



Rectifying spatiotemporal discrepancies for audio-visual emotion recognition via generative semantic anchors

Qiyin Zhou ^a, Ke Wang ^a, Yuanyuan Liu ^{a,*}, Kejun Liu ^a, Wujie Zhou ^b,
Chang Tang ^c

^a School of Computer Science, China University of Geosciences (Wuhan), Wuhan, China

^b School of Artificial Intelligence and Information Engineering, Zhejiang University of Science and Technology, Hangzhou, China

^c School of Software Engineering, Huazhong University of Science and Technology, Wuhan, China

ARTICLE INFO

Keywords:

Audio-visual emotion recognition

Spatiotemporal heterogeneity

Optimal transport

Affine transformation

ABSTRACT

Audio-visual Emotion Recognition (AVER) in the wild is heavily challenged by significant spatiotemporal discrepancies across modalities. Existing methods predominantly enforce explicit physical-level alignment, which overlooks the implicit spatiotemporal heterogeneity of emotional semantics and causes distorted cross-modal feature fusion. We argue that direct audio-visual alignment lacks a stable common ground, requiring a modality-agnostic semantic reference to serve as a unified benchmark. To this end, we propose the Rectifying Spatiotemporal Discrepancies (RSD) framework, which introduces generative text semantic anchors as this high-level reference. Rather than forcibly aligning audio and video to each other, RSD anchors both modalities to the text semantics at two distinct levels. Temporally, the Temporal Anchor Rectification (TAR) module formulates sequence synchronization as a semi-relaxed optimal transport problem to align asynchronous audio-visual streams with the text anchors. Spatially, the Spatial Anchor Rectification (SAR) module further rectifies heterogeneous spatial topologies through a Rank-Restricted Stochastic Alignment (RRSA) mechanism, enabling structurally consistent cross-modal representations. Building upon the rectified temporal and spatial representations, Multi-task Constrained Anchoring Fusion (MCAF) integrates these rectified representations under multi-task constraints to achieve robust emotion prediction. Extensive experiments demonstrate that RSD outperforms state-of-the-art methods, achieving 65.95% WAR on MAFW and 63.64% WAR on AFEW, validating its effectiveness in resolving implicit spatiotemporal discrepancies. Source code is available at <https://github.com/Tank9971/RSD>.

1. Introduction

Audio-visual emotion recognition (AVER) aims to infer human emotional states by jointly modeling facial expressions and speech signals. This approach provides richer and more robust emotion understanding than single-modal methods and supports a wide range of applications [1] such as human-computer interaction [2,3], affective computing [4,5] and social robotics [6]. However, due to the inherent heterogeneity between modalities, AVER faces fundamental challenges in effectively aligning. Fundamentally, this heterogeneity encompasses both explicit feature misalignments and implicit spatiotemporal discrepancies.

Although existing research has made significant progress in cross-modal learning, these efforts have predominantly focused on explicit cross-modal alignment. According to different learning paradigms, existing methods can be categorized into supervised learning, unsuper-

vised learning, and robustness-oriented multimodal fusion methods. Supervised learning methods leverage dense label guidance to capture complex pattern mappings, which can be viewed as explicitly aligning multimodal features under strong supervisory signals. For instance, VIPINN [7] incorporates domain-specific physical equations as structured priors to model spatial-temporal dynamics, while Graph-MDETR [8] utilizes relational graphs encoding component co-occurrence priors to enhance object detection under occlusion. However, these supervised priors are anchored to predefined physical or structural relationships and cannot adapt to the fluid semantic evolution of emotions across audio-visual streams. Unsupervised learning methods aim to extract transferable or domain-invariant representations without relying on labeled target data, essentially seeking explicit alignment across domains or modalities in a label-free manner. For example, SCMGIDTL [9] propagates information across graph structures to achieve unsupervised machinery fault diagnosis, yet its alignment remains at the distri-

* Corresponding author.

E-mail addresses: zhouqiyin@cug.edu.cn (Q. Zhou), liuyy@cug.edu.cn (Y. Liu).

<https://doi.org/10.1016/j.inffus.2026.104780>

Received 26 April 2026; Received in revised form 9 September 2026; Accepted 10 September 2026

Available online 14 September 2026

1566-2535/© 2026 Published by Elsevier B.V.

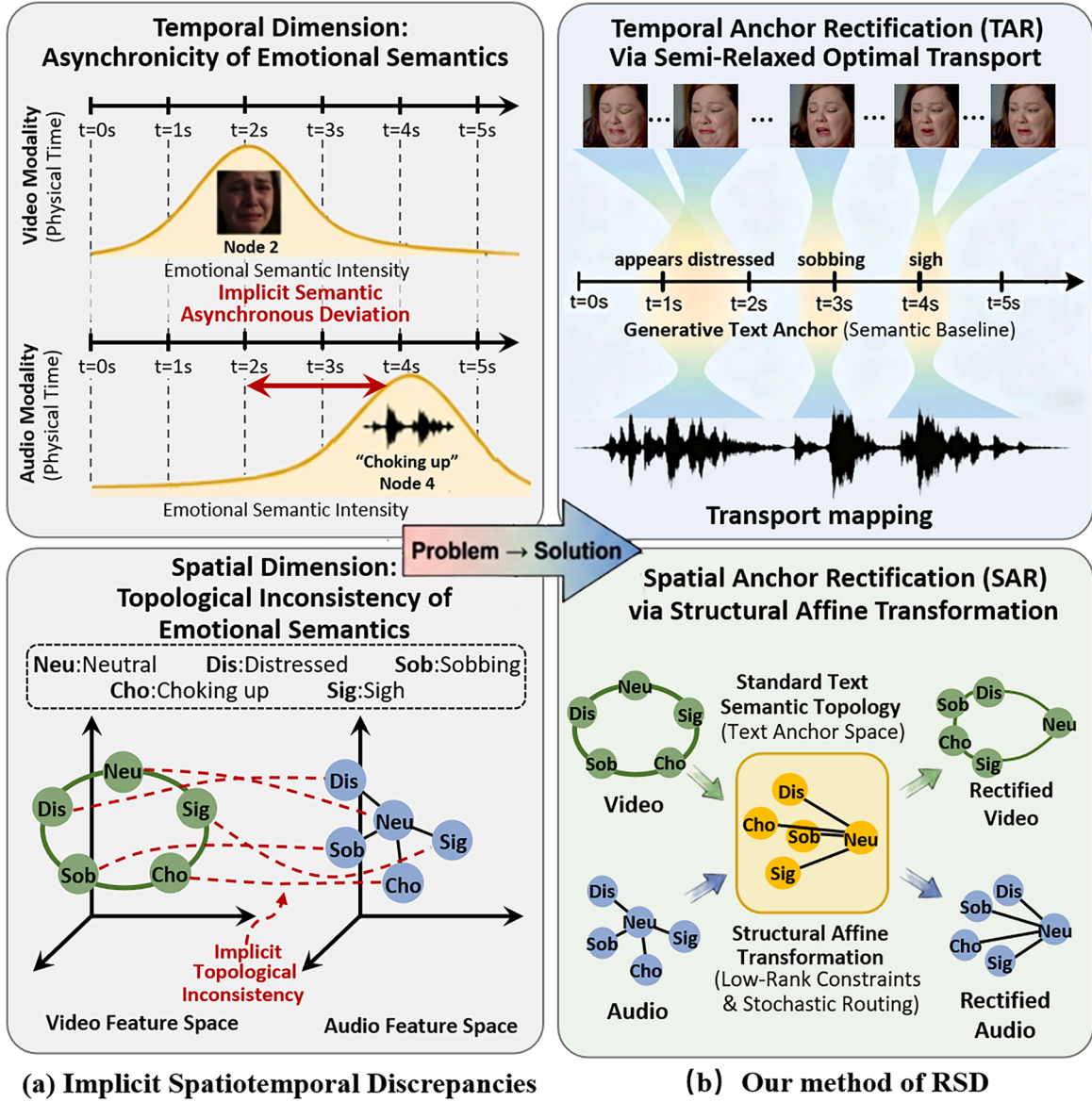


Fig. 1. Illustration of spatiotemporal discrepancies. (a) In multimodal emotion recognition, implicit spatiotemporal heterogeneity causes asynchronous temporal deviation and inconsistent spatial topologies, distorting representations in emotional semantic space. (b) Our RSD addresses it at two distinct levels. Temporally, TAR uses semi-relaxed optimal transport to synchronize audio-visual sequences with text anchors. Spatially, SAR structurally aligns modality-specific topologies with the text semantic space via low-rank constraints and stochastic routing.

bution level without reconciling the underlying temporal and geometric inconsistencies in emotional semantics. Robustness-oriented multimodal fusion methods specifically focus on aligning and fusing heterogeneous modalities under noisy or mismatched conditions. In audio-visual emotion recognition, M3-JEPA [10] employs a joint-embedding predictive architecture to map multimodal features into a unified latent space for cross-modal alignment, while MM-ORIENT [11] utilizes hierarchical interactive attention to achieve explicit interaction and alignment between decoupled multimodal features. Despite their effectiveness within respective paradigms, these methods primarily address explicit feature-level mismatches and fundamentally overlook the implicit spatiotemporal heterogeneity in underlying emotional semantics.

This implicit heterogeneity can be characterized along two critical dimensions. As illustrated in Fig. 1(a), in the temporal dimension, unlike explicit discrepancies (e.g., physical timing lags or raw timestamp misalignments), implicit heterogeneity manifests as **asynchronous de-**

viations in emotional semantics. As indicated by the red double-arrow connecting the semantic trajectories, even when raw physical timestamps align perfectly (e.g., $t=0s$ to $t=5s$), the underlying emotional states often evolve asynchronously. For instance, the emotional climax of a facial expression (visually "sobbing" at Node 2, $t=2s$) may substantially offset its corresponding vocal affect (audibly "choking up" at Node 4, $t=4s$). This non-linear desynchronization of emotional meaning eludes standard attention mechanisms, hindering the construction of temporally coherent representations. In the spatial dimension, distinct from explicit feature-level mismatches, implicit heterogeneity reflects the **topological inconsistency of emotional semantics**. As illustrated by red dashed lines connecting homologous states across the 3D feature spaces, heterogeneous modalities inherently organize emotional concepts into distinct latent geometric structures. For example, facial expressions form a continuous closed manifold, whereas vocal bursts form a distinct hub-and-spoke cluster anchored by a neutral state. Direct

cross-modal interaction without aligning this hidden emotional geometry introduces severe representational distortions, limiting fine-grained cross-modal integration.

To systematically address these challenges, we propose a novel framework, Rectifying Spatiotemporal Discrepancies (RSD), as illustrated in Fig. 1(b). Our framework utilizes generative text anchors as a unified high-level reference to address the issues of asynchronous deviations and spatial topological inconsistency in emotional semantics. Specifically, we first utilize a Multimodal Large Language Model (MLLM) to construct stable generative text anchors. We treat these anchors as a reliable semantic benchmark, and our framework operates across two fundamental levels corresponding to the two dimensions of implicit spatiotemporal heterogeneity. At the temporal level, we introduce the Temporal Anchor Rectification (TAR) module to resolve complex **asynchronous deviations in emotional semantics**. TAR leverages semi-relaxed optimal transport to implicitly synchronize heterogeneous audio-visual sequences with the text anchors. This process models the alignment problem as a cost-minimizing mapping problem defined by emotional semantic similarity, ensuring that asynchronous emotional cues can be aligned along a shared semantic timeline to facilitate the modeling of temporally coherent emotional states. At the spatial level, we propose the Spatial Anchor Rectification (SAR) module to address the **inconsistent topological distribution of emotional semantics**. SAR is built upon a Rank-Restricted Stochastic Alignment mechanism that leverages low-rank affine transformations together with stochastic routing. This mechanism adapts modality-specific feature geometries toward the shared semantic coordinate system, encouraging structural consistency while preserving modality-specific characteristics to improve cross-modal compatibility at the representation level. Finally, we employ a Multi-task Constrained Anchoring Fusion (MCAF) strategy to integrate these rectified representations, ensuring robust and fine-grained emotion prediction in complex, in-the-wild environments.

The main contributions are summarized as follows:

- We propose RSD, which leverages generative text anchors to rectify implicit spatiotemporal heterogeneity in audio-visual emotion recognition. RSD systematically resolves asynchronous deviations and inconsistent topological distributions in emotional semantics at both temporal and spatial levels.
- We introduce a Temporal Anchor Rectification (TAR) module based on semi-relaxed optimal transport theory. By utilizing the text anchors as benchmark, TAR implicitly synchronizes heterogeneous audio-visual sequences with the semantic timeline, resolving temporal asynchronous deviations in emotional semantics.
- We design a Spatial Anchor Rectification (SAR) module driven by a Rank-Restricted Stochastic Alignment (RRSA) mechanism. SAR aligns modality-specific topological structures with the semantic coordinate system via low-rank affine transformations and stochastic routing, ensuring consistency of topological distributions across the latent spaces of emotional semantics.
- Extensive experiments on the MAFW and AFEW benchmarks demonstrate that RSD significantly outperforms state-of-the-art methods, validating its effectiveness in resolving implicit spatiotemporal discrepancies for robust AVER.

2. Related work

2.1. Audio-visual emotion recognition

Audio-visual emotion recognition (AVER) has progressed from hand-crafted methods to modern deep-learning architectures that bridge the emotional gap through joint representation learning [12,13]. To capture the dynamic nature of affect, researchers have integrated spatiotemporal modeling and Multi-Task Learning (MTL) to optimize categorical and continuous emotion dimensions simultaneously [14,15]. Despite these advances, traditional fusion techniques often fail to fully exploit the

complementary relationships between heterogeneous modalities over time.

The emergence of Transformers and attention mechanisms has enabled models to focus on emotionally salient segments through cross-modal correlation [16,17]. Furthermore, self-supervised methods like HiCMAE [18] utilize masked modeling to learn generic representations from unlabeled data. The most recent trend involves leveraging Large Language Models (LLMs) and Vision-Language Models (VLMs) to provide descriptive semantic guidance [19–21]. However, real-world data still suffers from implicit spatiotemporal heterogeneity, characterized by asynchronous deviations in emotional semantics and inconsistent topological distributions of latent features [22]. Existing methods continue to struggle with resolving these spatiotemporal discrepancies, often leading to severe distortions in multimodal representations and hindering the precise alignment of heterogeneous features with a unified semantic coordinate system.

2.2. Cross-modal alignment

Cross-modal alignment is a fundamental prerequisite for multimodal fusion, aiming to synchronize information across heterogeneous modalities. Earlier research primarily utilized attention mechanisms to learn latent correspondences between unaligned sequences [23,24]. While flexible, these methods often fail to achieve sharp temporal synchronization, as they lack effective constraints to resolve complex asynchronous temporal dynamics. Another direction involves modality translation or mapping signals into a unified latent space to alleviate the modality gap [25,26]. However, these approaches typically focus on aligning feature vectors in terms of distance without addressing the fundamental inconsistencies in their spatial topological structures.

With the advancement of VLMs, recent works focus on achieving alignment through high-level semantic guidance and contrastive learning [27,28]. However, the increasing complexity of these models introduces risks of parametric knowledge conflicts, potentially leading to misinterpretation during alignment [29]. Despite these advancements, existing strategies often rely on coarse-grained matching, making them less effective in handling asynchronous deviations in emotional semantics where emotional events exhibit non-linear time offsets. Furthermore, most methods struggle to bridge the geometric gap caused by discordant topological distributions of emotional semantics across heterogeneous modalities, failing to achieve the spatial topological alignment required for robust affective computing. Current frameworks lack a robust mechanism integrating precise temporal synchronization and structural affine transformations to rectify underlying spatiotemporal discrepancies in the wild.

