

GeoMER: A Joint Euclidean–Hyperbolic Geometry Framework for Multimodal Emotion Recognition

Ting Zhu¹, Huizhi Liang¹, Boguslaw Obara¹, Shufei Duan², and Varun Ojha¹

¹ Newcastle University, Newcastle upon Tyne, United Kingdom

² Taiyuan University of Technology, Shanxi, China

{T.zhu11, Huizhi.Liang, Boguslaw.Obara, Varun.Ojha}@newcastle.ac.uk
duanshufei@tyut.edu.cn

Abstract. Multimodal Emotion Recognition (MER) aims to understand human emotions by integrating information from multiple modalities such as speech and text. Psychological studies suggest that emotions exhibit latent hierarchical structured relationships in their semantic representations [1]. However, most MER approaches rely on Euclidean representations, which capture fine-grained cross-modal interactions but do not explicitly provide a geometric prior for organizing hierarchical emotional structures. To mitigate this gap, we propose GeoMER, a joint Euclidean-Hyperbolic modeling framework that introduces hyperbolic emotion-core routing to capture higher-level, abstract emotional representations for MER. The proposed framework jointly models token-level cross-modal semantic interactions and latent emotion representations in two complementary geometric spaces. Specifically, GeoMER consists of a Euclidean Multimodal Fusion Module (EuclFusion) for multimodal representation fusion and a Hyperbolic Geometric Core Routing (HypGeoRoute) module that performs geometry-aware core assignment and aggregation in a hyperbolic manifold to capture more abstract feature representations. The HypGeoRoute module complements the Euclidean token-level modeling. We experiment on the IEMOCAP multimodal MER benchmark and demonstrate that the proposed method achieves the best performance, with 83.54% weighted accuracy and 83.35% unweighted accuracy, outperforming Euclidean baselines, including MemoCMT.

Keywords: Multimodal emotion recognition · Hyperbolic geometry · Speech emotion recognition · Multimodal feature fusion · Affective computing.

1 Introduction

Multimodal Emotion Recognition (MER) aims to enable machines to interpret human emotions more accurately by integrating multiple modalities, such as

speech and text [18]. Recent advances in deep neural architectures and cross-modal attention mechanisms [7, 11] have significantly improved MER performance. However, emotional expressions possess hierarchical structured relationships that go beyond simple categories. Psychological emotion theories [14, 15] suggest that emotions have multi-component hierarchical structures and continuous relationships, supporting emotion mixtures (e.g., joy + anticipation $\rightarrow$ optimism) and intensity variations (e.g., annoyance $\rightarrow$ anger $\rightarrow$ rage). While modern Transformer-based models can learn expressive representations in Euclidean space, standard Euclidean latent spaces do not explicitly provide a geometric structure for organizing hierarchical emotional structures. In contrast, hyperbolic geometry naturally excels at representing tree-like or multi-level relationships [12], making it a suitable complementary representation space for modeling structured underlying emotional cues. These limitations motivate exploring non-Euclidean representation spaces to model structured emotion in MER [3].

Hyperbolic geometry has recently attracted attention as a powerful representation space for hierarchical data due to its exponential expansion [5]. Hyperbolic representation learning has been successfully applied in various domains, including natural language processing and affective computing [3, 5]. Fine-grained emotion classification [20] suggests that emotion categories are not isolated labels, instead exhibiting multi-level and intensity-dependent relationships. HypEmo [3] introduces label-aware hyperbolic embeddings to model hierarchical structures among fine-grained emotion categories. Earlier works applied hyperbolic layers to multimodal sentiment and emotion recognition [1], implementing both sequential and parallel mixing schemes between Euclidean and hyperbolic dense layers. Although these studies demonstrate the potential of hyperbolic geometry for affective computing, they focus on the sample’s label hierarchy or on general Euclidean and hyperbolic layer composition. On the contrary, our GeoMER is a task-specific dual-geometry design framework.

This design stems from the fact that the MER requires both abstract structure modeling and cross-modal semantic alignment between heterogeneous modalities such as speech and text [5]. At the same time, mapping all representations into hyperbolic space may improve structural modeling, but weaken the advantages of Euclidean representations in cross-modal alignment and introduce an optimization stability problem. Therefore, developing a unified, geometry-aware hybrid model that preserves the strengths of Euclidean representations and hyperbolic geometry is more suitable for MER. Thus, unlike previous works, our work is not a simple replacement of Euclidean embeddings with hyperbolic ones or a direct reuse of these components for MER. Instead, GeoMER is a joint Euclidean-Hyperbolic framework for multimodal emotion recognition. It consists of two branches: a) the Euclidean branch learns discriminative cross-modal representations with cross-attention transformers, and b) the hyperbolic branch models the latent structure of emotions in a hyperbolic manifold. The two branches are jointly optimized and integrated via a unified fusion mechanism, enabling the model to simultaneously preserve both Euclidean discriminative

representations and more abstract hyperbolic features. The main contributions are as follows:

1. We propose GeoMER, a joint Euclidean-Hyperbolic MER framework that specializes both the discriminative semantic modeling (Euclidean space) and abstract hierarchical structural modeling (hyperbolic space).
2. We propose a HypGeoRoute module that softly assigns token representations to learnable geometric cores in the Poincaré ball via distance-aware routing.
3. We propose a core-wise hyperbolic self-attention mechanism that captures relationships among geometric cores in the Poincaré ball, preserving geometric consistency while enhancing contextual modeling.

