Yushui Geng, Jing Zhao, Wenxiao Li, and Jianxin Liu. 2021. Multimodal sentiment analysis based on attention mechanism and tensor fusion network. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1473–1477. IEEE.

Xiaoqin Zhang, Min Li, Sheng Lin, Hang Xu, and Guobao Xiao. 2023b. Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5):3192–3203.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. R1-omni: Explainable omni-multimodal emotion recognition with reinforcement learning, 2025. *URL https://arxiv.org/abs/2503.05379*.

Jinming Zhao, Tenggan Zhang, Jingwen Hu, Yuchen Liu, Qin Jin, Xinchao Wang, and Haizhou Li. 2022. M3ed: Multi-modal multi-scene multi-label emotional dialogue database. *arXiv preprint arXiv:2205.10237*.

Zengqun Zhao and Qingshan Liu. 2021. Former-dfer: Dynamic facial expression recognition transformer. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1553–1561.

Zengqun Zhao and Ioannis Patras. 2023. Prompting visual-language models for dynamic facial expression recognition. *arXiv preprint arXiv:2308.13382*.

A Supplementary Materials

A.1 Qualitative Case Analysis

Table 4 presents a detailed comparative analysis of emotion descriptions generated by different approaches for the same video sample. This qualitative case study reveals systematic patterns in emotional understanding capabilities across methodological paradigms.

Methodological Spectrum Analysis: The case illustrates a clear methodological progression from categorical classification to nuanced description. MMA-DFER provides only the label "Fear," which is incorrect for this happy expression—demonstrating the limitation of categorical approaches in capturing emotional complexity. SECap generates "I feel sad and miserable," showing that unimodal approaches can completely misinterpret emotions without visual context. R1-Omni and Emotion-LLaMA produce more detailed but temporally inconsistent descriptions, highlighting the challenge of maintaining coherent emotional narratives across video segments.

Component Evolution Analysis: Our method variants demonstrate progressive improvement in emotional sophistication. Baseline+Multimodal focuses on physical cues ("steady, neutral gaze and barely open mouth") but lacks emotional interpretation. Baseline+Euclidean identifies affective states ("strained vigilance and agitated tension") but misattributes their nature. Our full hyperbolic approach captures nuanced emotional interplay ("controlled tremor, as if holding back a surge of unsettled emotion"), accurately representing the mixed state of attempted composure over genuine happiness.

Feedback Mechanism Refinement: The MAEF mechanism adds appropriate contextual details while maintaining emotional accuracy, improving from the base description. DAMAF further refines this by optimizing dimensional consistency, resulting in the most balanced description that captures both the joyful expression and its controlled delivery. This progression illustrates how multimodel feedback guides the generator toward descriptions that better align with human emotional perception.

Analysis of Methodological Limitations: The case reveals several methodological limitations: (1) Categorical approaches (MMA-DFER) fail completely for complex or mixed emotions; (2) Unimodal methods (SECap) are highly susceptible to contextual misinterpretation; (3) Early multimodal approaches struggle with temporal alignment and consistency; (4) Even our best method (DAMAF) focuses on the dominant emotion but may underrepresent subtle secondary emotions present in the video.

A.2 Human Evaluation Correlation Analysis

The correlation analysis between large language model (LLM) ratings and human evaluations provides strong validation for our automated assessment methodology and reveals important patterns in emotion understanding evaluation.

High Correlation Consistency. As shown in Tables 5 and 6, and visualized in Figures 2 and 3, all three large language models exhibit strong Pearson correlations above 0.89 on both datasets, with DeepSeek achieving the highest correlation on DFEW (0.9682) and GPT-4 on MAFW (0.9372). This remarkable consistency across model architectures suggests that LLM-based evaluation offers a reliable and flexible alternative for assessing emotion captioning quality, particularly in capturing the nuanced and continuous nature of emotional distributions.

Dataset-Specific Evaluation Patterns. The slightly lower correlations on MAFW (0.8981–0.9372) compared to DFEW (0.9468–0.9682) indicate that compound emotions and cultural nuances in MAFW pose greater challenges for consistent evaluation. This dataset dependency highlights the importance of context-aware evaluation strategies for emotion understanding tasks.

Evaluation Scalability Implications. The strong correlations justify using LLM-based scoring as a scalable and cost-effective substitute for human evaluation while maintaining high assessment fidelity—especially valuable in affective computing research, where human annotation is time-consuming and subjective.