3. Method

3.1. Overview

To systematically address the implicit spatiotemporal heterogeneity of asynchronous deviations and inconsistent topological distribution in emotion semantics, we propose the RSD framework. As shown in Fig. 2, it comprises three main modules: Temporal Anchor Rectification (TAR), Spatial Anchor Rectification (SAR), and Multi-task Constrained Anchoring Fusion (MCAF). Specifically, to address the implicit temporal asynchronous deviation in emotion semantics, TAR generates stable text anchors as high-level references using MLLM, and then implicitly synchronizes heterogeneous audiovisual sequences along the semantic timeline using semi-relaxed optimal transport theory. Although TAR effectively resolves the temporal asynchronous deviations, the spatial topological inconsistencies across heterogeneous modalities still remain. To address this issue, SAR employs a rank-restricted stochastic alignment (RRSA) mechanism, utilizing low-rank affine transformations and stochastic routing to align the topological structure of specific modalities with a shared semantic coordinate system. Finally, to ensure robust

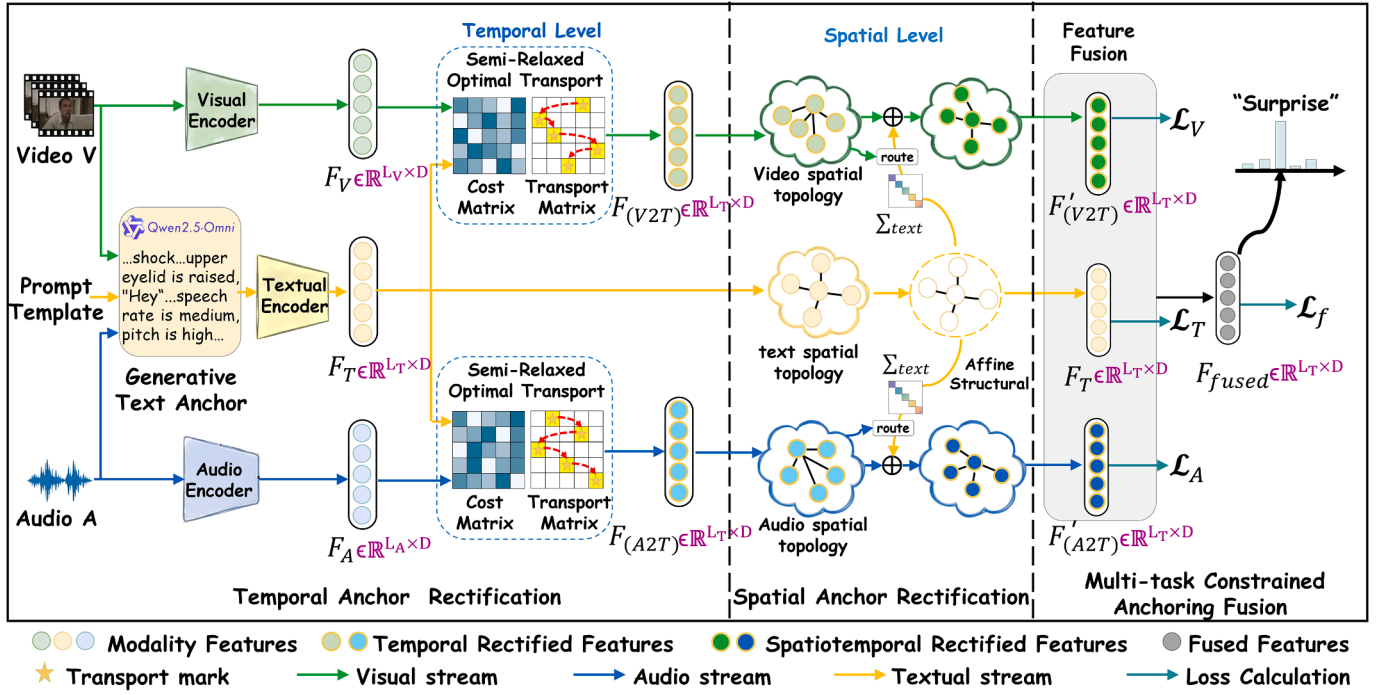


Fig. 2. Overview of RSD framework. Temporal Anchor Rectification (TAR) generates stable text anchors via an MLLM to implicitly synchronize heterogeneous audio-visual sequences along the semantic timeline using semi-relaxed optimal transport theory. Spatial Anchor Rectification (SAR) employs a Rank-Restricted Stochastic Alignment (RRSA) mechanism to align modality-specific topological structures with the semantic coordinate system through low-rank affine transformations and stochastic routing. Multi-task Constrained Anchoring Fusion (MCAF) integrates rectified representations to ensure robust emotion prediction under multi-task constraints. The stochastic switch *route* follows a Bernoulli distribution during training to prevent over-reliance on text geometry, while functioning deterministically as an identity residual connection during inference to guarantee stable predictions.

Table 1
Summary of key notation.

Symbol	Dimension	Description
F_T	$\mathbb{R}^{L_T \times D}$	Text anchor features
F_V, F_A	$\mathbb{R}^{L_V/L_A \times D}$	Video / Audio features
$F_{(V2T)}, F_{(A2T)}$	$\mathbb{R}^{L_T \times D}$	Synchronized features after TAR
$F'_{(V2T)}, F'_{(A2T)}$	$\mathbb{R}^{L_T \times D}$	Rectified features after SAR
F_{fused}	$\mathbb{R}^{L_T \times D}$	Fused features after MCAF
C_V, C_A	$\mathbb{R}^{L_T \times L_V/L_A}$	Semantic cost matrices
π_V, π_A	—	Deterministic Monge mappings
Γ	$\mathbb{R}^{L_T \times L_m}$	Transport coupling matrix
Σ_{text}	$\mathbb{R}^{D \times D}$	Diagonal scaling operator
W_{down}, W_{up}	$\mathbb{R}^{b \times D}, \mathbb{R}^{D \times b}$	Low-rank bottleneck projections ($b \ll D$)
$\mathbb{I}_{route}$	—	Stochastic routing switch
β, p_{drop}	—	Learnable scale / Dropout probability
$\mathcal{L}_{total}$	—	Total training objective
$\mathcal{L}_f$	—	Primary classification loss
$\mathcal{L}_V, \mathcal{L}_A, \mathcal{L}_T$	—	Multi-task auxiliary losses
D, b	—	Feature dimension / Bottleneck dimension
λ, α	—	Loss weights for multi-task constraints

and reliable emotion prediction, MCAF integrates these synchronized and corrected representations under multi-task learning constraints. In the following sections, we elaborate on these three components in detail. For easy reference, we have summarized the key mathematical symbols used in this section in the Table 1.

3.2. Temporal anchor rectification (TAR)

To address the implicit asynchronous deviation in emotion semantics between heterogeneous modalities, we propose the TAR module, which comprises two parts: **Emotion Semantic Text Anchor Generation and Sequence Synchronization via Semi-Relaxed Optimal Transport**. Specifically, the emotion semantic text anchor generation constructs sta-

ble generative text anchors, providing a robust emotion semantic timeline for subsequent sequence synchronization based on optimal transport. Detailed implementation specifications will be discussed below.

3.2.1. Emotion semantic text anchor generation

To provide a robust text anchor, we use MLLM to generate structured emotional text descriptions from raw audio and video. Specifically, considering that human emotion is jointly manifested by global affective tendency, facial physiological cues, verbal semantic content, and paralinguistic acoustic characteristics, we designed a comprehensive prompt template $\mathbf{P}_{template}$ guiding the model toward comprehensive, structured emotional perception from four complementary dimensions. $\mathbf{P}_{env}$ offers a global overview by summarizing the overall emotional tendency, laying a high-level emotional prior. $\mathbf{P}_{fau}$ focuses on the visual modality, requiring objective, interpretable facial muscle movement descriptions (i.e., Facial Action Units, AU) to establish observable visual physiological representations. $\mathbf{P}_{cont}$ targets the auditory modality, extracting phonetic or non-phonetic vocal content as the linguistic carrier of emotion. $\mathbf{P}_{acou}$ mines paralinguistic features (prosody, pitch, intensity variations) to supplement emotion's dynamic acoustic expression. The complete template aggregates these four components:

$$\mathbf{P}_{template} = \{\mathbf{P}_{env}, \mathbf{P}_{fau}, \mathbf{P}_{cont}, \mathbf{P}_{acou}\}. \quad (1)$$

Next, given the video V and audio A , the emotional semantic text description is generated as:

$$\mathbf{T} = \Phi(V, A, \mathbf{P}_{template}), \quad (2)$$

where $\mathbf{T}$ is the resulting emotion text description, and Φ denotes the generative function of the MLLM. The MLLM converts heterogeneous multimodal information into emotion-aware semantic descriptions, providing a stable emotional semantic reference for the entire framework. To guarantee deterministic reproducibility during text generation, we employ the locally deployed Qwen2.5-Omni-7B model with bf16at16

precision. The generation parameters are configured with a maximum new token length of 1024, `do_sample` set to false, and temperature set to 0.0. The full prompt template $P_{template}$ guiding the multimodal large language model is provided in Appendix A.

Since MLLM focuses on high-level semantic generation rather than fine-grained unimodal encoding, we adopt separate dedicated encoders for feature extraction. Thus, after obtaining the generated emotion text description T , we take it and the original audio-video as input, and use a dedicated pre-trained encoder to extract features for each modality. Specifically, we use ViT [30] as a visual encoder to process the original video frames V , Data2Vec [31] as an audio encoder to process the audio waveform A , and DistilBERT [32] as a text encoder to process the generated text anchors T . This process generates an initial unimodal feature sequence:

$$\mathbf{F}_V = \text{ViT}(V), \quad \mathbf{F}_A = \text{Data2Vec}(A), \quad \mathbf{F}_T = \text{DistilBERT}(T). \quad (3)$$

It is worth noting that the generated text descriptions may contain emotion-related adjectives as part of natural linguistic expressions. However, the MLLM prompt is strictly blind to the dataset's ground-truth label space. These outputs are free-form descriptive texts rather than categorical labels, and are never directly provided as category names to the classifier. In contrast to direct MLLM-based zero-shot classification, our framework utilizes the generated text exclusively as semantic anchors for spatiotemporal alignment, while the final prediction is determined by the fused multimodal representation. Consequently, the text anchors function purely as semantic guidance, intrinsically avoiding potential label leakage.

These extracted features ($\mathbf{F}_V, \mathbf{F}_A, \mathbf{F}_T$) are used as normalized input for subsequent spatiotemporal correction.

3.2.2. Sequence synchronization via semi-relaxed optimal transport

To address the core issue of implicit asynchronous bias in emotion semantics, we model the alignment task as a semi-relaxed discrete optimal transport problem. By utilizing $\mathbf{F}_T$ as a reference for the emotion semantic timeline, asynchronous audio and video are implicitly synchronized to a shared semantic timeline. Taking video modality as an example, the synchronization process includes **Semantic Cost Calculation** and **Deterministic Monge Mapping via Semi-Relaxed Optimal Transport**. Semantic cost calculation quantifies the similarity between video features and text anchors $\mathbf{F}_T$, providing a basis for cross-modal alignment. Based on the constructed cost matrix, we solve a semi-relaxed optimal transport problem to derive a deterministic Monge mapping between video frames and the semantic timeline, establishing a correspondence that enables synchronization of asynchronous emotional semantics. The implementation is as follows:

Semantic Cost Computation. To measure the semantic discrepancy between the text anchor and the video modality, we treat the text anchor $\mathbf{F}_T \in \mathbb{R}^{L_T \times D}$ and the video $\mathbf{F}_V \in \mathbb{R}^{L_V \times D}$ as discrete distributions. Based on these representations, we construct a pairwise semantic cost matrix $\mathbf{C}_V \in \mathbb{R}^{L_T \times L_V}$, which corresponds to lines 6 to 9 in Algorithm 1. Each element $\mathbf{C}_{V,(j,k)}$ quantifies the transport cost between the j -th text token and the k -th video frame. Specifically, we utilize the cosine distance as the metric:

$$\mathbf{C}_{V,(j,k)} = 1 - \frac{\mathbf{F}_T(j) \cdot \mathbf{F}_V(k)}{\|\mathbf{F}_T(j)\| \|\mathbf{F}_V(k)\|}, \quad (4)$$

where $j \in \{1, \dots, L_T\}$ and $k \in \{1, \dots, L_V\}$. A lower cost indicates a higher semantic correlation between the anchor and the modality frame.

Deterministic Monge Mapping via Semi-Relaxed Optimal Transport. Based on the semantic cost matrix $\mathbf{C}_{V,(j,k)}$ constructed above, to accurately align the asynchronous video features to the shared text anchors, we formulate the alignment as a semi-relaxed optimal transport problem. Let the empirical distributions of text anchors and video features be defined as:

$$\mu_T = \frac{1}{L_T} \sum_{j=1}^{L_T} \gamma_{\mathbf{F}_T(j)}, \quad \nu_V = \frac{1}{L_V} \sum_{k=1}^{L_V} \gamma_{\mathbf{F}_V(k)}, \quad (5)$$

where γ is the Dirac delta measure. The standard optimal transport seeks a coupling matrix $\Gamma \in \mathbb{R}_+^{L_T \times L_V}$ minimizing the total transport cost $\langle \mathbf{C}_V, \Gamma \rangle_F$ subject to the marginal constraints $\Gamma \mathbf{1} = \mu_T$ and $\Gamma^\top \mathbf{1} = \nu_V$, where $\mathbf{1}$ denotes the all-one vector. Specifically, $\Gamma \mathbf{1} = \mu_T$ enforces that the total mass transported from each text token sums to its empirical probability, while $\Gamma^\top \mathbf{1} = \nu_V$ ensures the same for each video frame. However, strictly enforcing both marginals is overly restrictive for emotion sequence alignment, as the temporal density of emotional events may differ across modalities. We therefore adopt a semi-relaxed formulation that relaxes the video frame marginal constraint (i.e., $\Gamma^\top \mathbf{1} = \nu_V$ is relaxed) while preserving the text anchor marginal constraint $\Gamma \mathbf{1} = \mu_T$. This relaxation allows each text semantic token to flexibly select its most relevant source frame without requiring every source frame to be consumed exactly once. In this way, multiple consecutive text tokens can anchor onto the same salient visual frame to establish a many to one correspondence. This formulation faithfully models the natural temporal persistence of emotional expressions such as a sustained smile or frown while the text sequentially unfolds fine-grained descriptive attributes. At the same time, emotionally uninformative, blurred, or neutral transitional frames are adaptively skipped rather than forcing the model to absorb redundant temporal noise. Under this formulation, because the column-wise coupling constraints are relaxed, the linear optimization decouples independently over the rows of the cost matrix $\mathbf{C}_V$ subject only to the preserved mass constraint $\sum_{k=1}^{L_V} \Gamma_{j,k} = 1/L_T$. To minimize this linear combination of costs in each row, the optimal plan assigns the entire available mass of $1/L_T$ to the column with the minimum cost. Consequently, the optimization reduces to a row-wise minimization, yielding a deterministic Monge mapping as formalized in lines 10 to 12 of Algorithm 1:

$$\pi(j) = \arg \min_{k \in \{1, \dots, L_V\}} \mathbf{C}_{V,(j,k)}, \quad \forall j \in \{1, \dots, L_T\}, \quad (6)$$

where π is a deterministic assignment that maps each text token position j to exactly one video frame k . This corresponds to a hard assignment $\Gamma_{j,k} = \mathbf{1}[\pi(j) = k]$, which is computationally efficient, semantically interpretable, and preferable to soft probabilistic alternatives. Soft transport plans compute convex combinations across multiple frames and inevitably dilute sharp localized emotional climaxes such as subtle micro-expressions or action unit bursts with surrounding neutral frames. In contrast, our hard assignment strictly preserves the intact feature representation and peak discriminability of salient frames. This index selection effectively anchors the relevant visual frames onto the unified text anchor. Based on these optimal assignments, we perform temporal rectification by re-organizing the source video features to strictly follow the emotion semantic progression of the text anchor. The synchronized video sequence $\mathbf{F}_{(V2T)}$ is reconstructed as presented in lines 13 to 15 of Algorithm 1:

$$\mathbf{F}_{(V2T)}(j) = \mathbf{F}_V(\pi(j)), \quad j \in \{1, \dots, L_T\}. \quad (7)$$

This reconstruction operation aligns the original video sequence to the temporal coordinate system defined by the text anchor. The original video frames are re-ordered and re-sampled according to the transport cost. Specifically, for each text token position j , the algorithm selects the video frame k that minimizes the transport cost $\mathbf{C}_{V,(j,k)}$ and places it at position j in the synchronized sequence. This process is repeated for all $j \in \{1, \dots, L_T\}$, generating a new sequence in which the order of video frames is determined by their semantic correspondence to the text anchor rather than by their original temporal indices. In this way, video frames that express similar emotional semantics are aligned to the same positions along the text-defined emotion semantic timeline, regardless of when they occurred in the original recording. Unlike monotonic sequence alignment methods such as Dynamic Time Warping that strictly enforce chronological progression, this unconstrained Monge mapping provides the semantic anchor with the topological freedom to resolve severe non-linear temporal offsets and asynchronous delays across modalities. The audio modality undergoes an identical process to generate

$\mathbf{F}_{(A2T)}$ with its own mapping $\pi(A)$. The complete procedure is summarized in [Algorithm 1](#).

