Robust Low-Rank Sparse Framework for Video-Based Affective Computing

Feng-Qi Cui, Jinyang Huang*, Sirui Zhao, Xinyu Li, Xin Yan, Ziyu Jia, Xiaokang Zhou

***Abstract*—Video-based Affective Computing (VAC) is central to emotion understanding and emerging consumer electronics applications, yet it often suffers from unstable optimization and representational degradation under complex, non-stationary affective dynamics in the wild. A key challenge is that short-term behavioral fluctuations can convey different affective meanings under different long-term emotional contexts, whereas most existing VAC models entangle heterogeneous factors in a single latent space, leading to component mixing and cross-scale inconsistency. We propose the Low-Rank Sparse Emotion Understanding Framework (*LSEF*), a plug-and-play framework grounded in the low-rank sparse principle that models affective dynamics as the composition of long-term emotional bases and sparse transient perturbations. Specifically, *LSEF* consists of four modules: the Stability Encoding Module (*SEM*) extracts low-rank bases by emphasizing slow-varying context patterns; the Dynamic Decoupling Module (*DDM*) isolates sparse transients via temporal routing and orthogonalized relational propagation; and the Consistency Integration Module (*CIM*) restores multi-scale coherence between stability and reactivity under hierarchical aggregation. We further propose a Rank Aware Optimization (*RAO*) strategy that adaptively modulates perturbation strength according to rank- and sparsity-sensitive structure to improve training stability and generalization. Experiments on discrete and continuous VAC benchmarks show consistent gains in robustness and dynamic discrimination across general CNN-based backbones, validating hierarchical low-rank sparse modeling as a principled foundation for video affect understanding.**

***Index Terms*—Video-based Affective Computing, low rank sparse modeling, representation learning.**

I. Introduction

Since showing great potential in real-world applications [1], *e.g.*, psychological assessment [2] and human–computer interaction [3]–[5], video-based Affective Computing (VAC) has attracted growing attention. It aims to model the dynamic evolution of human emotions from video sequences, enabling automatic perception and analysis of affective states [6], [7]. VAC is also increasingly relevant to consumer electronics, such as in-vehicle systems [8], [9] and service robots [10], where affect-aware interaction enhances personalization and user experience.

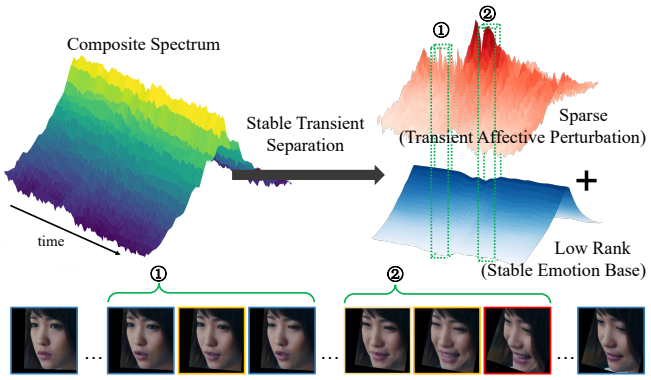


Fig. 1: Illustration of the inherent low-rank sparse structure of affective dynamics. The smooth low-rank curve (blue) represents stable emotional states, while the sparse curve (red) highlights transient expressive surges.

According to different affect representation paradigms, VAC research can be broadly categorized into two main directions, categorical emotion recognition and dimensional emotion regression [11], [12]. The former aims to recognize representative emotional expressions (*e.g.*, facial expressions [13], [14]) from videos to determine the overall emotion category, which is particularly useful for affective interaction and social signal recognition. The latter maps emotions onto a continuous multidimensional space, such as the valence arousal plane [15], to capture temporal variations in emotional intensity and psychological fluctuation, providing a more realistic characterization of continuous affect generation mechanisms. Despite different outputs, both settings require models to preserve long-term affective trends while being sensitive to short-lived discriminative cues. Benefiting from the rapid development of deep spatiotemporal modeling techniques, the performance of video-based affect analysis has seen significant improvements. For instance, representative studies have constructed unified 3D convolutional networks [16] or spatiotemporal attention networks [17], [18] to jointly capture spatial structures and temporal dynamics of video signals, thereby improving model robustness and generalization in complex real-world scenarios [19]. However, these methods typically learn heterogeneous dynamics in a shared representation pathway, which can be fragile under non-stationary affective evidence and noisy in-the-wild videos common in consumer devices [20], [21].