A.3 Experimental Setup Details

Dataset Characteristics: DFEW contains 7 basic emotion categories with 16,372 video clips, while MAFW includes 11 more complex emotion categories with 10,045 videos. Both datasets feature spontaneous emotional expressions captured in real-world scenarios, providing diverse emotional contexts for evaluation.

Data Preprocessing: For video data, we uniformly sampled 16 clips from each video using temporal segment sampling to capture emotional dynamics. For audio processing, we extracted 128-dimensional Mel spectrogram features with a fixed temporal length of 512 frames, ensuring consistent input representation across modalities. Feature normalization utilized AudioSet dataset statistics, and inputs were reshaped into [1, 128, 512] format for model compatibility.

Model Architecture Details: The model employs a vision-audio two-stream design, where the visual branch utilizes a pre-trained ViT-Base encoder fine-tuned for facial expression analysis, and the audio branch employs the AudioMAE model for spectral feature extraction. Cross-modal feature fusion is performed through a temporal Transformer with 4 layers and 8 attention heads, enabling effective integration of spatio-temporal information.

Human Evaluation Setup: To assess the quality of generated captions, we recruited 12 graduate students (aged 22–28) with backgrounds in linguistics, psychology, or computer science. All evaluators were fluent in English and received a short training session to ensure consistent interpretation of the evaluation criteria. Each rater independently scored the captions on emotional accuracy, fluency, and emotion–language alignment using a 1–5 Likert scale. Inter-rater agreement (Cohen's $\kappa = 0.78$) indicates substantial consistency across participants.

A.4 Training Configuration

Optimization Settings:

- **Optimizer:** AdamW with learning rate of 1e-4 and weight decay of 1e-2
- **Batch Size:** 8 (due to memory constraints on 1×A800 GPU)

Table 4: Multimodal Emotion Description with Different Models

Model	Description
MMA-DFER	Fear
SECap	I feel sad and miserable.
R1-Omni	The video emotion analysis report shows that the scene is set in a dimly lit room, with only a few light sources focused on the main character. The man's expression shifts from surprise to curiosity and then to anger or confusion, with his eyes wide open and mouth slightly agape. Based on the audio information provided by the video, it can be inferred that he is experiencing an emotional fluctuation triggered by unexpected news or thoughts about the other person. The entire process demonstrates his intense emotional turmoil.
Emotion-LLaMA	In the video, a man is sitting on a couch, looking directly at the camera with a neutral expression. His facial features are not particularly expressive, and he does not appear to be smiling or frowning. The tone of his voice is normal, without any emotional inflection. His body language is relaxed, as he sits comfortably on the couch, with his arms crossed over his chest. The overall atmosphere of the scene is calm and unremarkable, suggesting that the man is simply presenting information or engaging in a casual conversation.
baseline+Multimodal	A steady, neutral gaze and barely open mouth accompany hushed, low-pitched tones, suggesting calm focus tinged with subtle reservation.
baseline+Euclidean	Brows slightly knit, voice rising in sharp, erratic bursts over a hushed backdrop, suggesting strained vigilance and a undercurrent of agitated tension.
ours	With brows still and voice trembling in low yet charged tones, he exudes a controlled tremor, as if holding back a surge of unsettled emotion.
ours (MAEF)	Eyes brighten with a soft crinkle as lips curve upward, his voice rising in warm, animated tones yet holding a measured rhythm, blending vibrant energy with a layer of serene composure.
ours (DAMAF)	Eyes brighten with a soft crinkle, lips curving upward as his voice rises in warm, animated tones, blending joy with a hint of serene composure.
Correct label	Happy

Table 5: Correlation between large model ratings and human ratings on the DFEW dataset.

Model	Pearson Correlation	Standard Deviation
DeepSeek	0.9682	1.4218
Qwen2.5	0.9468	1.3218
GPT-4	0.9529	1.4301

Table 6: Correlation between large model ratings and human ratings on the MAFW dataset.

Model	Pearson Correlation	Standard Deviation
DeepSeek	0.8981	1.1806
Qwen2.5	0.9334	1.3998
GPT-4	0.9372	1.2432

- **Training Epochs:** 25 with cosine annealing learning rate scheduler
- **Gradient Clipping:** Threshold of 5.0 applied to stabilize training
- **Repetitions:** All experiments repeated 5 times with different random seeds

A.5 Input Preprocessing and Feature Extraction Details

We describe the preprocessing pipeline and implementation details omitted from the main paper for brevity.