Algorithm 1: Complete training process of RSD.

```

Input : Raw Video:  $V$ ; Raw Audio:  $A$ ; Prompt Template:
          $\mathbf{P}_{template}$ ; Labels:  $y$ ; Hyperparameters:  $\lambda, \alpha, p_{drop} = 0.3$ 
Output: Rectified Features:  $\mathbf{F}'_{(A2T)}, \mathbf{F}'_{(V2T)} \in \mathbb{R}^{L_T \times D}$ ; Anchor
         Features:  $\mathbf{F}_T \in \mathbb{R}^{L_T \times D}$ ; Fused Features:  $\mathbf{F}_{fused} \in \mathbb{R}^{L_T \times D}$ 

// Stage 1: Text Anchor Generation & Modality
// Feature Extraction
1  $\mathbf{T} \leftarrow \Phi(V, A, \mathbf{P}_{template})$ ;
2  $\mathbf{F}_T \leftarrow \text{Encoder}_T(\mathbf{T})$ ; // Anchor Features ( $L_T \times D$ )
3  $\mathbf{F}_V \leftarrow \text{Encoder}_V(V)$ ; // Source Features ( $L_V \times D$ )
4  $\mathbf{F}_A \leftarrow \text{Encoder}_A(A)$ ; // Source Features ( $L_A \times D$ )

// Stage 2: TAR - Temporal Synchronization via
// Semi-Relaxed OT
5 for Modality  $m \in \{V, A\}$  do
6   Initialize Cost Matrix  $\mathbf{C}_m \in \mathbb{R}^{L_T \times L_m}$ ;
7   for  $j \leftarrow 1$  to  $L_T$  do
8     for  $k \leftarrow 1$  to  $L_m$  do
9        $\mathbf{C}_{m,(j,k)} \leftarrow 1 - \text{CosineSimilarity}(\mathbf{F}_T(j), \mathbf{F}_m(k))$ ;
10  Initialize Mapping Indices  $\pi_m \in \mathbb{N}^{L_T}$ ;
11  for  $j \leftarrow 1$  to  $L_T$  do
12     $\pi_m(j) \leftarrow \arg \min_k \mathbf{C}_{m,(j,k)}$ ;
13  Initialize Output  $\mathbf{F}_{(m2T)}$ ;
14  for  $j \leftarrow 1$  to  $L_T$  do
15     $\mathbf{F}_{(m2T)}(j) \leftarrow \mathbf{F}_m(\pi_m(j))$ ;

// Stage 3: SAR - Spatial Alignment via RRSA
16  $\mathbf{g} \leftarrow \text{Mean}(\mathbf{F}_T, \text{dim} = 1)$ ; // Global text representation
17  $\mathbf{h} \leftarrow \text{GELU}(\text{LayerNorm}(W_{down} \cdot \mathbf{g}))$ ; // Low-rank bottleneck
18  $\Sigma_{text} \leftarrow \text{diag}(\text{Sigmoid}(W_{up} \cdot \mathbf{h}))$ ; // Diagonal scaling
   operator
19 for Modality  $m \in \{V, A\}$  do
20   if training then
21     Sample  $\mathbb{I}_{route} \sim \text{Bernoulli}(1 - p_{drop})$ ;
22     if  $\mathbb{I}_{route} = 0$  then
23        $\mathbf{F}'_{(m2T)} \leftarrow \mathbf{F}_{(m2T)}$ ; // Skip alignment
24     else
25        $\mathbf{F}'_{(m2T)} \leftarrow \mathbf{F}_{(m2T)} + \mathbb{I}_{route} \cdot \beta \cdot (\mathbf{F}_{(m2T)} \cdot \Sigma_{text})$ ;
26   else
27      $\mathbf{F}'_{(m2T)} \leftarrow \mathbf{F}_{(m2T)} + \beta \cdot (\mathbf{F}_{(m2T)} \cdot \Sigma_{text})$ ;
   // Deterministic

// Stage 4: MCAF - Multi-task Constrained Anchoring
// Fusion
28  $\mathbf{F}_{fused} \leftarrow \text{AttentionFusion}(\mathbf{F}_T, \mathbf{F}'_{(A2T)}, \mathbf{F}'_{(V2T)})$ ;

// Stage 5: Multi-task Loss Computation &
// Backpropagation
29  $\mathcal{L}_{total} \leftarrow \mathcal{L}_{cls}(\mathbf{F}_{fused}, y) + \lambda(\mathcal{L}_{cls}(\mathbf{F}'_{(V2T)}, y) + \mathcal{L}_{cls}(\mathbf{F}'_{(A2T)}, y)) + \alpha \cdot$ 
    $\mathcal{L}_{cls}(\mathbf{F}_T, y)$ ;
30 return  $\mathbf{F}'_{(A2T)}, \mathbf{F}'_{(V2T)}, \mathbf{F}_T, \mathbf{F}_{fused}$ 

```

3.3. Spatial anchor rectification (SAR)

Although TAR achieves temporal synchronization by aligning features along the emotion semantic timeline, it does not modify the underlying feature geometry. As a result, heterogeneous modalities may still exhibit spatial topological inconsistencies in embedding spaces, which hinders cross-modal interaction. To address this issue, we propose SAR,

which aligns modality-specific geometries to a unified text anchor coordinate system via rank-restricted stochastic alignment. Specifically, SAR consists of three components: **Affine Structural Extraction**, **Spatial Rectification via Anisotropic Scaling**, and **Modality-Specific Preservation**. Affine Structural Extraction first constructs a shared topological reference based on text anchors. Based on this reference, Spatial Rectification performs geometry alignment via anisotropic scaling to address inconsistent spatial topology. Finally, Modality-Specific Preservation regularizes the transformation process to mitigate feature degradation and preserve intrinsic modality characteristics, achieving a balance between structural alignment and modality fidelity. The implementation is as follows:

Affine Structural Extraction. To establish a consistent geometric reference and ensure dimensional consistency, we first obtain a global semantic vector by averaging the text sequence across tokens, defined as $\bar{\mathbf{f}}_T = \frac{1}{L_T} \sum_{j=1}^{L_T} \mathbf{F}_T(j) \in \mathbb{R}^D$. We then extract the underlying structural geometry by projecting this global text vector $\bar{\mathbf{f}}_T$ into a low-dimensional rank-restricted subspace for compression, extracting the main semantic axes and generating a compact low-rank representation. Subsequently, we project this low-rank representation back to the original high-dimensional space to generate the diagonal scaling operator $\Sigma_{text} \in \mathbb{R}^{D \times D}$. Specifically, $W_{down} \in \mathbb{R}^{b \times D}$ compresses the global text vector into a b -dimensional bottleneck space with $b \ll D$, and $W_{up} \in \mathbb{R}^{D \times b}$ reconstructs it back. The rank of the overall transformation is inherently bounded by b , meaning that $\text{rank}(W_{up}W_{down}) \leq b \ll D$. This bottleneck structure forces the network to discard noise and irrelevant variations, preserving only the most salient structural patterns of the text geometry. The diagonal structure of Σ_{text} further restricts the transformation to dimension-wise scaling rather than a full matrix transformation, implicitly regularizing the complexity of geometric adjustment. This encoding-decoding bottleneck structurally enforces a low-rank constraint without explicit regularization, enabling the operator to robustly encode the stable spatial geometry of the text modality. The extraction process is formalized in lines 16 to 18 of [Algorithm 1](#) as follows:

$$\hat{\mathbf{f}}_T = \delta(W_{down}\bar{\mathbf{f}}_T), \quad \Sigma_{text} = \text{diag}(\sigma(W_{up}\hat{\mathbf{f}}_T)). \quad (8)$$

where $\hat{\mathbf{f}}_T \in \mathbb{R}^b$ represents the low-rank semantic vector capturing the core geometry, δ and σ denote the activation functions, and W_{down} and W_{up} denote the projection matrices for dimension reduction and expansion, respectively.

Spatial Rectification via Anisotropic Scaling. To address inconsistent spatial topology, we achieve spatial rectification through a rank-restricted anisotropic scaling strategy, utilizing the diagonal scaling operator Σ_{text} to selectively calibrate the spatial coordinates of the audio and visual feature spaces. Specifically, the diagonal operator assigns different scaling weights to each feature dimension, with larger weights assigned to dimensions that contribute more to emotion discrimination and smaller weights to those that contribute less. When $\mathbf{F}_{(m2T)}$ is scaled by Σ_{text} , each feature dimension is independently multiplied by its corresponding scaling weight, amplifying dimensions with larger weights and suppressing those with smaller weights, effectively reorienting the feature vectors along the principal axes of the audio-visual feature space. Since a feature vector's orientation in high-dimensional space is determined by the relative magnitudes across its dimensions, this anisotropic scaling alters both the length and direction of each feature vector, reshaping the distribution of feature points in the latent space. Geometrically, this anisotropic scaling rescales the individual coordinate axes of the latent space, dynamically stretching or compressing the principal axes of the audio-visual feature distributions. This transformation alters the eccentricity and covariance orientations of the modality-specific feature clusters, aligning their inter-feature metric distances and directional correlations with the principal semantic axes of the text space. By applying the matrix multiplication $\mathbf{F}_{(m2T)} \cdot \Sigma_{text}$ where $\mathbf{F}_{(m2T)} \in \mathbb{R}^{L_T \times D}$ and $\Sigma_{text} \in \mathbb{R}^{D \times D}$, the spatial distribution of audio-visual modalities is geometrically harmonized into the unified text anchor coordinate system. In this way, the geometry of heterogeneous modalities is structurally

assimilated into a stable text anchor coordinate system. However, this strict alignment introduces a new challenge, as it may indiscriminately mask inherent modality-specific nuances with the dominant emotion semantic text geometry.

Modality-Specific Preservation. To compensate for the loss of fine-grained modality cues that may be caused by strict spatial alignment, we introduce a stochastic routing mechanism. This mechanism integrates the original modality representations with their corresponding affine-transformed components through a stochastically gated residual connection. The final rectified features $\mathbf{F}'_{(m2T)}$ during training are expressed in lines 20 to 25 of [Algorithm 1](#) as follows:

$$\mathbf{F}'_{(m2T)} = \mathbf{F}_{(m2T)} + \mathbb{I}_{route} \cdot \beta \cdot (\mathbf{F}_{(m2T)} \cdot \Sigma_{text}), \quad (9)$$

where β is a learnable scaling parameter initialized to 0.001, and the stochastic switch $\mathbb{I}_{route}$ is sampled from a Bernoulli distribution with dropout probability p_{drop} , i.e., $\mathbb{I}_{route} \sim \text{Bernoulli}(1 - p_{drop})$. During training, we set $p_{drop} = 0.3$, randomly skipping the spatial alignment with 30% probability to prevent over-reliance on the text anchor and enhance robustness against potential hallucinations. At inference time, the stochasticity is disabled by fixing $\mathbb{I}_{route} = 1$, and the alignment is applied deterministically as formalized in lines 26 to 27 of [Algorithm 1](#):

$$\mathbf{F}'_{(m2T)} = \mathbf{F}_{(m2T)} + \beta \cdot (\mathbf{F}_{(m2T)} \cdot \Sigma_{text}). \quad (10)$$

In addition, we conduct a sensitivity analysis of p_{drop} in our experiments to validate the robustness of this hyperparameter. Ultimately, this balancing strategy provides geometrically aligned and semantically diverse representations for the subsequent multi-task constrained anchoring fusion, thereby achieving more robust emotion prediction.

3.4. Multi-task constrained anchoring fusion (MCAF)

To obtain comprehensive and robust cross-modal representations for the final emotion classification, we employ the MCAF strategy. We integrate the spatiotemporally rectified features using an asymmetric, anchor-guided attention mechanism. Specifically, we project the refined text anchor $\mathbf{F}_T$, the rectified audio features $\mathbf{F}'_{(A2T)}$, and the rectified video features $\mathbf{F}'_{(V2T)}$ into a common subspace to serve as Query (Q), Key (K), and Value (V), respectively. The fused representation is computed via standard scaled dot-product attention as:

$$\mathbf{F}_{fused} = \text{Linear} \left(\text{Softmax} \left(\frac{\mathbf{QK}^T}{\sqrt{d_k}} \right) \mathbf{V} \right), \quad (11)$$

where d_k is the scaling dimension. This process allows generative text anchor to query acoustic variations and aggregate emotionally salient visual expressions.

However, while the generative text anchor provides powerful guidance for this fusion process, potential semantic deviations or hallucinated noise in the MLLM output could mislead the final prediction. Therefore, to mitigate this risk and prevent negative transfer, the core of MCAF relies on a joint optimization paradigm with hierarchical multi-task constraints. We apply independent supervisory constraints to both the final fused representation and the modality-specific features to ensure the overall robustness of the system. The total training objective $\mathcal{L}_{total}$ is formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_f + \lambda(\mathcal{L}_V + \mathcal{L}_A) + \alpha\mathcal{L}_T, \quad (12)$$

where $\mathcal{L}_f$ supervises the primary multimodal fused output $\mathbf{F}_{fused}$. Crucially, we introduce modality-specific losses ($\mathcal{L}_V$ and $\mathcal{L}_A$), governed by the weight λ , to force the upstream recalibration and rectification stages to preserve strong, independent discriminative power, ensuring the model does not over-rely on the text anchor. Meanwhile, $\mathcal{L}_T$ (weighted by α) is applied to constrain the text anchor itself, minimizing the impact of potentially unreliable information generated by the MLLM. Specifically, $\mathcal{L}_T$ acts as an auxiliary classification loss supervised directly by emotion labels, which naturally filters out MLLM hallucinations by forcing the encoder to extract only emotion-relevant cues.