2 Related Work

Early MER studies [18] relied on handcrafted features and employed simple fusion methods, which were sensitive to noise. Recent works have used deep neural networks with attention mechanisms [11] and Transformer-based architectures such as TDFNet [21] and MemoCMT [7] to capture dynamic interactions across modalities.

To better capture structured emotional relationships, researchers have introduced hyperbolic representation learning as it can model hierarchical data with exponential capacity [5]. The hyperbolic Transformer extends the self-attention mechanism to non-Euclidean spaces for hierarchical sequence modeling [19]. Several studies have explored the application of hyperbolic modeling in MER, including conversational emotion recognition with hyperbolic hypergraph fusion [22], hyperbolic fusion of heterogeneous speech representations [13], and label-aware hyperbolic embeddings for fine-grained emotion classification [3]. Meanwhile, hybrid geometric learning has been preliminarily studied, in which Euclidean encoders are combined with hyperbolic components for multimodal representation learning [1]. Emotion Collider [5], HypEmo [3], and HyperPoint [17] further explore hyperbolic modeling for emotion mixture, label-aware classification, and multimodal representation learning.

However, there are two key limitations. First, most existing MER methods operate in Euclidean space, which struggles to capture the more structured and abstract emotional representations. Although hyperbolic approaches have been explored, they remain limited in scope and often focus on specific components, such as embeddings or labels. Second, hybrid geometric methods remain underdeveloped, typically combining Euclidean and hyperbolic representations in a loosely coupled manner without a unified framework for joint optimization.

3 Methodology

We propose GeoMER, a Euclidean-Hyperbolic modeling framework for MER that models discriminative cross-modal semantics in Euclidean space and structured emotion representations on a hyperbolic manifold. As shown in Figure 1,

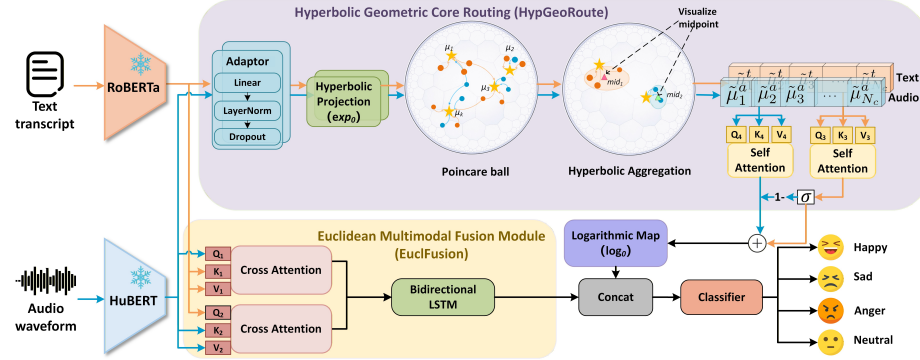


Fig. 1. Overview of the GeoMER framework. Text and audio features extracted by pre-trained encoders are first aligned by a Euclidean adaptor. EuclFusion models multi-modal interactions in Euclidean space, while HypGeoRoute captures abstract emotional features via geometric core routing in a hyperbolic manifold. The resulting hyperbolic representations are mapped back to Euclidean space and concatenated with Euclidean features for the final emotion classification.

the framework consists of feature extraction, Euclidean multimodal fusion, hyperbolic core routing, and final emotion classification. Specifically, *Euclidean Multimodal Fusion Module (EuclFusion)* models cross-modal interactions, and the *Hyperbolic Geometric Core Routing (HypGeoRoute)* module projects multimodal token representations onto a hyperbolic manifold and organizes them around learnable geometric cores via geometry-aware routing and core-wise aggregation. The final representation is obtained by concatenating the two representations, enabling GeoMER to jointly exploit multimodal interactions and abstract geometric features for emotion prediction.

3.1 Feature Embedding

We use the HuBERT encoder to encode the raw audio into a series of acoustic embeddings. All audio signals are 16 kHz. The RoBERTa model is employed to obtain contextual tokens from the transcriptions.