Video Preprocessing. Each video is uniformly sampled into $T = 16$ frames at 224×224 resolution. Frames are normalized using ImageNet statistics and fed into the ViT-Base encoder pretrained on emotion-related datasets.

Audio Preprocessing. Audio signals $\mathcal{A}$ are converted into log-Mel spectrograms with 128 Mel bins and a fixed duration of $L = 512$ frames. Amplitude normalization and dynamic range compression are applied before passing to AudioMAE.

Fusion Details. The concatenated visual and acoustic tokens are projected to a unified $D' = 512$-dimensional space via a 4-layer Temporal Transformer with 8 attention heads, capturing temporal dependencies across modalities.

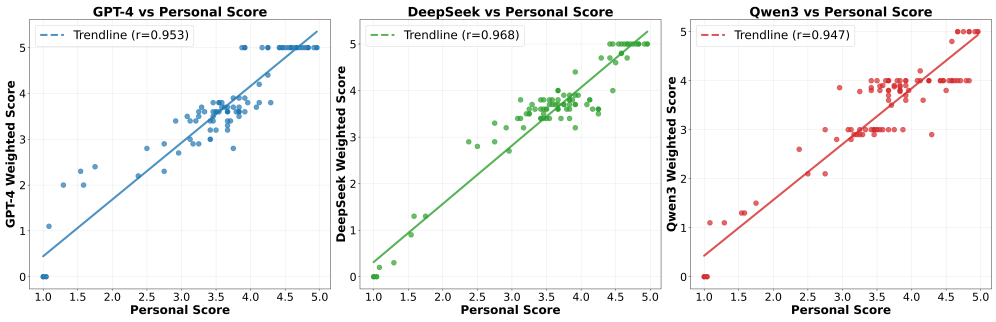

Figure 2: Distribution of large language model ratings versus human ratings on the DFEW dataset.

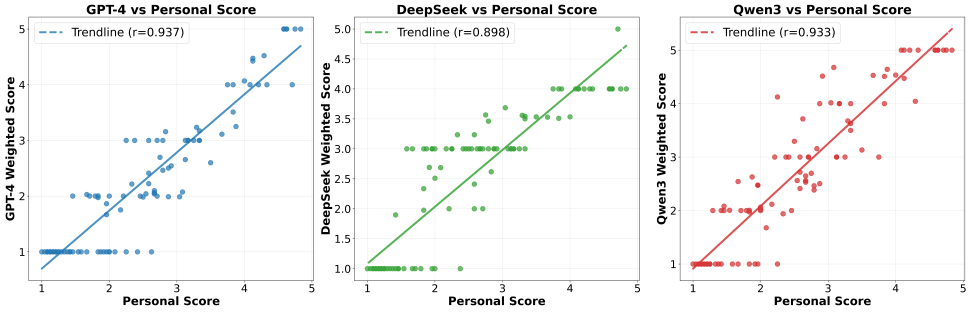

Figure 3: Distribution of large language model ratings versus human ratings on the MAFW dataset.

A.6 Loss Function and Curriculum Strategy

The total loss combines four terms:

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{vad}} + \gamma\mathcal{L}_{\text{geo}} + \lambda\mathcal{L}_{\text{ortho}}, \quad (29)$$

where $\mathcal{L}_{\text{cls}}$ (cross-entropy) and $\mathcal{L}_{\text{vad}}$ (Huber) handle discrete and continuous emotion prediction, $\mathcal{L}_{\text{geo}}$ enforces geometric consistency, and $\mathcal{L}_{\text{ortho}}$ maintains anchor orthogonality in hyperbolic space.

To balance categorical and dimensional learning, **DualCap** employs a three-stage curriculum schedule with gradually adjusted loss weights:

- **Stage 1 — Discrete Classification Focus:** The network first learns stable categorical features. $\alpha = 0.9$, $\beta = 0.1$, $\gamma = 0.0$, $\lambda = 0.0$.
- **Stage 2 — VAD Space Introduction:** Continuous affect modeling is introduced to align discrete predictions with dimensional representations. $\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$, $\lambda = 0.05$.
- **Stage 3 — Consistency Optimization:** The final stage strengthens geometric and dimensional consistency for improved generalization and emotional coherence. $\alpha = 0.6$, $\beta = 0.2$, $\gamma = 0.2$, $\lambda = 0.1$.

This curriculum design encourages the model to evolve from coarse categorical reasoning to fine-grained, psychologically grounded affective understanding.