Through these multi-task constraints, the framework guarantees robust decision-making and maintains high recognition precision even when facing significant semantic biases in complex, in-the-wild environments.

4. Experiments and analysis

4.1. Datasets

MAFW [33]: This dataset is a challenging real-life multi-modal dataset covering audio, vision, and text modalities. It contains 10,045 video-audio clips annotated with compound emotions, which span 11 common categories including anger, disgust, fear, happiness, neutral, sadness, surprise, contempt, anxiety, helplessness, and disappointment. Each clip is annotated 11 times by professional annotators, and the unreliable annotations are filtered out using the Expectation Maximization algorithm. To ensure reliable evaluation, consistent with standard practices, we employ 5-fold cross-validation on the MAFW dataset for all experiments.

AFEW [34]: This benchmark dataset comprises 1809 video clips extracted from movies and TV series, covering seven basic emotions (happiness, sadness, neutral, anger, surprise, disgust, and fear). The dataset provides an official split into training, validation, and test sets. In line with prior works [21], we evaluate our method on the official validation set for standard comparison. Furthermore, to ensure a more robust and comprehensive assessment given the dataset's limited scale, we combined the training and validation splits to conduct a supplementary 5-fold cross-validation. This rigorous protocol is consistently applied to our method, the re-implemented Baseline, and other re-evaluated models to guarantee fair comparisons under identical conditions.

4.2. Implementation details

Data Processing: For each video clip, we uniformly sample 8 frames and resize all images to 224×224 pixels. For audio processing, we extract waveforms at a sampling rate of 16 kHz, with audio length set to 16,000 samples per clip. All input features are normalized using modality-specific statistics.

Text Anchor Generation: The MLLM-generated text anchors are generated offline prior to model training using Qwen2.5-Omni-7B [35], with the structured prompt template detailed in [Section 3.2.1](#). The generated descriptions are stored locally and loaded as fixed text anchors during both training and inference stages. This offline preprocessing avoids the significant GPU memory overhead that would arise from simultaneously loading the MLLM and the main RSD model during training, and also ensures consistent text anchors across multiple experimental runs, further enhancing reproducibility.

Training and Inference Settings: We train our model for 25 epochs with a batch size of 4 on a single NVIDIA RTX 4090 (24GB) GPU. We adopt the AdamW optimizer with an initial learning rate of 1×10^{-5} , following a cosine annealing schedule with a warmup ratio of 0.1. The weight decay is set to 0.005. The loss weights for multi-task constraints are set to $\lambda = 2.0$ and $\alpha = 1.0$, as validated in our ablation study ([Section 4.4.5](#)). During inference, the stochastic routing mechanism in SAR ([Section 3.3](#)) is disabled by fixing $\mathbb{I}_{route} = 1$, ensuring deterministic and reproducible predictions. All experiments are conducted with a fixed random seed of 42 to ensure reproducibility. The PyTorch version is 1.12.1 with CUDA 12.2, and the code is built on Python 3.10.

To assess the practical feasibility of RSD, we report the resource consumption of RSD across its three stages in [Table 2](#). The offline MLLM generation requires 19.08 GB of peak GPU memory and 115 minutes to process all 10,045 videos (achieving a throughput of 1.45 samples/s). Because the semantic anchors are pre-computed, the subsequent RSD model training and inference stages are highly lightweight, requiring only 6.59 GB and 1.14 GB of GPU memory, with throughputs of 12.7 samples/s and 13.1 samples/s, respectively.

Table 2
Resource consumption of the RSD framework.

Stage	Total Time (min)	Peak GPU Memory (GB)	Throughput (samples/s)
MLLM Text Generation	115	19.08	1.45
Model Training	265	6.59	12.7
Model Inference	2.54	1.14	13.1

Table 3

Comparison with SOTA methods on MAFW (11-class). Ang: anger. Dis: disgust. Fea: fear. Hap: happiness. Neu: neutral. Sad: sadness. Sur: surprise. Con: contempt. Anx: anxiety. Hel: helplessness. Disa: disappointment. UAR: unweighted average recall. WAR: weighted average recall. The best results are in bold, the second-best results are underlined.

Method	Accuracy of Each Emotion (%)											WAR (%)	UAR (%)
	Ang	Dis	Fea	Hap	Neu	Sad	Sur	Con	Anx	Hel	Disa		
C3D+LSTM [33]	62.47	3.17	15.74	77.30	42.20	65.30	42.67	0.00	19.14	0.00	0.00	44.15	30.47
AMH [41]	51.73	18.68	28.13	<u>79.14</u>	52.55	52.26	46.29	0.26	29.62	1.74	2.39	48.83	32.98
T-MEP [21]	57.04	24.85	36.09	78.96	50.83	61.85	51.28	1.29	38.47	6.46	1.70	51.15	37.17
HiCMAE [18]	69.24	29.73	34.72	78.32	<u>59.15</u>	77.69	<u>60.65</u>	6.78	<u>31.11</u>	8.08	13.74	56.17	42.65
MMA-DFER [36]	86.31	76.12	38.05	59.28	12.74	66.32	43.51	51.68	6.36	35.37	11.07	58.52	44.25
Qwen2.5-Omni [35]	55.19	6.22	54.68	77.72	72.90	76.95	45.70	51.93	13.83	28.35	<u>14.29</u>	53.55	45.25
Qwen3 [38]	57.63	6.11	59.36	87.83	30.37	90.34	68.56	17.80	10.94	1.53	15.38	53.27	40.53
Gemma3 [39]	46.12	5.95	37.76	53.42	100.00	<u>85.29</u>	59.29	23.73	11.71	4.60	2.21	52.18	39.10
VAEmo [37]	-	-	-	-	-	-	-	-	-	-	-	58.91	<u>45.67</u>
Baseline [40]	<u>92.15</u>	<u>80.63</u>	<u>65.05</u>	72.44	0.00	61.36	6.94	<u>52.18</u>	0.00	43.64	0.00	60.04	43.16
RSD (Ours)	94.09	82.22	70.16	75.27	2.23	71.21	17.03	56.87	2.21	60.31	7.66	65.95	49.02

4.3. Overall performance

To ensure fair comparisons, all compared methods are evaluated under the same evaluation protocol and data splits, with identical input-frame sampling strategy. For MAFW, we adopt 5-fold cross-validation. For AFEW, we follow prior works by using the official validation set, while additionally adopting 5-fold cross-validation for fair comparison among re-implemented methods. For MLLM-based baselines (Qwen2.5-Omni, Qwen3, Gemma3), we adopt zero-shot inference with a structured prompt adapted to 11 classes for MAFW and 7 classes for AFEW. For our RSD and the re-implemented Baseline, we report results as mean $\pm$ standard deviation over 5 folds. For other compared methods, we report results from our re-implementation or their original papers.

4.3.1. Evaluation on MAFW datasets

As shown in Table 3, we evaluate the performance of our proposed RSD framework on the MAFW dataset. To ensure a rigorous and fair comparison, we contrast RSD with several existing SOTA methods, including recent traditional architectures like MMA-DFER [36] and VAEmo [37], as well as advanced MLLMs such as Qwen2.5-Omni [35], Qwen3 [38], and Gemma3 [39]. Furthermore, we implement a strong Baseline adapted from the recent ETG-AVER [40] framework. To ensure fairness, we upgrade the Baseline's emotion-aware text generation module with the exact same MLLM used in our approach. The Baseline shares the same unimodal encoders, MLLM-based text generation, and MCAF fusion mechanism as RSD. However, it differs from RSD in two critical aspects: it lacks TAR for temporal synchronization and SAR for spatial topological alignment.

On the MAFW dataset, RSD achieves state-of-the-art performance with 65.95% WAR and 49.02% UAR, surpassing the advanced VAEmo [37] method by substantial margins of 7.04% and 3.35%, respectively. Notably, our framework also significantly outperforms directly utilizing recent powerful MLLMs (e.g., Qwen3 [38] and Gemma3 [39]), demonstrating its superior adaptation capability for multimodal emotion recognition. Furthermore, RSD substantially surpasses the strong Baseline (+5.91% in WAR and +5.86% in UAR). To assess the statistical reliability of these gains, we conduct paired t-tests over the 5 folds, yielding p-values of 0.0065 for WAR and 0.00008 for UAR, both well below the standard 0.05 threshold, demonstrating the statistical significance of the improvements. These results indicate that merely using MLLM-generated emotion semantic text for standard fusion is insuffi-

cient. By employing generative text anchors as unified references to perform implicit temporal synchronization and spatial topological rectification, RSD effectively captures more discriminative emotional features, ensuring robust decision-making in diverse and complex scenarios.

To further analyze the model's discriminative power, we compare the confusion matrices of the Baseline and RSD presented in Fig. 3 on the MAFW dataset. The comparison reveals that RSD provides a significant accuracy improvement in distinguishing challenging and complex categories. Specifically, the recognition accuracy for "Sadness" improves from 61% to 71%, and the accuracy for "Helplessness" increases from 44% to 60%. By resolving implicit synchronous deviations and rectifying inconsistent spatial topologies before cross-modal interaction, RSD successfully mitigates severe misclassifications. For instance, the Baseline mistakenly predicts 18% of "Helplessness" samples as "Disgust", whereas RSD drastically reduces this confusion to merely 5%. Similarly, the misclassification rate of "Disappointment" being mistaken for "Fear" drops from 35% to 25%. While certain challenges persist for the "Surprise" category, which is frequently confused with "Helplessness" at a rate of 30% and "Fear" at 19% due to highly dynamic and abrupt facial variations, the overall results remain compelling. These findings confirm that the proposed RSD framework effectively resolves spatiotemporal discrepancies by implicitly synchronizing deviations and rectifying discordant spatial topologies prior to cross-modal fusion.

4.3.2. Evaluation on AFEW datasets

As shown in Table 4, we further evaluate the performance of RSD on the AFEW dataset. The upper section reports results under the same evaluation protocol as existing methods, where RSD achieves 63.64% WAR and 61.99% UAR, outperforming the competitive PFFEI method by 0.46% in WAR and 1.78% in UAR. The lower section reports results under 5-fold cross-validation, where RSD achieves 66.52% WAR and 60.80% UAR, surpassing the Baseline by 7.66% and 7.52%, respectively. Paired t-tests over the 5 folds confirm the statistical significance of these gains, with p-values of 0.0426 for WAR and 0.0432 for UAR, both below 0.05.

To further validate the fairness of our comparisons, Table 5 reports the mean and standard deviation of all methods that we re-implemented under the 5-fold cross-validation protocol on both datasets. RSD consistently achieves the highest WAR and UAR across both MAFW and AFEW, with competitive standard deviations, confirming the robustness and statistical reliability of our method.

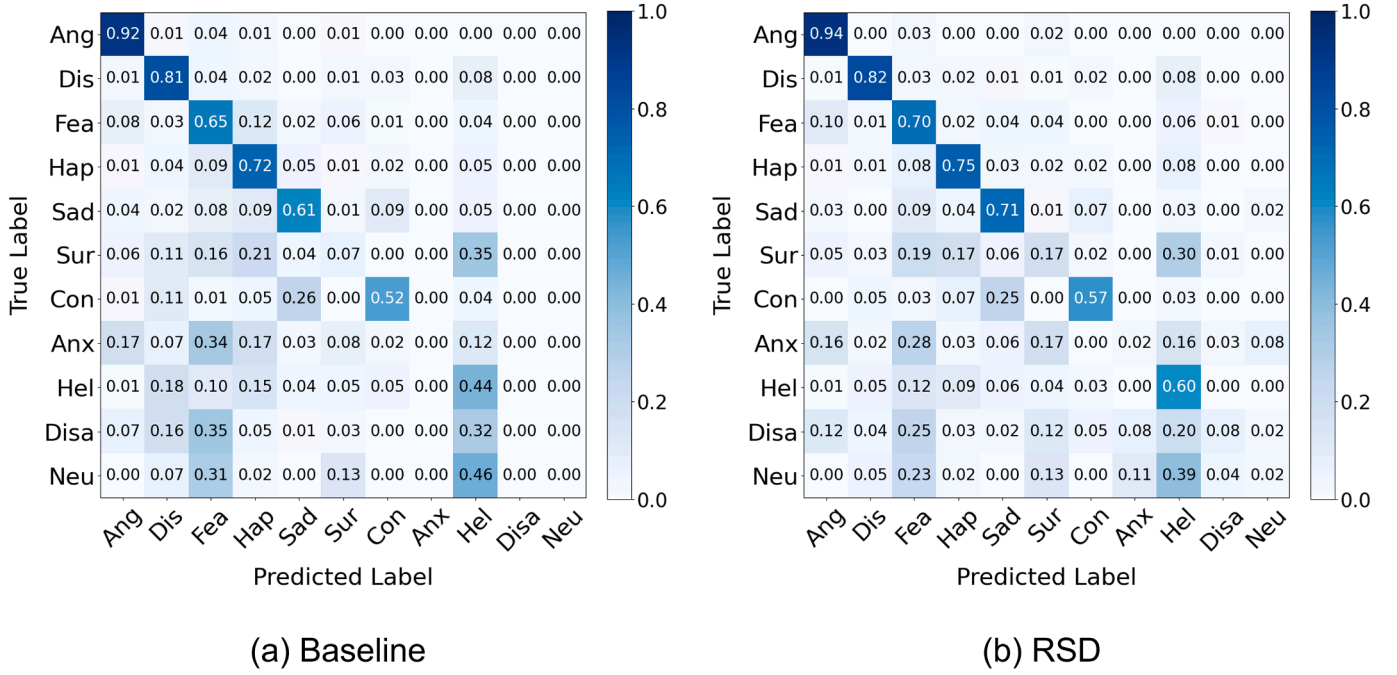


Fig. 3. Confusion matrices of baseline and RSD on the MAFW dataset.

Table 4

Comparison with SOTA methods on AFEW (7-class). Ang: anger. Dis: disgust. Fea: fear. Hap:happiness. Neu: neutral. Sad: sadness. Sur: surprise. The upper section reports results using the same evaluation protocol as existing methods; the lower section reports results under 5-fold cross-validation (marked with †). Best results in each section are in bold, second-best underlined.