3.2 Euclidean Multimodal Fusion Module

The EuclFusion is designed to model cross-modal interactions in Euclidean space. The architecture follows MemoCMT [7] while improving on simple concatenation by introducing a bidirectional Long Short-Term Memory (BiLSTM) network for sequential aggregation. Two multi-head cross-attention mechanisms are first employed to effectively fuse speech and text cues. This step enhances token-level alignment across modalities. While cross-attention focuses on aligning and exchanging information between speech and text tokens, the subsequent BiLSTM serves a different role by modeling the temporal evolution of the fused

multimodal sequence at the utterance level. Specifically, the fused multimodal token sequence is processed by a BiLSTM. This step dynamically integrates cross-modal cues over time, which is crucial for capturing the evolving nature of emotional expression in speech [8, 10]. Let the audio and text token sequences $V_a \in \mathbb{R}^{n_a \times d}$ and $V_t \in \mathbb{R}^{n_t \times d}$, where a is for audio and t is for text, n_a and n_t are the token lengths, and d is the embedding dimension. The resulting hidden states are aggregated to produce a fixed-length utterance-level embedding $h_{euc} \in \mathbb{R}^d$. This Euclidean representation captures multimodal interactions in embedding space and provides complementary information to the subsequent hyperbolic modeling branch.

3.3 Hyperbolic Geometric Core Routing

The HypGeoRoute module is designed to capture latent abstract information through geometry-aware token routing in a hyperbolic manifold. Unlike traditional Euclidean representation modeling, HypGeoRoute operates in the Poincaré ball to organize multimodal token embeddings around learnable geometric cores. Tokens are softly routed to multiple cores based on hyperbolic distances, enabling geometrically consistent aggregation of multimodal abstract information.

Feature Alignment and Hyperbolic Emotion Representation. The audio and text token embeddings $V_a \in \mathbb{R}^{n_a \times d}$ and $V_t \in \mathbb{R}^{n_t \times d}$ extracted from frozen pre-trained encoders have modality-specific scale and distribution mismatches. To stabilize subsequent geometric operations, we introduce lightweight Euclidean adaptors composed of a linear projection (Linear), layer normalization (LN), and dropout (Drop). Let $W_a, W_t \in \mathbb{R}^{d \times d}$ and $b_a, b_t \in \mathbb{R}^d$ denote learnable projection parameters for audio and text modalities. The adapted audio and text embeddings $\tilde{V}_a$ and $\tilde{V}_t$ are therefore computed as:

$$\tilde{V}_a = \text{Drop}(\text{LN}(V_a W_a + b_a)), \quad \tilde{V}_t = \text{Drop}(\text{LN}(V_t W_t + b_t)). \quad (1)$$

These audio and text features are then projected onto the Poincaré ball model $\mathbb{B}_c^d$ via an exponential map at the origin as: $H_a = \exp_0(\tilde{V}_a)$ and $H_t = \exp_0(\tilde{V}_t)$, where

$$\exp_0(\mathbf{x}) = \tanh(\sqrt{c} \|\mathbf{x}\|) \frac{\mathbf{x}}{\sqrt{c} \|\mathbf{x}\|}, \quad (2)$$

$\mathbf{x} \in \mathbb{R}^d$ is $\tilde{V}_a$ for audio, and $\mathbf{x} \in \mathbb{R}^d$ is $\tilde{V}_t$ for text. We use $c > 0$ to control the magnitude of the negative curvature.

Mixed Abstract Emotion Representation via Geometric Core Routing. Traditional Euclidean approaches typically aggregate all tokens into a single utterance-level vector with a single emotion label. However, in real-world conversations, it may contain multiple emotional cues with varying degrees of dominance, resulting in an emotional mixture rather than a single, discrete emotion. Therefore, we introduce N_c learnable geometric cores $\{\mu_k \mid k = 1, \dots, N_c\} \subset \mathbb{B}_c^d$,

shared across modalities and acting as latent emotional components in hyperbolic space. This routing mechanism organizes token-level emotional cues into higher-level emotional components, enabling the model to learn abstract emotional representations in a hyperbolic manifold. For each modality $m \in \{a, t\}$ with hyperbolic token embeddings $H_m = \{h_i^m\}_{i=1}^{L_m}$, where L_m denotes the sequence length of modality m , we compute the hyperbolic distance d_{ik}^m between a token i and a core k in the Poincaré ball as:

$$d_{ik}^m = \text{arcosh} \left(1 + \frac{2 \|h_i^m - \mu_k\|^2}{(1 - \|h_i^m\|^2)(1 - \|\mu_k\|^2)} \right), \quad (3)$$

where $\|\cdot\|$ denotes the Euclidean norm in the ambient space.

Instead of assigning each utterance to a single core, tokens are softly routed to multiple cores according to their hyperbolic distance, allowing multiple tokens (utterances) to form a hierarchical abstract representation in hyperbolic space. The routing weight α_{ik}^m of token i to core k under modality m is defined as:

$$\alpha_{ik}^m = \frac{\exp(-d_{ik}^m)}{\sum_{j=1}^{N_c} \exp(-d_{ij}^m)}, \quad \sum_{k=1}^{N_c} \alpha_{ik}^m = 1. \quad (4)$$

where j iterates over all N_c cores for normalization.