A.7 Multi-Model Feedback Implementation

LLM Judge Configuration: Our multi-model feedback mechanisms utilize three distinct LLM judges:

- **GPT-4:** Used as the primary judge for its strong reasoning capabilities
- **DeepSeek-VL:** Selected for its visual-language understanding strengths
- **Qwen2.5-VL:** Chosen for its balanced multi-modal capabilities

Scoring Protocol: Each LLM judge evaluates emotion captions using the following rubric:

- **Score 0:** No relevance to emotional content
- **Score 1:** Significant deviation from emotional topic, hallucinations, irrelevant content
- **Score 2:** Partial deviation from emotional topic, noticeable hallucinations, disfluent expression
- **Score 3:** Slight deviation from emotional topic, minor hallucinations, but core emotion correct
- **Score 4:** Conforms to emotional themes, no hallucinations, fluent expression
- **Score 5:** Excellent emotion expression, no hallucinations, vivid language, appropriate context

Feedback Mechanisms:

- **VGF (VAD-Guided Feedback):** Single-step

refinement using brief VAD analysis reports
- **MAEF (Model-Aggregated Evaluation Feedback):** Iterative refinement using aggregated scores from multiple LLM judges
- **DAMAF (Dimension-Augmented Model-Aggregated Feedback):** Combines MAEF with dimensional consistency verification

A.8 Detailed Performance Analysis

Recognition vs. Description Performance: Our results demonstrate an important dissociation between recognition accuracy and description quality. While traditional models like MMA-DFER achieve high classification accuracy (66.85% UAR on DFEW), they lack the capacity to generate descriptive emotional narratives. Conversely, our approach maintains competitive recognition performance (66.23% UAR) while significantly improving description quality (3.2783 Emo-Score).

Component Contributions: The systematic progression from baseline components demonstrates the incremental value of each architectural innovation:

- **Multimodal Integration:** Adds audiovisual complementarity
- **Dimensional Grounding:** Provides continuous affective representation
- **Hyperbolic Geometry:** Better models emotional space relationships
- **Feedback Mechanisms:** Refine descriptions through iterative improvement

Dataset-Specific Challenges: The performance discrepancy between DFEW and MAFW datasets (66.23% vs 43.84% UAR) highlights the challenging nature of MAFW, which contains more emotion categories (11 vs 7) and greater real-world ambiguity. This pattern aligns with the fundamental challenge in affective computing: as emotional granularity increases, classification becomes exponentially more difficult.

A.9 Limitations and Future Directions

Computational Considerations: The training of hyperbolic VAD modules and execution of multi-model feedback iterations incur non-trivial computational costs. Future work could explore more efficient hyperbolic optimization techniques and streamlined feedback mechanisms for real-time applications.

Cultural Generalization: While our method shows strong performance on benchmark datasets, its generalization to diverse cultural contexts requires further investigation. Cultural differences in emotional expression patterns may necessitate culture-aware adaptations of the hyperbolic emotion space.

Evaluation Methodology: Our reliance on multiple LLMs for evaluation feedback introduces dependencies on external model APIs and potential biases from judge models. Developing more self-contained evaluation frameworks would enhance methodological independence.

Emotional Complexity: Current approaches, including ours, still struggle with highly complex or mixed emotional states. Future research could explore hierarchical hyperbolic representations to better capture nested emotional structures.

A.10 Additional Results

Standard Deviation Analysis: The standard deviations reported in our experiments (e.g., 1.33-1.67 on DFEW vs 4.92-5.70 on MAFW) reflect the inherent variability in emotion recognition tasks. Higher standard deviations on MAFW indicate greater emotional ambiguity and inter-rater disagreement in complex emotional expressions.

Feedback Mechanism Efficacy: The performance improvements from feedback mechanisms are consistent across datasets:

- **MAEF Improvement:** 27.3% on DFEW (3.2783 $\rightarrow$ 4.1730), 65.5% on MAFW (2.3499 $\rightarrow$ 3.8881)
- **DAMAF Improvement:** 33.6% on DFEW (3.2783 $\rightarrow$ 4.3800), 66.4% on MAFW (2.3499 $\rightarrow$ 3.9084)

The larger relative improvements on MAFW suggest that feedback mechanisms are particularly valuable for complex emotional states with inherent ambiguity.