Method	Accuracy of Each Emotion (%)							WAR(%)	UAR(%)
	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Neutral		
C3D + LSTM [33]	-	-	-	-	-	-	-	50.32	46.81
AMH [41]	-	-	-	-	-	-	-	51.64	47.96
T-MEP [21]	-	-	-	-	-	-	-	52.96	50.22
HiCMAE [18]	80.33	18.42	19.79	76.04	48.27	51.21	85.48	58.81	54.22
MMA-DFER [36]	83.33	21.05	10.52	<u>77.05</u>	68.96	14.63	67.74	54.55	49.04
PFFEI [42]	<u>81.30</u>	37.50	<u>39.10</u>	88.90	57.40	43.50	73.00	<u>63.18</u>	<u>60.21</u>
Baseline [40]	63.33	<u>39.47</u>	<u>35.12</u>	50.81	<u>58.62</u>	<u>56.59</u>	43.55	50.81	49.64
RSD (Ours)	69.23	44.00	57.14	66.70	54.55	61.76	80.56	63.64	61.99
HiCMAE † [18]	87.62	14.68	20.37	<u>78.72</u>	73.63	<u>63.95</u>	53.64	<u>62.39</u>	<u>56.09</u>
MMA-DFER † [36]	80.69	21.10	34.26	72.34	<u>65.17</u>	65.70	30.00	57.16	52.79
Qwen2.5-Omni † [35]	46.99	4.76	<u>31.78</u>	57.71	39.64	55.45	63.08	45.98	42.77
Qwen3 † [38]	48.62	0.97	26.47	61.81	42.26	60.91	65.64	47.73	43.81
Gemma3 † [39]	41.49	0.92	18.52	54.95	51.74	59.09	59.70	44.40	40.92
Baseline † [40]	74.75	<u>37.61</u>	11.11	76.60	56.98	42.31	<u>73.63</u>	58.86	53.28
RSD (Ours) †	<u>83.17</u>	46.51	29.63	80.85	56.81	39.42	89.05	66.52	60.80

To gain deeper insights into the discriminative capabilities of our model, we analyze the confusion matrices of the Baseline and RSD presented in Fig. 4, which are aggregated from the 5-fold cross-validation results. The comparison illustrates that RSD significantly improves the recognition accuracy for particularly challenging emotions. Specifically, the recognition accuracy for the “Fear” category experiences a remarkable increase from 11% to 30%, and the accuracy for “Neutral” rises substantially from 74% to 89%. Furthermore, the accuracy for “Anger” and “Disgust” also shows steady improvements. By effectively rectifying spatial and temporal misalignments before cross-modal interaction, RSD successfully mitigates severe cross-category confusions. For example, the Baseline mistakenly classifies 25% of “Fear” samples as “Sadness”, whereas RSD drastically reduces this misclassification rate to merely 6%. Additionally, the confusion of “Disgust” being falsely predicted as “Happiness” drops from 31% to 20%. While recognizing “Surprise” remains relatively challenging due to its high transient similarity with other dynamic expressions in unconstrained

environments, the overall findings confirm that utilizing generative text anchors allows the RSD framework to extract more discriminative spatiotemporal features, thereby alleviating cross-emotion ambiguity.

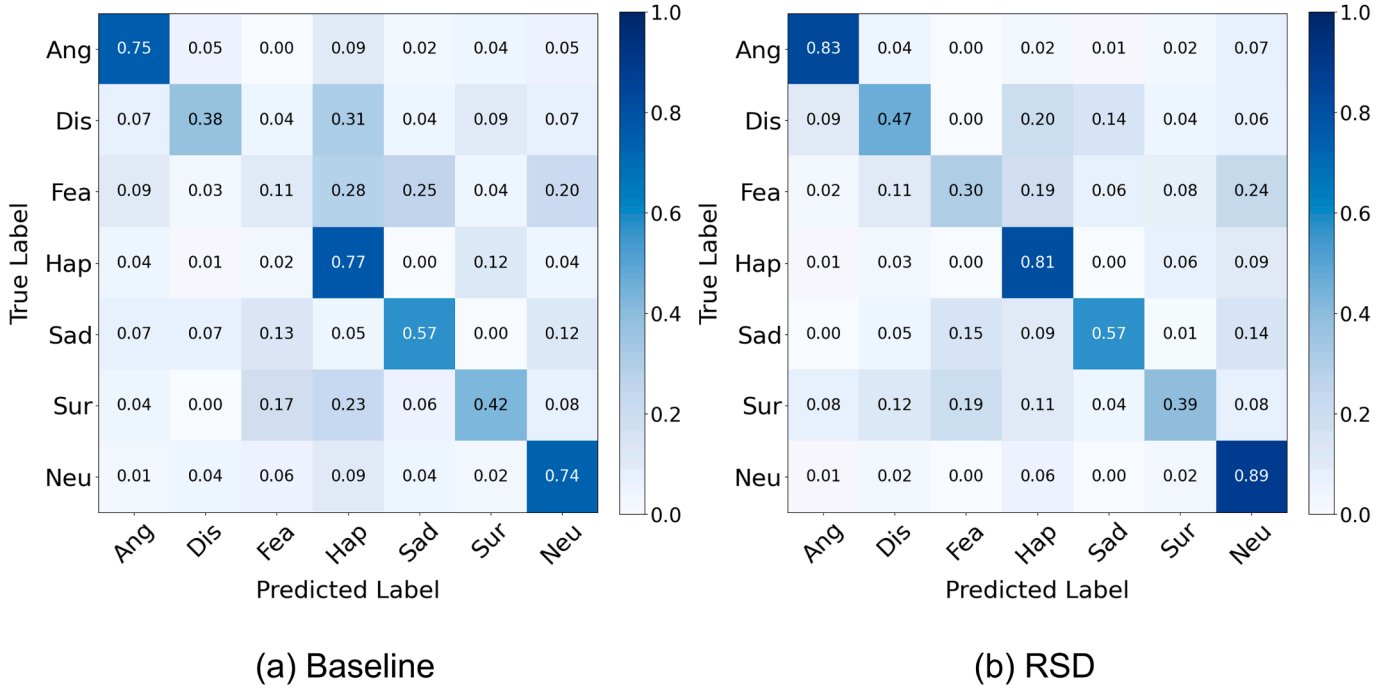
4.3.3. Computational efficiency analysis

Beyond recognition accuracy, we further evaluate the practical deployment efficiency of RSD by comparing its model parameters, computational complexity, inference latency, and GPU memory consumption with representative SOTA methods, as summarized in Table 6. Compared to other SOTA approaches, RSD achieves competitive computational efficiency. Notably, although RSD has a larger parameter footprint than MMA-DFER due to the integration of powerful pre-trained backbones, it achieves lower computational complexity, reducing the FLOPs by 33.8% from 35.60G to 23.56G. This indicates that RSD requires fewer runtime floating-point operations during forward propagation. Furthermore, RSD maintains a relatively modest memory footprint,

Table 5

Comparison of methods with reported standard deviations on MAFW and AFEW datasets.

Method	MAFW		AFEW	
	WAR (%)	UAR (%)	WAR (%)	UAR (%)
HiCMAE [18]	56.17 $\pm$ 5.96	42.65 $\pm$ 4.67	62.39 $\pm$ 2.10	56.09 $\pm$ 2.03
MMA-DFER [36]	58.52 $\pm$ 7.29	44.25 $\pm$ 7.04	57.16 $\pm$ 2.91	52.79 $\pm$ 3.01
Qwen2.5-Omni [35]	53.55 $\pm$ 3.50	45.25 $\pm$ 3.47	45.98 $\pm$ 4.07	42.77 $\pm$ 3.44
Qwen3 [38]	53.27 $\pm$ 3.30	40.53 $\pm$ 2.80	47.73 $\pm$ 2.78	43.81 $\pm$ 2.21
Gemma3 [39]	52.18 $\pm$ 3.89	39.10 $\pm$ 3.65	44.40 $\pm$ 2.58	40.92 $\pm$ 1.59
Baseline [40]	60.04 $\pm$ 7.71	43.16 $\pm$ 5.33	58.86 $\pm$ 4.82	53.28 $\pm$ 4.85
RSD (Ours)	65.95 $\pm$ 5.97	49.02 $\pm$ 5.17	66.52 $\pm$ 5.12	60.80 $\pm$ 5.36

**Fig. 4.** Confusion matrices of baseline and RSD on the AFEW dataset.**Table 6**

Computational efficiency comparison with SOTA methods.

Method	Total Params (M)	FLOPs (G)	Latency (ms)	Peak Memory (MB)
HiCMAE [18]	81.00	46.00	–	–
VAEmo [37]	39.00	49.33	–	–
MMA-DFER [36]	180.00	35.60	–	–
Baseline [40]	251.30	22.80	4.28	1002.14
RSD (Ours)	265.74	23.56	5.87	1067.46

requiring a peak GPU memory of 1067.46 MB, which is compatible with consumer-grade hardware or resource-constrained deployment environments. Relative to the Baseline, RSD introduces a modest increase of approximately 14.44M parameters, 0.76G FLOPs, and 1.59ms in latency, corresponding to relative increases of 5.7%, 3.3%, and 37.1%, respectively. These results suggest that RSD offers a favorable balance between recognition accuracy and computational efficiency, making it a viable candidate for practical deployment scenarios where both recognition performance and hardware constraints are critical considerations.

4.4. Ablation studies

4.4.1. Effects of different modules in RSD

To evaluate the progressive integration of the Temporal Anchor Rectification (TAR) and Spatial Anchor Rectification (SAR) modules into the baseline, Table 7 details the ablation results. Specifically, TAR achieves

implicit temporal synchronization via semi-relaxed optimal transport, while SAR rectifies spatial topology through low-rank affine scaling.

Relying solely on the Baseline, the model achieves a WAR of 60.04% and a UAR of 43.16% on the MAFW dataset. The independent introduction of the TAR module yields a significant performance leap, elevating the WAR and UAR to 62.95% and 46.36%, respectively. This absolute improvement of 2.91% in WAR clearly indicates that resolving temporal synchronous deviations is a critical bottleneck in audio-visual emotion recognition. Similarly, applying only the SAR module to the Baseline improves the performance to 62.07% in WAR and 45.37% in UAR, validating the necessity of rectifying inconsistent spatial topologies before cross-modal interaction.

Finally, the complete framework incorporating both TAR and SAR achieves the optimal performance of 65.95% WAR and 49.02% UAR. Notably, the combined performance gain of 5.91% in WAR significantly exceeds the sum of their individual improvements, which are 2.91% and

Table 7
The Effect of different modules in RSD on the MAFW dataset.

Method	TAR	SAR	WAR(%)	UAR(%)
Baseline	×	×	60.04	43.16
Baseline + TAR	✓	×	62.95	46.36
Baseline + SAR	×	✓	62.07	45.37
Baseline + TAR + SAR	✓	✓	65.95	49.02

2.03%. This mathematical evidence demonstrates a profound synergistic effect between the two modules. Specifically, the synchronized temporal foundation provided by TAR enables the SAR module to perform much more precise affine transformations across the strictly aligned audio-visual frames. Together, these mutually reinforcing modules successfully resolve complex spatiotemporal discrepancies, ensuring highly robust decision-making.

4.4.2. Reliability analysis of descriptive text generation

To comprehensively evaluate the reliability of the emotion description text generated by the **Emotion Semantic Text Anchor Generation** in Section 3.2.1, we follow Knight [43] by evaluating ROUGE-L for structural fluency and METEOR for vocabulary diversity. Because our long-form descriptions naturally incur length penalties in traditional exact-match metrics, we further introduce SBERT and BERTScore. These deep semantic metrics specifically calculate the underlying semantic consistency between our generated texts and the ground-truth textual descriptions provided by the MAFW dataset, by evaluating global sentence similarities and token-level equivalence.

As shown in Table 8, while all combined descriptions ($P_{template}$) face a length penalty in ROUGE-L, they achieve a METEOR score of 11.98%, outperforming Knight and proving strong foundational lexical correctness. Furthermore, the impressive SBERT similarity of 0.7202 and BERTScore of 0.4679 confirm a reasonable level of deep semantic consistency with the ground truth. Dimension-wise, the global emotional tendency (P_{env}) and facial physiological cues (P_{fau}) exhibit the highest similarities, reflecting the macro-level and visual bias in original human annotations. Conversely, the lower scores in the acoustic and verbal content dimensions (P_{acou} and P_{cont}) highlight the MLLM's capability to generate rich paralinguistic and linguistic details completely absent in the sparse ground truth. Consequently, these robust evaluations demonstrate that our generated descriptions are semantically accurate and highly reliable.

While the generated text anchors are semantically reliable, MLLMs may still produce occasional hallucinations. To mitigate this, RSD incorporates two complementary mechanisms. First, the text anchor loss $\mathcal{L}_T$ directly supervises the text encoder with ground-truth emotion labels, forcing it to retain only emotion-relevant information while filtering out hallucinated or irrelevant content. Second, the stochastic routing mechanism (RRSA) randomly skips spatial alignment with $p_{drop} = 0.3$ during training, preventing the model from over-relying on potentially noisy text anchors. To validate their effectiveness, we conduct ablation experiments on MAFW by removing $\mathcal{L}_T$ and disabling the stochastic routing (i.e., $\mathbb{I}_{route} = 1$ with $p_{drop} = 0$). As shown in Table 9, removing $\mathcal{L}_T$ causes a significant performance drop of 0.56% in WAR and 1.78% in UAR, confirming that the text anchor loss plays a crucial role in filtering hallucinated content. Similarly, disabling stochastic routing leads to a 2.50% WAR drop and a 3.90% UAR drop, as the model becomes vulnerable to noise in the text anchors. These results verify that both $\mathcal{L}_T$ and stochastic routing are essential for robustly mitigating the impact of potential hallucinations in MLLM-generated text anchors.

To further verify the robustness of RSD to variations in text anchor generation, we additionally evaluate our framework using a different MLLM backbone, LLaMA3 [44], for generating text anchors. As shown in Table 10, RSD with LLaMA3-generated anchors achieves 64.67% WAR and 48.08% UAR on MAFW, which is slightly lower than the Qwen2.5-Omni-7B based result but still significantly outperforms the baseline.

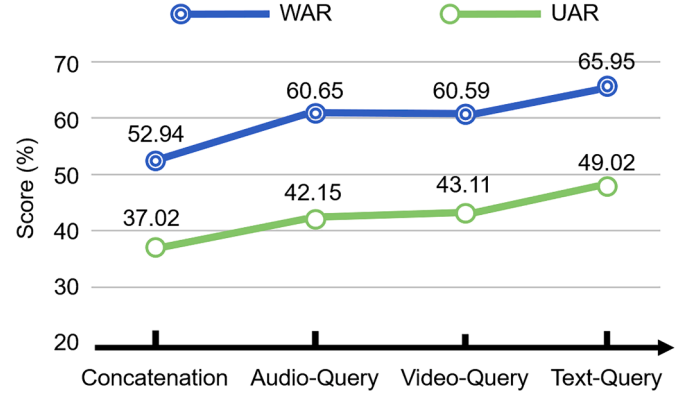


Fig. 5. The effect of different fusion strategies in the multi-task constrained anchoring fusion on the MAFW dataset.

This confirms that RSD maintains strong performance across different MLLM backbones, demonstrating its robustness to fluctuations in text anchor generation.

4.4.3. Effects of different prompt components

Table 11 presents the impact of different prompt template configurations within the Emotion Semantic Text Anchor Generation module on the MAFW dataset. Since the quality and composition of the generative text anchor play a crucial role in instructing the subsequent spatiotemporal rectification, we systematically removed specific textual components to evaluate the individual contribution of each contextual cue. The results demonstrate that the global affective tendency (P_{env}) serves as the most critical element. Its removal leads to the sharpest decline in overall performance, plummeting the WAR from 65.95% to 61.50%. This substantial decrease of 4.45% confirms that a high-level emotional prior provides the most stable semantic reference for aligning noisy and heterogeneous sensory data. Furthermore, the exclusion of audio-driven components, specifically the vocal content (P_{cont}) and acoustic analysis (P_{acou}), results in significant UAR drops to 46.33% and 44.07%, respectively. Notably, the absence of P_{acou} causes the most severe UAR degradation across all configurations. This phenomenon clearly indicates that fine-grained linguistic content and dynamic paralinguistic cues are indispensable for distinguishing ambiguous or minority emotion categories. Additionally, omitting detailed Facial Action Units (P_{fau}) decreases the WAR to 64.51% and the UAR to 44.89%. While this decline validates their necessary contribution to the fine-grained discriminative ability of the framework, their overall impact remains secondary compared to the broader affective context synthesized by the large language model. Ultimately, these findings confirm that the Full Template ensures the most comprehensive text anchor, enabling precise temporal synchronization and robust predictions across all emotion classes.