Token Aggregation for Hyperbolic Cores. Given the routing weights α_{ik}^m , we aggregate modality-specific token embeddings into N_c core-level representations. For each modality m and core k , we employ the Einstein midpoint [6] for computing a geometrically consistent weighted mean in the Poincaré ball. Specifically, hyperbolic token embeddings are first mapped to the tangent space at the origin using $\log_{\mathbf{0}}(\cdot)$. The routing-weighted aggregation is then performed in this space, followed by projection back to the hyperbolic manifold via $\exp_{\mathbf{0}}(\cdot)$:

$$\tilde{\mu}_k^m = \exp_{\mathbf{0}} \left(\sum_{i=1}^{L_m} \alpha_{ik}^m \log_{\mathbf{0}}(h_i^m) \right), \quad (5)$$

where $h_i^m \in \mathbb{B}_c^d$ is the i -th hyperbolic token embedding of modality m , and $\tilde{\mu}_k^m \in \mathbb{B}_c^d$ is the aggregated core representation for the k -th abstract emotion prototype under modality m . This operation is applied for each modality $m \in \{a, t\}$ and each core $k \in \{1, \dots, N_c\}$, obtaining two sets of core-level representations:

$$M^m = [\tilde{\mu}_1^m, \tilde{\mu}_2^m, \tilde{\mu}_3^m, \dots, \tilde{\mu}_{N_c}^m]. \quad (6)$$

Hyperbolic Self-Attention over Emotion Cores. After getting the core-level representations $M^m \in \mathbb{B}_c^{N_c \times d}$ for each modality $m \in \{a, t\}$, we apply Hyperbolic Self-Attention (HSA) across the N_c emotion cores to model inter-core semantic correlations between different abstract emotion prototypes. This mechanism enables interactions among latent emotional features, allowing related abstract emotional features to influence one another on the hyperbolic

manifold. Since linear transformations are not directly defined on Riemannian manifolds, Queries (Q), Keys (K), and Values (V) are projected by Poincaré linear layers ($\log_{\mathbf{0}} \rightarrow \text{Linear} \rightarrow \exp_{\mathbf{0}}$), where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are learnable projection matrices:

$$\begin{aligned} Q^m &= \exp_{\mathbf{0}}(\log_{\mathbf{0}}(M^m) W_Q), \quad K^m = \exp_{\mathbf{0}}(\log_{\mathbf{0}}(M^m) W_K), \text{ and} \\ V^m &= \exp_{\mathbf{0}}(\log_{\mathbf{0}}(M^m) W_V). \end{aligned} \quad (7)$$

We adopt a multi-head mechanism with R attention heads. For the r -th head, the attention score $s_{pq}^{(m,r)}$ and attention weights $a_{pq}^{(m,r)}$ (for $r = 1, 2, \dots, R$) between the p -th query core and the q -th key core are defined as:

$$s_{pq}^{(m,r)} = -\frac{d_{\mathbb{B}}(Q_p^{(m,r)}, K_q^{(m,r)})}{\exp(\tau)}, \quad a_{pq}^{(m,r)} = \frac{\exp(s_{pq}^{(m,r)})}{\sum_{q'=1}^{N_c} \exp(s_{pq'}^{(m,r)})}, \quad (8)$$

where $d_{\mathbb{B}}(\cdot, \cdot)$ denotes the geodesic distance on the Poincaré ball, and τ is a learnable parameter, called temperature in literature.

Given the value representations $V_q^{(m,r)}$, the output $O_p^{(m,r)}$ is computed using the Einstein midpoint as:

$$O_p^{(m,r)} = \text{HypAgg}\left(\{V_q^{(m,r)}\}_{q=1}^{N_c}, \{a_{pq}^{(m,r)}\}_{q=1}^{N_c}\right), \quad (9)$$

where $\text{HypAgg}(\cdot)$ denotes weighted midpoint aggregation on the manifold.

The outputs of all attention heads are first concatenated and projected through a hyperbolic linear transformation $\text{HypLinear}(\cdot)$, and then the resulting representation is combined with the input M^m through a hyperbolic residual connection using Möbius addition. The output Z_m of HSA is defined as:

$$Z_m = \frac{1}{\sqrt{2}} \otimes \left(M^m \oplus \text{HypLinear}\left(\text{Concat}(O^{(m,1)}, \dots, O^{(m,R)})\right) \right), \quad (10)$$

where $\oplus$ denotes Möbius addition and $\otimes$ denotes Möbius scalar multiplication. This residual design preserves manifold consistency under Möbius operations while stabilizing feature propagation across layers.

Cross-modal Gated Fusion. After core-wise hyperbolic self-attention, each modality produces N_c core-specific abstract emotion representations. Given $Z_t = \{z_k^t\}_{k=1}^{N_c}$ and $Z_a = \{z_k^a\}_{k=1}^{N_c}$, we fuse the text and audio representations of each core k directly in the Poincaré ball through Möbius operations, ensuring that the representations remain on the hyperbolic manifold. For each core k , we compute a gate signal $g_k \in (0, 1)$ from the difference of two modality representations in the tangent space at the origin, which measures cross-modal inconsistency:

$$g_k = \sigma(W_g (|\log_{\mathbf{0}}(z_k^t) - \log_{\mathbf{0}}(z_k^a)|) + b_g), \quad (11)$$

where $W_g \in \mathbb{R}^{d \times d}$ and $b_g \in \mathbb{R}^d$ are learnable parameters, and $\sigma(\cdot)$ denotes the sigmoid function. A larger discrepancy produces a stronger gating response, allowing the model to emphasize the more informative modality.