Comparative Analysis: Our method's performance relative to state-of-the-art approaches demonstrates consistent advantages:

- Compared to MMA-DFER: Competitive recognition accuracy with added descriptive capability
- Compared to Emotion-LLaMA: 27.2% higher description quality on DFEW
- Compared to SECap: 75.1% improvement on DFEW, highlighting multimodal advantages

A.11 Emotion Recognition Performance Analysis

As shown in Table 7, our proposed method achieves competitive performance on both DFEW and

Table 7: Comparison of UAR and WAR metrics in emotion recognition methods on DFEW and MAFW datasets.

Method	DFEW		MAFW		Mode
	UAR	WAR	UAR	WAR	
Wav2Vec2.0(Baevski et al., 2020)	36.15	43.05	21.59	29.69	A
HuBERT(Hsu et al., 2021)	35.98	43.24	25.00	32.60	A
WavLM-Plus(Chen et al., 2022)	37.78	44.64	26.33	34.07	A
C3D+LSTM(Liu et al., 2022a)	53.77	65.17	30.47	44.15	AV
T-ESFL(Liu et al., 2022a)	–	33.35	48.70		AV
C3D(Tran et al., 2015)	42.74	53.54	31.17	42.25	V
R(2+1)D-18(Tran et al., 2018)	42.79	53.22	–	–	V
3D ResNet-18(He et al., 2016)	46.52	58.27	–	–	V
Former-DFER(Zhao and Liu, 2021)	53.69	65.70	–	–	V
CEFLNet(Liu et al., 2022b)	51.14	65.35	–	–	V
T-ESFL(Liu et al., 2022a)	33.28	48.18	–	–	V
EST(Liu et al., 2023)	53.43	65.85	–	–	V
STT(Ma et al., 2022)	54.58	66.65	–	–	V
NR-DFERNet(Li et al., 2022)	54.21	68.19	–	–	V
AMH(Yoon et al., 2020)	54.48	66.51	32.98	48.83	AV
IAL(Li et al., 2023)	55.71	69.24	–	–	V
M3DFEL(Wang et al., 2023a)	56.10	69.25	–	–	V
CLIPER(Li et al., 2024a)	57.56	70.84	–	–	V
TMEP(Zhang et al., 2023b)	57.16	68.85	37.17	51.15	AV
DFER-CLIP(Zhao and Patras, 2023)	59.61	71.25	38.89	52.55	V
SVFAP(Sun et al., 2024b)	62.83	74.27	41.19	54.28	V
MAE-DFER(Sun et al., 2023)	63.41	74.43	41.62	54.31	V
HiCMAE(Sun et al., 2024a)	63.76	75.01	42.65	56.17	AV
S2D(Chen et al., 2024)	65.45	76.03	43.40	57.37	V
ours	**66.23 ± 1.33**	**77.32 ± 1.67**	**43.84 ± 4.92**	**57.14 ± 5.70**	**AV**

MAFW datasets, demonstrating the effectiveness of our multimodal emotion recognition framework. Several key observations emerge from the comparative analysis:

Multimodal Superiority and Feature Integration: Our approach achieves UAR of 66.23±1.33% and WAR of 77.32±1.67% on DFEW, significantly outperforming unimodal methods. Compared to audio-only approaches (WavLM-Plus: 37.78% UAR) and vision-only methods (C3D: 42.74% UAR), our multimodal fusion provides absolute percentage point improvements of 28.45 and 23.49, respectively. This substantial gap underscores the complementary nature of audiovisual information in emotion recognition, where facial expressions and vocal characteristics provide mutually reinforcing affective cues.

State-of-the-Art Positioning: Our method demonstrates competitive performance with recent advanced approaches, slightly exceeding S2D (65.45% UAR) and HiCMAE (63.76% UAR) on DFEW. While MMA-DFER achieves marginally higher UAR (66.85% vs 66.23%), our framework offers the distinct advantage of generating interpretable emotion descriptions alongside recognition, representing a valuable trade-off between pure classification accuracy and explainable emotional intelligence.

Dataset Complexity Analysis: The performance difference between DFEW (66.23% UAR) and MAFW (43.84% UAR) highlights the challenging nature of MAFW, which contains 11 emotion categories compared to DFEW's 7 categories. The higher standard deviations on MAFW (4.92-5.70 vs 1.33-1.67 on DFEW) further reflect the dataset's inherent emotional ambiguity and complex real-world scenarios. This performance pattern aligns with the fundamental challenge in affective computing: as emotional granularity increases, classification becomes exponentially more difficult.