4.4.4. Effects of different fusion strategies in the multi-task constrained anchoring fusion

To validate the effectiveness of our fusion strategy, we conducted an ablation study comparing different cross-modal integration mechanisms, as illustrated in Fig. 5. The evaluation begins with a simple Concatenation of the three modalities. As anticipated, this unguided approach yields the lowest performance, achieving a WAR of 52.94% and a UAR of 37.02%. This substantial performance drop indicates that naive concatenation fails to adequately capture the complex interactions among different modalities, thereby highlighting the absolute necessity for a guided fusion mechanism even after spatial rectification and temporal synchronization have been performed.

To address this limitation, we further evaluated various asymmetric cross-attention configurations by altering the source of the Query (Q).

Table 8

Comparison of text generation reliability on the MAFW dataset.

Method / Description	ROUGE-L(%)	METEOR(%)	SBERT	BERTScore
Knight [43]	17.73	9.06	-	-
Only Tendency (P_{env})	18.52	10.36	0.6725	0.4778
Only Acoustic (P_{acu})	13.48	4.47	0.5961	0.4046
Only Facial Cues (P_{fau})	15.47	11.41	0.7066	0.4859
Only Content (P_{cont})	2.85	1.19	0.5072	0.3388
All Descriptions ($P_{template}$)	15.00	11.98	0.7202	0.4679

Table 9

Ablation study on hallucination mitigation mechanisms on the MAFW dataset.

Configuration	WAR (%)	UAR (%)	Performance Drop
Full RSD	65.95	49.02	-
w/o $\mathcal{L}_T$	65.39	47.24	-0.56% / -1.78%
w/o Stochastic Routing ($\mathcal{I}_{route} = 1$)	63.45	45.12	-2.50% / -3.90%

Table 10

Performance of RSD with different MLLM backbones for text anchor generation on the MAFW dataset.

MLLM Backbone	Method	WAR (%)	UAR (%)
Qwen2.5-Omni-7B [35]	RSD	65.95	49.02
LLaMA3 [44]	RSD	64.67	48.08
Qwen2.5-Omni-7B [35]	Baseline	60.04	43.16
LLaMA3 [44]	Baseline	60.23	41.92

Table 11

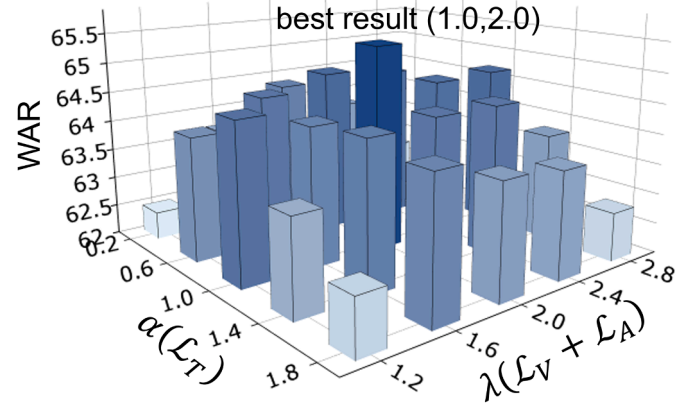
The Effect of different Prompt Template in Emotion Semantic Text Anchor Generation on the MAFW dataset.

Configuration	Description	WAR(%)	UAR(%)
Full Template	All components	65.95	49.02
w/o P_{env}	No Global Affective Tendency	61.50	47.48
w/o P_{cont}	No Audio Content	62.81	46.33
w/o P_{acu}	No Acoustic Analysis	63.30	44.07
w/o P_{fau}	No Facial AU	64.51	44.89

Using the audio modality as the Query yields a WAR of 60.65% and a UAR of 42.15%, while using the video modality produces similar results with a WAR of 60.59% and a UAR of 43.11%. Although these individual query-based approaches provide noticeable improvements over simple concatenation, they encounter a clear performance bottleneck. In stark contrast, our proposed Text-Query strategy shatters this bottleneck, achieving the optimal performance of 65.95% WAR and 49.02% UAR. This remarkable performance gap conclusively demonstrates the superiority of utilizing the text anchor as the Query. Instead of relying on raw sensory modalities that may contain inherent noise, the generated emotion semantic text establishes a stable, high-level semantic guide to instruct the fusion process. This mechanism ensures that the model actively retrieves and prioritizes the most emotionally salient visual and acoustic segments from the synchronized sequences. By implicitly grounding the cross-modal interaction in this unified semantic-temporal reference, our framework guarantees highly robust predictions in complex multimedia environments.

4.4.5. Effects of different multi-task loss weights setting in multi-task constraints

To evaluate the impact of the hierarchical multi-task constraints introduced during the anchoring fusion stage, we conducted a grid search ablation study on the loss hyper-parameters α and λ . As depicted in Fig. 6, RSD performance exhibits a clear sensitivity to the balance between semantic guidance and modality-specific discriminability. The optimal performance is achieved at the configuration of ($\lambda=2.0$, $\alpha=1.0$), reaching a peak WAR of 65.95%. Specifically, when the audio-visual penalty weight λ is set too low (e.g., $\lambda=1.2$), the performance sig-

**Fig. 6.** The effect of different multi-task loss weights setting in multi-task constraints on the MAFW dataset.

nificantly degrades, yielding a minimum WAR of 62.46% at $\alpha=0.2$. This confirms that without sufficient independent supervision, the audio-visual branches lose their modality-specific discriminative power, making the fusion process overly dependent on the text anchor. Conversely, an excessively high λ (e.g., $\lambda=2.8$) also leads to a performance decline, dropping the WAR to 62.81% when α is high. This suggests that prioritizing unimodal weakens the cross-modal spatiotemporal alignment learned by TAR and SAR.

Similarly, the text supervision weight α is crucial for maintaining system robustness. When α is near zero ($\alpha=0.2$), the model lacks direct ground-truth supervision for the textual branch, resulting in sub-optimal performance across all λ settings (e.g., 64.33% at $\lambda=2.0$). As α increases towards 1.0, we observe a consistent upward trend in performance. However, once α exceeds the optimal threshold (e.g., $\alpha=1.8$), the WAR again decreases to 63.86% (at $\lambda=2.0$), demonstrating that an appropriately regularized text anchor is essential to prevent the fusion process from being misled by potential deviations in the MLLM output. Ultimately, the optimal combination ($\lambda=2.0$, $\alpha=1.0$) effectively balances semantic guidance with physical nuances, ensuring robust emotion recognition.

4.4.6. Effects of dropout probability in stochastic routing

To validate the robustness of the stochastic routing mechanism in SAR (Section 3.3), we conduct a sensitivity analysis on the dropout probability p_{drop} , which controls the probability of randomly skipping spatial alignment during training. We vary p_{drop} from 0.0 to 0.5 with an interval of 0.1, while keeping all other hyperparameters fixed. As shown in Table 12, the model achieves optimal performance at $p_{drop}=0.3$. When $p_{drop}=0.0$ (i.e., deterministic alignment without stochastic protection), the performance drops to 65.24% WAR and 47.82% UAR, indicating that the stochastic routing effectively prevents over-reliance on the text anchor. Similarly, when p_{drop} is increased to 0.4 or 0.5, performance decreases to 64.10% WAR / 46.94% UAR and 65.41% WAR / 48.35% UAR, respectively, suggesting that excessive skipping weakens the spatial alignment and introduces instability. These results demonstrate that an appropriate dropout probability ($p_{drop}=0.3$) strikes a balance be-

Table 12
Sensitivity analysis of the dropout probability p_{drop} on the MAFW dataset.

p_{drop}	Description	WAR (%)	UAR (%)
0.0	Deterministic alignment (no stochastic protection)	65.24	47.82
0.1	Weak stochastic protection	65.76	47.73
0.2	Moderate stochastic protection	65.46	48.72
0.3	Optimal balance (Ours)	65.95	49.04
0.4	Strong stochastic protection	64.10	46.94
0.5	Excessive protection	65.41	48.35

Table 13
Robustness analysis under text noise in MLLM-generated anchors on the MAFW dataset.

Perturbation Ratio	Baseline		RSD (Ours)	
	WAR (%)	UAR (%)	WAR (%)	UAR (%)
0.0	60.04	43.16	65.95	49.02
0.1	58.89	42.31	64.97	46.61
0.2	56.73	40.52	63.36	46.18
0.3	54.10	38.86	63.41	46.60
0.4	52.68	37.83	60.64	45.48
0.5	51.20	36.72	58.59	44.91

tween leveraging the text anchor for spatial guidance and preserving modality-specific information, confirming the robustness of our RRSA module.

4.4.7. Robustness analysis

Text Perturbation Robustness. To evaluate the robustness of RSD against unreliable MLLM-generated text anchors, we conduct experiments under text degradation on the MAFW dataset. Specifically, words in the generated text anchors are randomly replaced with semantically unrelated noise words at perturbation ratios ranging from 0.1 to 0.5 with an interval of 0.1. As shown in Table 13, the performance of both models declines as text perturbation intensifies. However, RSD exhibits a more gradual degradation than the Baseline. Under 30% perturbation, RSD drops 2.54% WAR from its clean performance of 65.95% to 63.41%, while the Baseline drops 5.94% from 60.04% to 54.10%. Under 50% perturbation, RSD maintains 58.59% WAR, compared to 51.20% for the Baseline. These results suggest that the multi-task constraints and stochastic routing mechanism help mitigate the impact of corrupted text anchors, though the degradation still persists under severe perturbation. This indicates that while RSD is relatively more robust to text degradation, it does not fully eliminate the influence of unreliable MLLM-generated text anchors, which we discuss as a limitation in Section 5.

Temporal Shift Robustness. To further evaluate the robustness of RSD against temporal asynchrony, we conduct experiments with synthetic temporal shifts on the MAFW dataset. Specifically, we shift the video frames forward by offsets ranging from 0.1 to 1.5 seconds, simulating the asynchronous deviations commonly encountered in real-world scenarios. As shown in Table 14, RSD maintains relatively stable performance under shifts up to 0.3 seconds, with WAR fluctuating around 65%, while the Baseline shows a gradual decline from 60.04% to 58.28% under the same offsets. Under larger offsets (1.0 and 1.5 seconds), both methods exhibit performance degradation. Nevertheless, RSD consistently outperforms the Baseline across all offset levels, suggesting its resilience to temporal perturbations.

In-the-Wild Robustness. In addition to quantitative robustness verifications based on text perturbation and temporal offset, we supplemented our analysis with qualitative case studies across multiple real-world scenarios, covering common degradation conditions in field data such as audio noise, visual occlusion, low-quality video, and missing audio-visual modalities. All of these audiovisual input distortions simultaneously cause biases and incompleteness in the semantic anchors generated by MLLM.

Table 14
Robustness analysis under temporal shift on the MAFW dataset.

Shift Offset (s)	Baseline		RSD (Ours)	
	WAR (%)	UAR (%)	WAR (%)	UAR (%)
0.0	60.04	43.16	65.95	49.04
0.1	59.54	42.60	65.40	48.17
0.2	59.03	42.04	65.17	47.40
0.3	58.28	41.55	65.45	48.05
0.4	58.44	41.43	63.14	45.83
0.5	58.41	41.49	63.04	45.76
1.0	57.50	41.12	59.92	43.95
1.5	52.26	37.85	54.96	40.44

As shown in Fig. 7, HiCMAE [18], MMA-DFER [36], and the baseline [40] model all exhibited significant misclassifications in various degraded samples; while RSD, through TAR temporal alignment, SAR spatial correction, and multi-task constraints, effectively counteracted input distortion interference, achieving relatively accurate sentiment discrimination and verifying the robustness of RSD in real-world complex scenarios. (Fig. 7)

4.5. Qualitative analysis and visualization

4.5.1. Visualization analysis of temporal rectification in TAR module

To verify the effectiveness of resolving implicit asynchronous deviations, we visualize the dynamic emotional trajectories of the audio and visual modalities against the shared text semantic timeline in Fig. 8. The X-axis represents the progression of the text anchor, while the Y-axis denotes the corresponding emotional intensity for each modality. This visualization illustrates how emotional dynamics from different modalities evolve in accordance with the shared text-based semantic timeline. As shown in the *Before Synchronization* plot (top), the raw audio and video trajectories exhibit severe asynchronous temporal dynamics, confirming the presence of asynchronous deviations in underlying emotional semantics. As highlighted by the dashed red circle, the peaks and valleys of the two modalities occur at completely disjoint semantic moments, clearly illustrating the inherent implicit asynchronous deviations. Such non-linear desynchronization of emotional meaning would inevitably lead to semantic confusion during direct cross-modal interaction.

Conversely, in the *After Synchronization* plot (bottom), the sequences rectified via our TAR module demonstrate a highly consistent evolutionary trend. Driven by the semi-relaxed optimal transport mapping mechanism, the peaks and valleys of the audio and visual cues now rise and fall synchronously along the text semantic timeline. This alignment confirms that TAR effectively rectifies implicit asynchronous deviations by mapping heterogeneous sequences to a stable semantic-temporal reference. By successfully capturing this harmonized emotional trajectory, our framework ensures that the subsequent spatial topological rectification and multimodal fusion are performed on synchronized representations, thereby preserving the fine-grained affective dynamics.

4.5.2. Visualization analysis of SAR module

To intuitively demonstrate the impact of our spatial rectification, we visualize the evolution of feature space topology in Fig. 9. In the *Before Spatial Rectification* space, although the sequences have been temporally synchronized by TAR, the feature centers of the audio and visual modalities for each emotion category remain scattered, exhibiting inconsistent spatial topological distributions of emotional semantics. This mismatch confirms that temporal alignment alone is insufficient to bridge the geometric gap of hidden emotional geometry between heterogeneous modality distributions, as the internal feature structures remain discordant across different modality spaces.

After processing by SAR, the *After Spatial Rectification* space shows significant structural transformation. With low-rank diagonal affine

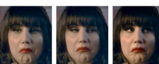
Scene	Input	True Label	Methods			
			HiCMAE	MMA-DFER	Baseline	RSD(Ours)
Audio Noise	 ... state of shock, ... wide eyes and open mouth. "How did you ~~get in~~ " ... , the pitch is high.	Fear	Surprise	Surprise	Fear	Fear
Visual Occlusion	 ... calm. ...upper eyelid sinks down, "psst, you got my vision on you." ... , the pitch is low, ...	Helplessness	Neutral	Sadness	Sadness	Helplessness
Low-Quality Video	 ... serious and focused. ... mouth is set in a firm line... "Target practice." The speech is ... neutral pitch.	Contempt	Happiness	Anxiety	Angry	Contempt
Missing Visual	 "with better... wares." The speech rate is moderate, the pitch is low, and the stress is low.	Angry	Disappointment	Sadness	Angry	Angry
Missing Audio	 ...state of concern or worry, ... upper eyelid is raised, eyebrows are furrowed.	Disgust	Fear	Disgust	Helplessness	Disgust

Fig. 7. Qualitative robustness comparison under five challenging wild conditions: audio noise, visual occlusion, low-quality video, missing visual modality, and missing audio modality.