The fused representation of core k is then computed using Möbius scalar multiplication and Möbius addition:

$$f_k = \frac{1}{\sqrt{2}} \otimes ((g_k \otimes z_k^t) \oplus ((1 - g_k) \otimes z_k^a)), \quad (12)$$

where $\otimes$ and $\oplus$ denote Möbius scalar multiplication and Möbius addition, respectively.

The fused core representations $F^{hyp} = \{f_k\}_{k=1}^{N_c}$ are mapped back into Euclidean space to get $F^{euc} = \{f_k^{euc}\}_{k=1}^{N_c}$ via the logarithmic map function. Since an utterance may express multiple latent emotional cues with different contributions, we further estimate the importance of each core at the utterance level. Therefore, we compute a core importance distribution $\{\pi_k\}_{k=1}^{N_c}$ using the softmax function, which reflects how strongly each emotional core contributes to the overall emotional representation. The final representation of the hyperbolic branch (h_{hyp}) is obtained by weighted aggregation of the fused cores as:

$$h_{hyp} = \sum_{k=1}^{N_c} \pi_k f_k^{euc}, \quad (13)$$

where π_k denotes the importance weight of the k -th core.

Finally, we concatenate the representations from the Euclidean branch (h_{euc}) and the hyperbolic branch (h_{hyp}) and then feed them into a linear classifier followed by a softmax function:

$$\hat{y} = \text{Softmax}(W_c \text{Concat}(h_{euc}, h_{hyp}) + b_c), \quad (14)$$

where $\text{Concat}(\cdot)$ is the concatenation operation, W_c and b_c are learnable classifier parameters, and $\hat{y}$ is the predicted probability distribution over emotion classes.

4 Experiments

MER Benchmark We use the IEMOCAP benchmark [2] that consists of 12 hours of audio-visual emotional conversations performed by 10 professional actors (5 male, 5 female). Each session includes both scripted and improvised dialogues designed to elicit natural emotional expressions. It provides multimodal recordings, including audio, text, facial motion capture, and videos. Only speech and text modalities are used in this paper. Following prior studies, four emotion classes, including *angry*, *happy*, *sad*, and *neutral*, are chosen. IEMOCAP contains 5,531 utterances in total, 4479 training set, 498 validation samples, and 554 test samples. The distribution of the training set, test set, and validation set is shown in Table 1. It can be observed that the IEMOCAP is moderately class-imbalanced, where the *happy* and *neutral* categories contain more samples than *angry* and *sad*.

Table 1. Emotion distribution of IEMOCAP under the four-class setting. Percentages are computed within each split.

Emotion	Train		Val		Test		Total	
	#Utterances	(%)	#Utterances	(%)	#Utterances	(%)	#Utterances	(%)
Angry	902	20.1	99	19.9	102	18.4	1103	19.9
Happy	1321	29.5	144	28.9	171	30.9	1636	29.6
Sad	857	19.1	98	19.7	129	23.3	1084	19.6
Neutral	1399	31.2	157	31.5	152	27.4	1708	30.9
Total	4479	100	498	100	554	100	5531	100

Evaluation Metrics. Given the class imbalance in IEMOCAP, we evaluate model performance using both overall and class-balanced metrics. Weighted accuracy (WA), unweighted accuracy (UA), and Macro-F1 are reported. WA measures the overall classification accuracy, while UA averages recall across all classes to better reflect performance under class imbalance. In addition, Macro-F1 is obtained by averaging the F1-scores of all classes.

Experiment Setup. All experiments are implemented in PyTorch and conducted on one NVIDIA RTX 3090 GPU. HuBERT-base and RoBERTa-base are used as frozen acoustic and text encoders, producing 768-dimensional features. The model is optimized using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and weight decay 5×10^{-3} , and the cross-entropy loss. The initial learning rate is 1×10^{-4} and decayed by a factor of 0.1 every 30 epochs. The model is trained for up to 60 epochs with early stopping based on validation UA. The hyperbolic curvature is $c = 1.0$, the hyperbolic routing core is $N_c = 4$, and the scaling coefficient α is learnable.