Although VAC has achieved remarkable progress in both categorical recognition and dimensional regression, pioneer

Feng-Qi Cui and Sirui Zhao are with the MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China.

Feng-Qi Cui, Jinyang Huang, Xinyu Li are with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, and the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China.

Xin Yan is with Cylingo Group, Beijing, China.

Ziyu Jia is with the Beijing Key Laboratory of Brainnetome and Brain-Computer Interface, and the Brainnetome Center, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

Xiaokang Zhou is with the Faculty of Business Data Science, Kansai University, Osaka 565-8585, Japan, and also with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan

Corresponding author*: Jinyang Huang (Email: hjy@hfut.edu.cn).

approaches still struggle to achieve robust and interpretable affect understanding due to the lack of structured modeling of emotional dynamics. This is because there are not only long-term emotional tones but also short-term emotional fluctuations that affect recognition results. Psychostatistical theories decompose emotions into stable components that reflect persistent psychological states and transient components driven by momentary stimuli [22]. From the perspective of structured signal modeling, this emotional process can be characterized using a low-rank sparse framework, where long-term trends and short-term fluctuations jointly contribute to the affective dynamics [23]. As illustrated in Fig. 1, the low-rank emotional field describes smooth and persistent affective baselines, while sparse high-frequency perturbations correspond to brief, localized, and discriminative emotional variations. This view also matches practical usage: users exhibit relatively stable baselines during a session, while key affective events appear as bursts. Without explicit disentanglement, models may overreact to spurious high-frequency artifacts or oversmooth subtle fluctuations, harming reliability in real-time applications.

However, mainstream methods typically learn these heterogeneous components jointly within a unified feature space, leading to multi-level representational entanglement and optimization instability. Specifically, this fundamental mechanism deficiency manifests in three key issues for visual affect analysis: *1) Structural Component Mixing:* Stable emotional trends and localized transient variations interfere with each other within the representation space, making it difficult to preserve coherent affective trajectories while capturing fine-grained changes. *2) Dynamic Propagation Coupling:* Without isolating stable and transient components during propagation, the two dynamics become cross-contaminated across spatiotemporal dimensions, weakening independent pathways for long-term tones and transient bursts. *3) Cross-scale Structural Inconsistency:* As features propagate to deeper layers and are aggregated across scales, stable trends may be diluted while sparse variations are over-smoothed, especially under efficiency-driven operations (downsampling/pooling) often used for on-device deployment. These biases hinder maintaining global stability and local sensitivity simultaneously, resulting in unstable optimization and degraded representations.

To address these challenges, we revisit affective video modeling from the perspective of low-rank sparse emotional organization and develop the ***Low-Rank Sparse Emotion Understanding Framework*** **(*LSEF*)**, a unified architecture grounded in the low-rank sparse principle that can be seamlessly integrated into general video backbones. *LSEF* follows a structured dynamic hierarchical decomposition paradigm, constructing a systematic pathway from stable encoding to dynamic decoupling and consistency integration. First, the ***Stability Encoding Module*** **(*SEM*)** extracts stable emotional bases via frequency-aware decomposition with low-rank priors, suppressing random perturbations to form a coherent affective field. Second, the ***Dynamic Decoupling Module*** **(*DDM*)** decouples dynamics through temporal gating and graph orthogonalization, disentangling sparse transient signals to improve discriminability and generalization. Third, the ***Consistency Integration Module*** **(*CIM*)** aligns stability and reactivity via multi-scale fusion and relational aggregation, mitigating scale drift and over-smoothing. Finally, ***Rank Aware Optimization*** **(*RAO*)** adaptively regulates perturbations to balance low-rank smoothness and sparse sensitivity, improving training stability and generalization under heterogeneous and noisy affective patterns relevant to consumer videos.