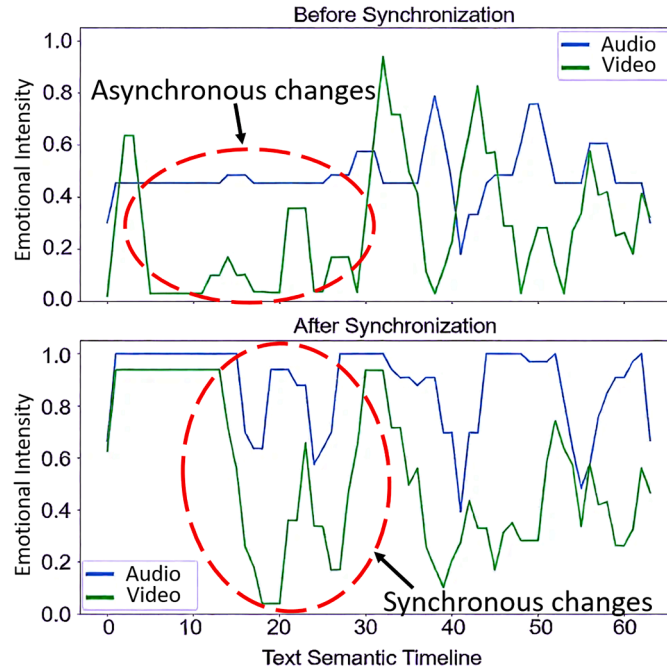


Fig. 8. Visualization of emotion intensity before and after temporal rectification on MAFW. (Top) Raw sequences show severe asynchrony. (Bottom) After TAR-based rectification, modalities achieve synchronous dynamics, verifying effective implicit temporal synchronization.

scaling in the RRSA mechanism, audio and visual feature spaces are geometrically aligned with text anchors. Specifically, for each emotion category, the feature centers of the audio, visual, and textual modalities converge toward a shared class-specific text anchor, resulting in consistent star-like topological distributions. This visual progression demonstrates that the SAR module effectively rectifies spatial topological discrepancies by implicitly aligning modality-specific features to the shared text anchors. This process establishes a structurally harmonized repre-

sentation for the subsequent fusion stage while maintaining the discriminative integrity of each modality.

4.5.3. Visualization analysis of cross-modal attention rectification

To comprehensively verify the fine-grained cross-modal semantic alignment achieved by our RSD framework, we visualize the cross-modal attention heatmap in Fig. 10, which explicitly illustrates the se-

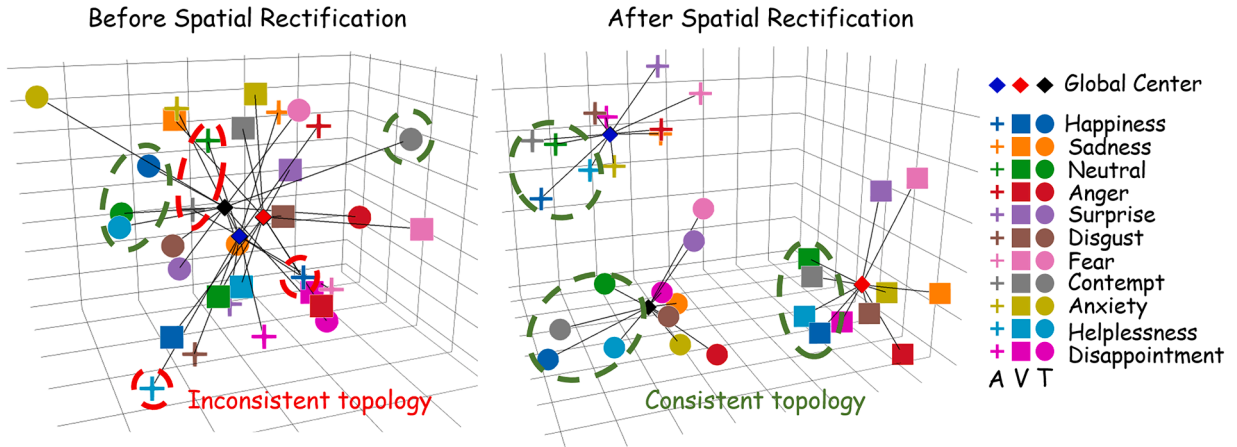


Fig. 9. Visualization of feature space topology before and after spatial rectification. (Left) Raw features show scattered and inconsistent spatial structures across modalities. (Right) SAR-rectified features form a consistent topology aligned with text anchors, exhibiting strong spatial consistency.

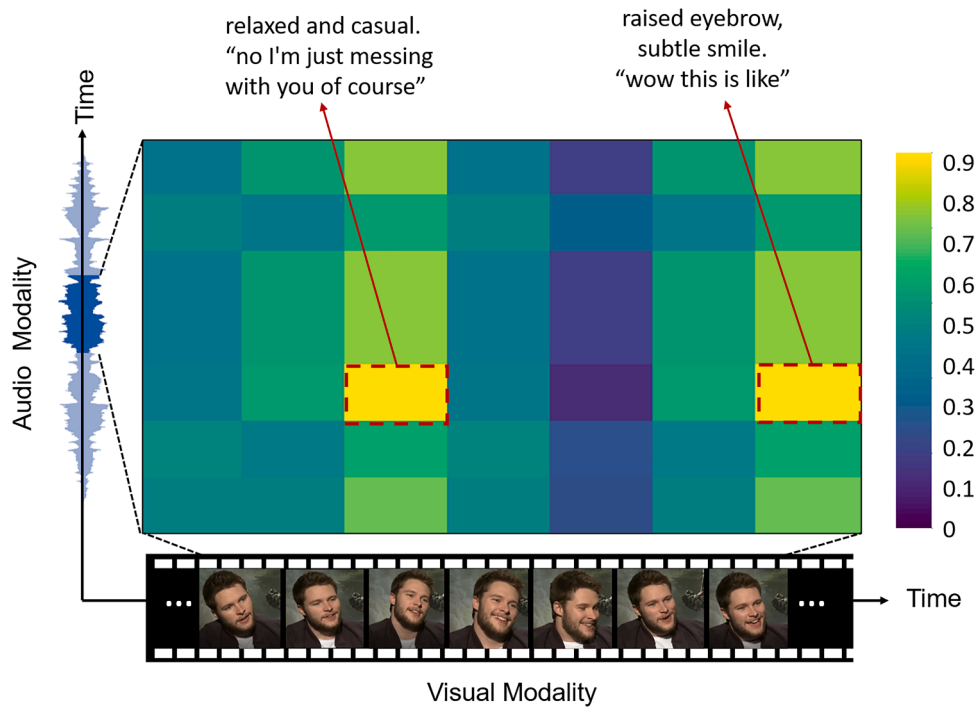


Fig. 10. Cross-modal attention visualization after temporal-spatial rectification. The heatmap illustrates the fine-grained alignment between audio waveforms, visual frames, and text semantics, where high-attention regions correspond to consistent emotional expressions across modalities, verifying the effectiveness of our framework.

mantic correspondence between audio and visual modalities after unified temporal-spatial rectification.

The heatmap axis corresponds to the temporal progression of visual frames and audio waveforms, with the color bar indicating attention weights, where brighter yellow regions denote higher cross-modal attention scores. Both the audio sequence (left) and visual frame sequence (bottom) are aligned to the shared emotion semantic text timeline, enabling direct observation of emotional consistency across heterogeneous modalities. As highlighted by the red dashed boxes, two representative high-attention regions exhibit a high degree of semantic coherence: in the first region, the audio segment with a relaxed, casual tone ("no I'm just messing with you of course") aligns with the visual frame showing a relaxed facial expression; in the second region, the audio conveying a surprised tone ("wow this is like") matches the visual frame with a raised eyebrow and subtle smile. This visualization confirms that after rectification by our

TAR and SAR modules, the cross-modal attention mechanism accurately captures fine-grained emotional semantic alignment. The high-attention regions consistently correspond to coherent emotional expressions, verifying that our framework effectively eliminates implicit temporal synchronous deviations and spatial topological discrepancies, establishing a unified semantic space that ensures reliable multimodal emotion recognition with preserved fine-grained affective dynamics.

4.5.4. Visualization analysis of multimodal representation

To validate the discriminative capability of our RSD framework, we compare emotional feature distributions across methods in Fig. 11. The HiCMAE [18] and MMA-DFER [36] show feature overlap and indistinct class separation, failing to capture discriminative emotional semantics. In contrast, our RSD achieves clear, compact clusters for each emotion category with minimal inter-class overlap. This visualization

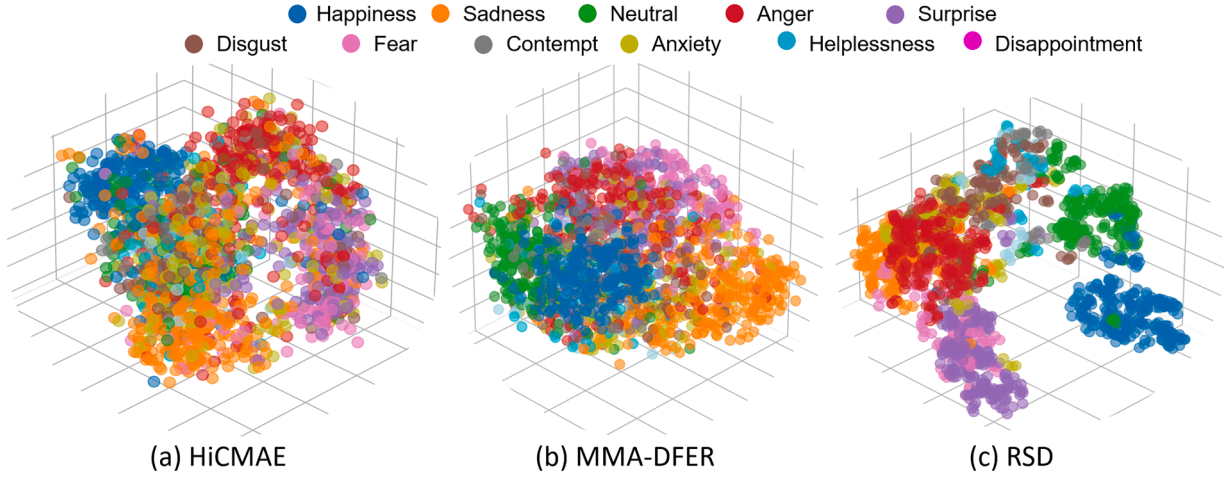


Fig. 11. Comparative visualization of emotional feature distributions on the MAFW dataset. (a) HiCMAE, (b) MMA-DFER, (c) our RSD. Each color denotes a distinct emotion category. Clearer class separation indicates better discriminative performance.

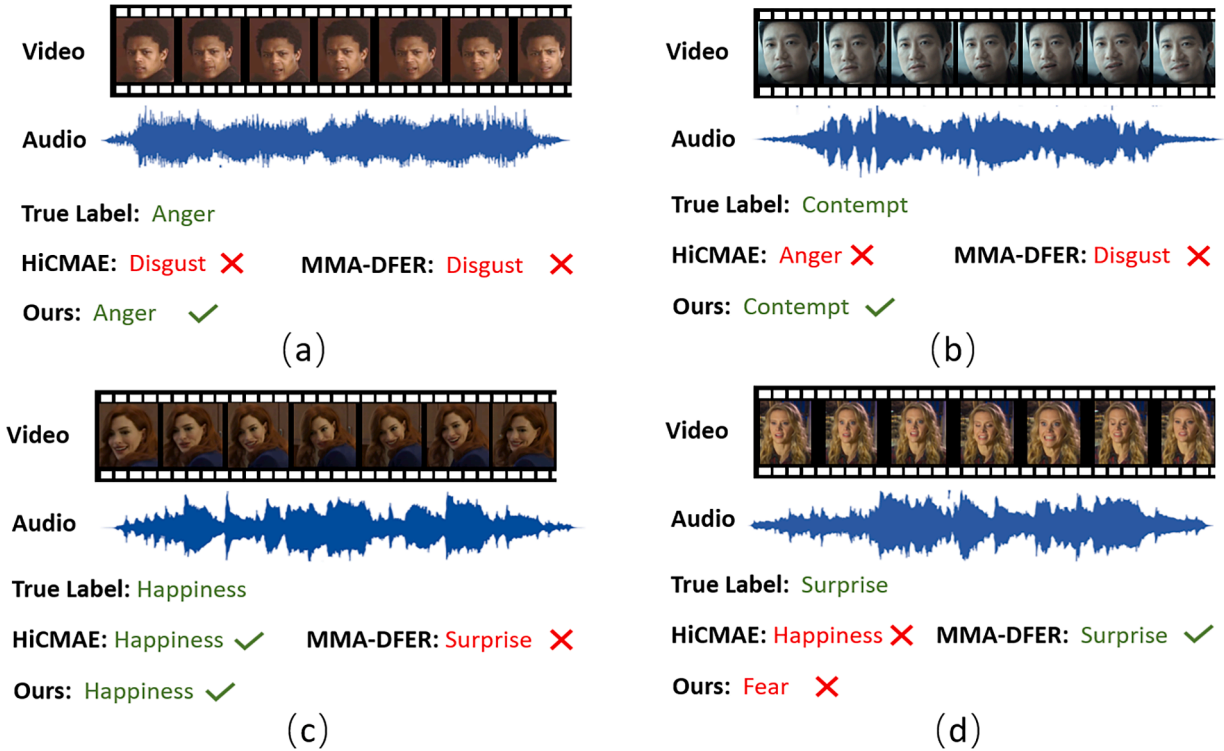


Fig. 12. Comparison of emotion recognition results using different methods. Each case includes video frames and audio waveforms, with ground-truth emotion labels. Predictions from HiCMAE, MMA-DFER, and our proposed RSD method are presented. RSD achieves higher prediction accuracy across diverse emotion categories.

confirms that RSD effectively enhances feature discriminability by resolving temporal-spatial discrepancies, laying a foundation for accurate multimodal emotion recognition.

4.5.5. Qualitative analysis of emotion recognition results

Fig. 12 presents a qualitative comparison of emotion recognition outcomes across different multimodal methods. In Fig. 12 (a), RSD correctly predicts “Anger”, while both HiCMAE and MMA-DFER misclassify it as “Disgust”. TAR precisely synchronizes the intense acoustic and facial cues of anger via the semantic cost matrix, while SAR’s anisotropic scaling reinforces the discriminative dimensions that separate anger from disgust. In Fig. 12 (b), RSD accurately identifies “Contempt”, whereas HiCMAE and MMA-DFER output “Anger” and “Disgust”, respectively. SAR effectively highlights the asymmetric facial geometry of contempt,

distinguishing it from the symmetric expressions of anger and disgust. In Fig. 12 (c), both RSD and HiCMAE correctly predict “Happiness”, while MMA-DFER misclassifies it as “Surprise”. In Fig. 12 (d), RSD mispredicts “Surprise” as “Fear”, while MMA-DFER correctly recognizes it. Surprise and fear share highly overlapping visual patterns, and surprise is intrinsically transient, with its emotional peak lasting only 1–2 frames. TAR’s frame re-ordering mechanism dilutes these sparse surprise frames across the semantic timeline, while SAR’s scaling cannot fully separate their overlapping geometries. Beyond these inherent challenges, we further examine the quality of the MLLM-generated text anchor for this sample. The generated description reads: “The character displays dramatic facial muscle movements, starting with the inner and outer eyebrows raised, upper eyelids severely raised exposing the whites of her eyes, and a sudden wide jaw drop. This quickly transitions into a tensed, hor-

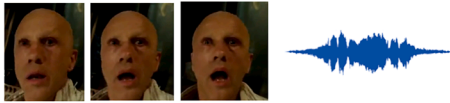
	Sample ID: 11170, Fear	Sample ID: 20009, Disappointment
Prompt Dimension		
Affective Tendency (P_{env})	The man appears to be in a state of shock, as indicated by his wide eyes and open mouth.	The woman appears to be in a state off frustration or annoyance, as indicated by her facial expression and body language.
Facial AU (P_{fau})	The man's upper eyelids are raised, his eyebrows are slightly raised, and his mouth is open, revealing his teeth. His cheeks are slightly puffed out, and his jaw is slightly clenched.	The woman's eyebrows are slightly furrowed, her mouth is set in a firm line, and her lips are pursed, suggesting a sense of irritation or displeasure. Her eyes are looking downward, possibly indicating a feeling of disappointment or frustration.
Speech Content (P_{cont})	"How did you get in?"	"This is crazy to me that you're still defending him."
Audio Features (P_{acou})	The speech rate is medium, the pitch is high, and the stress is medium.	The speech rate is moderate, the pitch is medium, and the stress is high, indicating a strong emotional delivery.