4.1 Main Results

Table 2 presents the comparison between the proposed method and several representative baselines on the IEMOCAP benchmark under the four-class setting. When only acoustic features are used, our GeoMER achieves 70.80% WA and 70.36% UA, outperforming classical audio baselines such as CNN and LSTM. Compared with MemoCMT (Audio-only), our method improves WA by 4.24% and UA by 5.02%, indicating its benefit even in a single-modality setting. For text-based emotion recognition, GeoMER achieves 73.24% WA and 72.89% UA, outperforming both BERT and MemoCMT (Text-only). This demonstrates that it effectively captures semantic emotional cues from textual inputs. When using the hyperbolic-only (HypGeoRoute), the model achieves competitive performance, with 82.78% WA and 82.64% UA, outperforming the Euclidean baseline, MemoCMT. When both audio and text modalities are used, the proposed GeoMER achieves the best performance, with $83.54 \pm 0.77\%$ WA and $83.35 \pm 0.57\%$ UA across five runs. Compared with the reported Euclidean baseline MemoCMT result, our method improves WA by 1.69% and UA by 2.02%. Compared

Table 2. Comparison with representative multimodal methods on IEMOCAP (4-class). Results marked with [†] are reproduced under our experimental setting and reported as mean $\pm$ standard deviation over five runs; other results are taken from the corresponding published papers. WA and UA indicated the weighted and unweighted accuracies.

Model	WA (%)	UA (%)
Audio-only Modality		
CNN	65.30	64.20
LSTM	66.80	65.90
MemoCMT (Audio-only)	66.56	65.34
GeoMER (Audio-only)	70.80	70.36
Text-only Modality		
BERT	70.50	69.30
MemoCMT (Text-only)	71.10	71.12
GeoMER (Text-only)	73.24	72.89
Text + Audio Modalities		
Fan [4]	78.51	79.31
Sun [16]	78.42	79.71
Lei [9]	78.60	77.60
MemoCMT(reported) [7]	81.85	81.33
MemoCMT(reproduced) [†]	81.05 $\pm$ 0.45	81.33 $\pm$ 0.34
HypGeoRoute(Ours)	<i>82.78$\pm$0.71</i>	<i>82.64$\pm$0.67</i>
GeoMER(Ours)	83.54$\pm$0.77	83.35$\pm$0.57

with our reproduced MemoCMT baseline, GeoMER improves WA by 2.49% and UA by 2.02%. These results demonstrate the effectiveness of multimodal fusion and evidence that incorporating hyperbolic geometry benefits multimodal representation learning. Our method models abstract, latent relationships among emotional representations via hyperbolic geometric routing. Therefore, GeoMER does not replace Euclidean representation learning; instead, it complements it by modeling hyperbolic structure. Overall, the results verify the advantage of this hybrid Euclidean–hyperbolic design for multimodal emotion representations.

4.2 Ablation Study

Table 3 presents the ablation results of the proposed components on the IEMOCAP benchmark. Based on the MemoCMT baseline (81.05% WA and 81.33% UA), adding the Euclidean Multimodal Fusion Module (EuclFusion) improves the performance to 81.95% WA and 82.38% UA. This corresponds to an improvement of 0.90% in WA and 1.05% in UA, demonstrating that enhanced cross-modal interaction can effectively improve emotion representation learning. We further explore several variants of hyperbolic modeling. When the Hyperbolic Geometric Core Routing (HypGeoRoute) module is further incorporated, the full GeoMER model achieves the best performance with 83.54% WA and 83.35% UA. Compared with the MemoCMT baseline, the proposed method improves WA by 2.49% and UA by 2.02%. Similar improvements are also observed on Macro-F1, which increases from 80.84% to 83.17%. These results indicate

Table 3. Ablation study of different components on IEMOCAP (4-class). WA, UA, and Macro-F1 denote weighted accuracy, unweighted accuracy, and macro-averaged F1-score, respectively. Δ WA/ Δ UA reports the absolute gain over the reproduced MemoCMT baseline.

Model	WA (%)	UA (%)	Macro-F1 (%)	Δ WA/ Δ UA
MemoCMT(reproduced)	81.05	81.33	80.84	—
+EuclFusion	81.95	82.38	81.77	+0.90/+1.05
+Hyperbolic mean pooling	82.16	82.53	81.95	+1.11/+1.20
+Hyp self-attention (w/o Routing)	82.51	82.99	82.31	+1.46/+1.66
+Hyp cross-attention (w/o Routing)	82.54	82.60	82.42	+1.49/+1.27
GeoMER (+HypGeoRoute) (Ours)	83.54	83.35	83.17	+2.49/+2.02

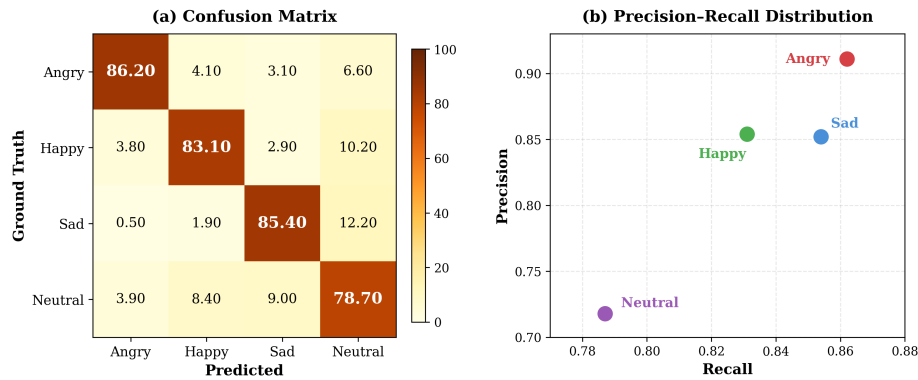


Fig. 2. Per-class performance analysis of GeoMER on the IEMOCAP benchmark. (a) Confusion matrix, where rows denote ground-truth labels and columns denote predicted labels. The values indicate the classification percentage for each emotion category. (b) Precision–recall distribution of each emotion category.