Fig. 13. Case study of text anchors across distinct prompt dimensions.

horizontal stretching of the lip corners...The speech rate is high, the pitch is high, and the stress is high." While the anchor faithfully captures both surprise-like and fear-like cues, it fails to emphasize the transient nature of surprise, introducing semantic ambiguity that leads to high transport costs for both categories in TAR. This explains why surprise exhibits systematic confusion with fear and helplessness in our results. Similarly, disappointment shows the lowest recognition accuracy due to its semantic ambiguity with sadness and fear, which SAR's low-rank scaling cannot fully resolve. These observations demonstrate that RSD achieves reliable recognition for emotions with strong, distinctive patterns (e.g., anger, contempt) by effectively leveraging TAR and SAR, while emotions that are visually overlapping and temporally transient (e.g., surprise, disappointment) remain challenging.

4.5.6. Case study on prompt fine-grainedness and diversity

To illustrate the role of our structured prompt template, Fig. 13 presents a qualitative case study examining the generated semantic text anchors across two representative samples from the benchmark, specifically comparing fear against disappointment.

Specifically, our multidimensional prompt template successfully guides the model to capture fine-grained physical and acoustic evidence across distinct emotional states. In the visual stream, the template reliably isolates localized facial muscle activations such as raised upper eyelids and clenched jaw in fear, contrasting clearly with furrowed eyebrows, pursed lips, and downward gaze in disappointment. In the auditory stream, the template explicitly quantifies speech rate, pitch, and acoustic stress variations across contrasting speech deliveries.

Furthermore, the two samples exhibit clear semantic diversity in lexical choice, conversational content, and paralinguistic tone, confirming that the structured template prevents the model from generating repetitive boilerplate text. These distinct, high-fidelity multimodal descriptions provide reliable reference benchmarks for subsequent temporal synchronization in TAR and geometric calibration in SAR.

5. Limitations and discussion

Despite the encouraging results, several limitations of RSD should be acknowledged. The framework relies on MLLM-generated text anchors, whose quality inherently bounds the overall performance. The text anchor loss $\mathcal{L}_T$ provides some resilience against textual noise, but cannot fully eliminate the impact of severely distorted anchors. This dependency also introduces computational overhead, as the offline MLLM generation requires considerable GPU memory and processing time for large-scale data, with subsequent model training adding further hours. On the experimental front, MAFW and AFEW are limited in scale and emotion categories, and class imbalance remains a persistent issue, biasing the model toward frequent categories. Furthermore, transient emotions like "Surprise" have peaks lasting only 1–2 frames, which TAR's frame re-ordering struggles to preserve, while semantically ambiguous categories like "Disappointment" overlap substantially with "Sadness" and "Fear", challenging SAR's ability to fully separate their geometric distributions. Finally, although RSD's core principles are potentially generalizable to other multimodal tasks, different applications may require distinct temporal granularities, fusion strategies, or text anchor definitions, necessitating task-specific adaptations. We view these limitations as constructive directions for future work.

6. Conclusion

In this work, we propose a novel Rectifying Spatiotemporal Discrepancies framework to systematically resolve the implicit spatiotemporal heterogeneity inherent in audio-visual emotion recognition, which refers to asynchronous deviations and inconsistent topological distributions in emotional semantics. Leveraging generative text anchors as a unified high-level reference, RSD addresses both temporal and spatial discrepancies through two specialized modules. Specifically, the Temporal Anchor Rectification (TAR) module utilizes an MLLM to establish stable text anchors and applies semi-relaxed optimal transport theory

to implicitly synchronize asynchronous audio-visual sequences along a shared semantic timeline. Furthermore, the Spatial Anchor Rectification (SAR) module employs a Rank-Restricted Stochastic Alignment (RRSA) mechanism to bridge the geometric gap of latent emotional structures between modalities, aligning modality-specific topological structures with the shared semantic coordinate system. Finally, a Multi-task Constrained Anchoring Fusion strategy integrates these rectified representations to ensure robust emotion prediction in unconstrained environments.

Extensive experiments on the competitive MAFW and AFEW benchmarks demonstrate that RSD outperforms existing state-of-the-art methods. Beyond performance gains, our framework enables robust emotion recognition by effectively rectifying spatiotemporal discrepancies while maintaining the discriminative integrity of heterogeneous modalities in the wild. By leveraging generative text anchors to guide both temporal synchronization and spatial alignment, our work provides an effective strategy for addressing the inherent spatiotemporal discrepancies in audio-visual emotion data. Future research will explore the extensibility of the RSD framework to a broader range of complex, unconstrained affective computing tasks, and further investigate more robust mechanisms to mitigate the reliance on MLLM-generated text anchors.

CRedit authorship contribution statement

Qiyin Zhou: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization; **Ke Wang:** Writing – review & editing, Validation, Software; **Yuanyuan Liu:** Writing – review & editing, Supervision, Conceptualization; **Kejun Liu:** Writing – review & editing, Validation; **Wujie Zhou:** Writing – review & editing, Supervision; **Chang Tang:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Research Link Provided.

Appendix A. Prompt template for text-anchor generation

The exact prompt template $P_{template}$ used to instruct the multimodal large language model is as follows:

"You are an extremely meticulous multimodal AI analyst. Your task is to analyze the provided video and generate a structured report. The report must contain four specific sections and strictly follow the following format.

1. Environmental Description: Please describe the character's (man/woman's) emotions in one sentence. Do not use the text 'in the video' and directly describe the person in the video.

2. Facial Action Unit: Analyze the character's facial expression in detail. Describe the specific facial muscle movements you observed (e.g., upper eyelid raised, mouth corner raised, eyebrow drooped). Do not use high-level emotion words such as 'happiness' or 'sadness'. Just describe the specific facial muscle movements, no need to describe the meaning of these facial muscle movements.

3. Speech Transcription: Transcribe the speech from the audio track. If there is no speech, please start with 'Translated speech said' or 'No speech detected'.

4. Audio Features: Analyze the acoustic characteristics of the speech. Describe the speech rate (e.g.: slow, medium, high), pitch (e.g.: low, medium, high), and stress (e.g.: low, medium, high). If there is no speech, please indicate 'Audio feature not applicable'. Just describe the rate, pitch and stress without further explanation of additional content.

Finally, please concatenate the above formatted content into a paragraph. Now, generate a report based on the provided video."

References

- [1] Y. Jiang, W. Li, M.S. Hossain, M. Chen, A. Alelaiwi, M. Al-Hammadi, A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition, *Inf. Fusion* 53 (2020) 209–221.
- [2] A. Hussain, E. Cambria, S. Poria, A. Hawalah, F. Herrera, Information fusion for affective computing and sentiment analysis, *Inf. Fusion* 71 (2021) 97–98.
- [3] K. Wang, Y. Liu, K. Liu, Y. Xia, C. Tang, Y. Zhan, Z. Chen, et al., EmWorld: emotion world model with latent state evolution for scenario-incremental dynamic facial expression recognition, in: *Forty-third International Conference on Machine Learning*.
- [4] K. Liu, Y. Liu, L. Wei, C. Tang, Y. Zhan, Z. Chen, Z. Chen, Smile on the face, sadness in the eyes: bridging the emotion gap with a multimodal dataset of eye and facial behaviors, *IEEE Trans. Multimed.* 28 (2026) 7331–7342.
- [5] K. Liu, Z. Chen, Y. Liu, Y. Zhan, W. Xiang, H. Zhang, et al., World-model inspired emotion-aware token refinement for training-free multimodal emotion recognition, in: *Forty-third International Conference on Machine Learning*.
- [6] C. Palmero, M. Develasco, M.A. Hmani, A. Mtibaa, L.B. Letaifa, P. Buch-Cardona, R. Justo, T. Amorese, E. González-Fraile, B. Fernández-Ruano, et al., Exploring emotion expression recognition in older adults interacting with a virtual coach, *IEEE Trans. Affect. Comput.* 16 (2025) 2303–2320.
- [7] J. Yan, N. Zhou, Y. Cheng, F. Zhang, H. Wang, M. Wang, B. Jin, M. Li, Q. Lu, W. Zhang, Application of machine-vision-driven physics-informed neural networks in pantograph-catenary system state detection, *Mech. Syst. Signal Process.* 257 (2026) 114577.
- [8] H. Yang, Z. Liu, W. Liu, H. Wang, Y. Zhang, H. Wang, Graph-MDETR: a graph-guided Mamba-DETR network for UAV catenary support components detection in electrified railways, *IEEE Trans. Intell. Transp. Syst.* 27 (2026) 6319–6332.
- [9] X. Wang, H. Jiang, Y. Dong, M. Mu, Spatial-channel collaborative multi-scale graph interaction deep transfer learning for unsupervised rotating machinery fault diagnosis, *Eng. Appl. Artif. Intell.* 176 (2026) 114691.
- [10] H. Lei, X. Cheng, Q. Qin, D. Wang, K. Fan, H. Huang, Q. Gu, Y. Wu, Z. Jiang, Y. Chen, et al., M3-JEPA: multimodal alignment via multi-gate MoE based on the joint-embedding predictive architecture, *arXiv preprint arXiv:2409.05929*, (2024).
- [11] M.Z.U. Rehman, D. Raghuvanshi, U. Jain, S. Bansal, N. Kumar, A multimodal-multitask framework with cross-modal relation and hierarchical interactive attention for semantic comprehension, *Inf. Fusion* 126 (2025) 103628.
- [12] S. Zhang, S. Zhang, T. Huang, W. Gao, Multimodal deep convolutional neural network for audio-visual emotion recognition, in: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 281–284.
- [13] M. Hao, W.-H. Cao, Z.-T. Liu, M. Wu, P. Xiao, Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features, *Neurocomputing* 391 (2020) 42–51.
- [14] F. Kuhnke, L. Rumberg, J. Ostermann, Two-stream aural-visual affect analysis in the wild, in: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, IEEE, 2020, pp. 600–605.
- [15] B.T. Atmaja, M. Akagi, Multitask learning and multistage fusion for dimensional audio-visual emotion recognition, in: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 4482–4486.
- [16] R.G. Praveen, E. Granger, P. Cardinal, Cross attentional audio-visual fusion for dimensional emotion recognition, in: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, IEEE, 2021, pp. 1–8.
- [17] E. Ghaleb, J. Niehues, S. Asteriadis, Joint modelling of audio-visual cues using attention mechanisms for emotion recognition, *Multimed. Tools Appl.* 82 (8) (2023) 11239–11264.
- [18] L. Sun, Z. Lian, B. Liu, J. Tao, Hicmae: hierarchical contrastive masked autoencoder for self-supervised audio-visual emotion recognition, *Inf. Fusion* 108 (2024) 102382.
- [19] Z. Zhao, I. Patras, Prompting visual-language models for dynamic facial expression recognition, in: *British Machine Vision Conference (BMVC)*, 2023, pp. 1–14.
- [20] H. Chen, H. Huang, J. Dong, M. Zheng, D. Shao, Finecliper: multi-modal fine-grained clip for dynamic facial expression recognition with adapters, in: *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 2301–2310.
- [21] X. Zhang, M. Li, S. Lin, H. Xu, G. Xiao, Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild, *IEEE Trans. Circuits Syst. Video Technol.* 34 (5) (2023) 3192–3203.
- [22] Y. Li, W. Gan, K. Lu, D. Jiang, R. Jain, AVES: an audio-visual emotion stream dataset for temporal emotion detection, *IEEE Trans. Affect. Comput.* 16 (2024) 438–450.

- [23] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, vol. 2019, 2019, pp. 6558.
- [24] M. Li, D. Yang, Y. Lei, S. Wang, S. Wang, L. Su, K. Yang, Y. Wang, M. Sun, L. Zhang, A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 10074–10082.
- [25] Z. Liu, B. Zhou, D. Chu, Y. Sun, L. Meng, Modality translation-based multimodal sentiment analysis under uncertain missing modalities, *Inf. Fusion* 101 (2024) 101973.
- [26] Y. Zeng, W. Yan, S. Mai, H. Hu, Disentanglement translation network for multimodal sentiment analysis, *Inf. Fusion* 102 (2024) 102031.
- [27] S. Feng, Q. Zhou, Y. Liu, K. Wang, K. Liu, C. Tang, Unsupervised multimodal emotion-unified representation learning with dual-level language-driven cross-modal emotion alignment, *Pattern Recognit.* 176 (2026) 113160.
- [28] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.
- [29] T. Zhu, Q. Liu, F. Wang, Z. Tu, M. Chen, Unraveling cross-modality knowledge conflicts in large vision-language models, *arXiv preprint arXiv:2410.03659*, (2024).
- [30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, *ICLR*, 2021.
- [31] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, M. Auli, Data2vec: a general framework for self-supervised learning in speech, vision and language, in: *International Conference on Machine Learning*, PMLR, 2022, pp. 1298–1312.
- [32] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, *arXiv preprint arXiv:1910.01108*, (2019).
- [33] Y. Liu, W. Dai, C. Feng, W. Wang, G. Yin, J. Zeng, S. Shan, 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*, 2022, pp. 24–32.
- [34] J. Kossai, G. Tzimiropoulos, S. Todorovic, M. Pantic, AFEW-VA database for valence and arousal estimation in-the-wild, *Image Vis. Comput.* 65 (2017) 23–36.
- [35] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, et al., Qwen2. 5-Omni Technical Report, *arXiv preprint arXiv:2503.20215*, (2025).
- [36] K. Chumachenko, A. Iosifidis, M. Gabbouj, 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*, 2024, pp. 4673–4682.
- [37] H. Cheng, Z. Zhao, Y. He, Z. Hu, J. Li, M. Wang, R. Hong, Vaemo: efficient representation learning for visual-audio emotion with knowledge injection, in: *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 5547–5556.
- [38] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 Technical Report, *arXiv preprint arXiv:2505.09388*, (2025).
- [39] A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, et al., Gemma 3 technical report, *CoRR abs/2503.19786* (2025).
- [40] L. Wei, Y. Liu, K. Liu, Y. Huang, S. Feng, Emotion-aware text generation for instructing audio-visual emotion recognition, in: *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–8. <https://doi.org/10.1109/IJCNN64981.2025.11227919>
- [41] S. Yoon, S. Dey, H. Lee, K. Jung, Attentive modality hopping mechanism for speech emotion recognition, in: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 3362–3366.
- [42] S. Li, C. Lu, Y. Zong, H. Lian, W. Zheng, Feature evaluation and joint interaction for audio-visual emotion recognition, *IEEE Trans. Circuits Syst. Video Technol.* 36 (2025) 5146–5160.
- [43] J. Wang, M. Yan, Y. Zhang, J. Sang, From association to generation: text-only captioning by unsupervised cross-modal mapping, *arXiv preprint arXiv:2304.13273*, (2023).
- [44] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al., The llama 3 herd of models, *arXiv preprint arXiv:2407.21783*, (2024).