that the hyperbolic routing mechanism can effectively capture latent hierarchical structures in emotional representations. All ablation results show that Euclidean multimodal fusion and hyperbolic geometric modeling play complementary roles. The Euclidean branch strengthens speech and text cross-modal interaction, and the hyperbolic branch provides a geometry-aware structure for organizing latent abstract representations. This supports the effectiveness of the hybrid Euclidean–hyperbolic design for multimodal emotion recognition.

Figure 2 (a) shows the confusion matrix of GeoMER on the IEMOCAP benchmark. Angry achieves the highest recall of 86.20%, followed by Sad (85.40%) and Happy (83.10%). Neutral remains the most challenging category, with a recall of 78.70%. Specifically, neutral samples are more frequently misclassified as Sad (9.00%) and Happy (8.40%), reflecting the inherent ambiguity of neutral emotional expressions in natural conversations. As shown in Figure 2 (b), the precision–recall distribution for each emotion category on the IEMOCAP benchmark has the same pattern. Angry achieves the best overall balance, with

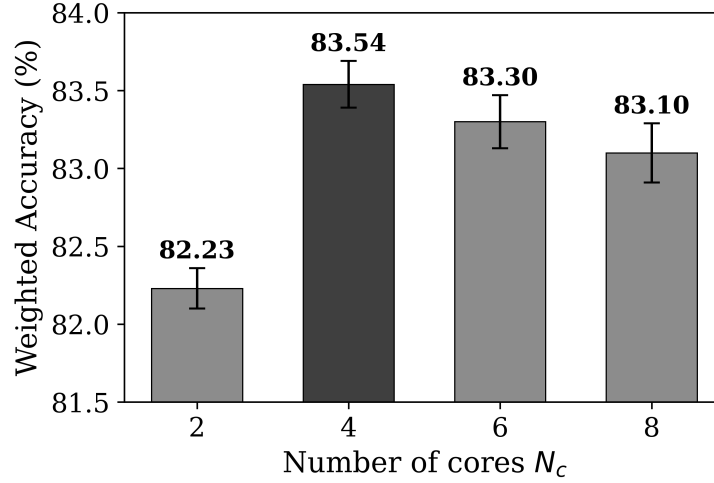


Fig. 3. Effect of the number of emotion cores N_c in HypGeoRoute on weighted accuracy(%). Performance is best when $N_c = 4$, but larger values also yield good results. This suggests that the learned emotion cores are not restricted to a strict one-to-one mapping to emotion classes, but instead serve as latent affective anchors for flexible representation learning.

91.10% precision and 86.20% recall. Sad also shows strong performance with 85.20% precision, and Happy reaches 85.40% precision. While Neutral obtains the lowest precision (71.80%), mainly because samples from other categories are often predicted as Neutral. Overall, GeoMER achieves relatively balanced recognition across all four categories, with a per-class recall range of only 7.50%. This suggests that the proposed hyperbolic core routing mechanism helps the model capture more discriminative representations.

As shown in Figure 3, we further analyze the effect of the number of emotion cores N_c in HypGeoRoute. Performance improves WA from 82.23% to 83.54% as N_c increases from 2 to 4, suggesting that a small number of cores may be insufficient to capture diverse affective structures. Larger values ($N_c = 6$ or 8) are slightly lower but still achieve competitive performance, with weighted accuracies of 83.30% and 83.10%, respectively. It indicates that the performance of GeoMER does not strictly rely on a one-to-one correspondence between the number of cores and the number of emotion classes. Instead, the learned cores serve as latent anchors that support soft routing and flexible representation learning. In addition, the soft routing mechanism assigns each token to all cores with continuous weights via softmax, so tokens from a single utterance naturally distribute across multiple cores rather than to one. Therefore, we set $N_c = 4$ in the final model, as it achieves the best balance between performance and model complexity.

5 Conclusion

In this paper, we propose GeoMER, a dual Euclidean–hyperbolic framework for multimodal emotion recognition. The proposed method combines a Euclidean Multimodal Fusion Module (EuclFusion) for modeling cross-modal interactions with a Hyperbolic Geometric Core Routing (HypGeoRoute) module to capture latent hierarchical structures in emotional representations. Unlike existing hyperbolic emotion embedding methods that mainly focus on label-level or representation-level modeling, distance-aware assignment, manifold-consistent aggregation, core-wise hyperbolic self-attention, and cross-modal gated fusion on the Poincaré ball, our HypGeoRoute module improves performance by capturing hierarchical structure across modalities. By integrating Euclidean multimodal fusion with hyperbolic geometric modeling, GeoMER provides a more expressive representation for complex emotional structure. Experimental results on the IEMOCAP benchmark show that the proposed method performs well, achieving 83.54% WA and 83.35% UA, outperforming the MemoCMT baseline by 2.49% WA and 2.02% UA. The ablation study further validates the effectiveness of the proposed components.

References

1. Araño, K.A., Orsenigo, C., Soto, M., Vercellis, C.: Multimodal sentiment and emotion recognition in hyperbolic space. *Expert Systems with Applications* **184**, 115507 (2021)
2. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation* **42**(4), 335–359 (2008)
3. Chen, C.Y., Hung, T.M., Hsu, Y.L., Ku, L.W.: Label-aware hyperbolic embeddings for fine-grained emotion classification. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 10947–10958 (2023)
4. Fan, W., Xing, X., Cai, B., Xu, X.: Mgat: Multi-granularity attention based transformers for multi-modal emotion recognition. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5 (2023)
5. Fu, R., Wang, Z., Yin, S., Zhang, W., Wei, H., Liu, K., Li, X., Su, Z., Fong, S.: Emotion collider: Dual hyperbolic mirror manifolds for sentiment recovery via anti emotion reflection. *arXiv preprint arXiv:2602.16161* (2026)
6. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic neural networks. *Advances in Neural Information Processing Systems* **31** (2018)
7. Khan, M., Tran, P.N., Pham, N.T., El Saddik, A., Othmani, A.: Memocmt: multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific reports* **15**(1), 5473 (2025)
8. Lee, Y., Yoon, S., Jung, K.: Multimodal speech emotion recognition using cross attention with aligned audio and text. In: *21st Annual Conference of the International Speech Communication Association, INTERSPEECH 2020*. pp. 2717–2721. International Speech Communication Association (2020)

9. Lei, J., Zhang, T., Wang, Y., Wang, Y.: Scmi-net: Semantic constraints and modal interaction network for multimodal emotion recognition. *Journal of Intelligent Information Systems* pp. 1–23 (2026)
10. Li, H., Ding, W., Wu, Z., Liu, Z.: Learning fine-grained cross modality excitement for speech emotion recognition. In: *Proceedings of Interspeech*. pp. 3375–3379 (2021)
11. Makhmudov, F., Kultimuratov, A., Cho, Y.I.: Enhancing multimodal emotion recognition through attention mechanisms in BERT and CNN architectures. *Applied Sciences* **14**(10), 4199 (2024)
12. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems* **30** (2017)
13. Phukan, O.C., Akhtar, M.M., Behera, S.R., Reddy, P.B., Buduru, A.B., Sharma, R., et al.: Hyfuse: Aligning heterogeneous speech pre-trained representations in hyperbolic space for speech emotion recognition. In: *Proceedings of Interspeech 2025*. pp. 131–135 (2025)
14. Plutchik, R.: *The emotions*. Bloomsbury Publishing PLC (1991)
15. Russell, J.A.: A circumplex model of affect. *Journal of Personality and Social Psychology* **39**(6), 1161 (1980)
16. Sun, D., He, Y., Han, J.: Using auxiliary tasks in multimodal fusion of wav2vec 2.0 and bert for multimodal emotion recognition. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 1–5. IEEE (2023)
17. Sun, Y., Cheng, H., Lu, C., Li, Z., Wu, M., Lu, H., Zhu, J.: Hyperpoint: Multimodal 3d foundation model in hyperbolic space. *Pattern Recognition* p. 112800 (2025)
18. Wu, C., Cai, Y., Liu, Y., Zhu, P., Xue, Y., Gong, Z., Hirschberg, J., Ma, B.: Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. pp. 6257–6274 (2025)
19. Yang, M., Verma, H., Zhang, D.C., Liu, J., King, I., Ying, R.: Hypformer: Exploring efficient transformer fully in hyperbolic space. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. pp. 3770–3781 (2024)
20. Zhao, B., Okawa, M., Bigelow, E.J., Yu, R., Ullman, T., Lubana, E.S., Tanaka, H.: Emergence of hierarchical emotion organization in large language models. *arXiv preprint arXiv:2507.10599* (2025)
21. Zhao, Z., Wang, Y., Shen, G., Xu, Y., Zhang, J.: Tdfnet: Transformer-based deep-scale fusion network for multimodal emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **31**, 3771–3782 (2023)
22. Zheng, Y., Chen, G., Song, W., Liu, Y., Zhang, P.: Multimodal hyperbolic embedding and hyperbolic hypergraph fusion for emotion recognition in conversation. In: *Proceedings of the 7th ACM International Conference on Multimedia in Asia*. pp. 1–8 (2025)