Our main contributions are summarized as follows:

- To the best of our knowledge, this paper is the first to introduce the low-rank sparse principle into VAC and propose the *LSEF* framework, reformulating affective dynamics as a low-rank sparse compositional process and providing a unified lens bridging stability and sensitivity for both categorical and dimensional affect tasks.
- We design hierarchical plug-and-play modules *SEM*, *DDM*, and *CIM* to enable structured decomposition, independent propagation, and multi-scale integration, addressing component entanglement, propagation coupling, and scale imbalance.
- We propose a *RAO* strategy that regulates perturbations in a rank- and sparsity-aware manner, offering an effective trade-off between stability and adaptability during optimization.
- Extensive experiments on multiple video affective benchmarks demonstrate that *LSEF* achieves SOTA performance in both classification and regression, with superior robustness and generalization, indicating practical potential for affect-aware consumer electronics systems.

II. Related Work

A. Video-based Affective Computing

VAC aims to understand human emotions from dynamic video sequences by modeling temporal and structural variations in affective expressions. According to the affective representation paradigm, VAC research can be categorized into Discrete and Continuous VAC tasks [24].

Discrete VAC focuses on mapping facial and related cues to a finite set of emotional categories (*e.g.*, happiness, anger, sadness). Early studies relied on handcrafted features [25], which were later replaced by CNN–RNN frameworks integrating spatial extraction and temporal modeling [26]. The emergence of 3D CNNs [27] and Transformer architectures [17], [18] further advanced spatiotemporal representation, enabling models to capture complex dynamics more effectively. In addition, recent works explored multi-instance learning, part-based modeling, temporal attention, and cross-clip aggregation to mitigate sparse evidence issues, where only a few frames provide discriminative cues while the rest are neutral or noisy [16]. Beyond backbone upgrades, recent DFER studies have also leveraged stronger pretraining and fine-grained adaptation to better exploit scarce discriminative frames, *e.g.*, FineCLIPER introduces CLIP-based fine-grained modeling for dynamic expression recognition, highlighting the benefit of richer visual priors under in-the-wild noise [28]. To address practical challenges such as sample heterogeneity, label noise, and domain shift, a growing line of research incorporated robust optimization, uncertainty-aware learning, and weak supervision [20], [29], [30], improving model stability

and generalization. Nevertheless, despite stronger backbones and training recipes, most existing methods still treat affective dynamics as a unified latent process and propagate features through coupled pathways. This design tends to mix slowly-varying affective trends with short-lived bursts, leading to representational aliasing, sensitivity to nuisance perturbations, and limited interpretability when handling rapidly changing affective patterns.

Continuous VAC, based on dimensional emotion theory, projects emotions into a continuous valence–arousal space and models temporal evolution via regression. Mainstream approaches employ recurrent structures or temporal Transformers [24] to capture long-term dependencies and smooth affect trajectories. Recent works further explore alternative long-sequence modeling architectures for continuous VA estimation, such as Mamba-style state space modeling to improve temporal dependency capture under non-stationary affect dynamics [31]. However, continuous VAC is often challenged by non-stationary dynamics: affective states may drift slowly over time, but can also exhibit abrupt transitions triggered by momentary stimuli. Purely sequential modeling without structural decomposition often over-smooths transient shifts or becomes sensitive to short-term noise, resulting in trajectory jitter and optimization instability. Moreover, continuous regression is typically more sensitive to cross-subject and cross-device variability and thus benefits from representations that explicitly preserve global context while remaining responsive to salient local changes. Overall, a key limitation persisting through these advances is the lack of a structured modeling principle that explicitly disentangles stable emotional baselines and transient fluctuations, and further ensures their independent yet coordinated propagation across depth and scale.

B. Low-Rank Sparse Modeling for Emotion Representation Learning

Low-rank sparse modeling originates from high-dimensional statistics and signal decomposition theory, aiming to separate complex observations into a low-rank global structure and a sparse local perturbation for structured information extraction and noise suppression. Robust principal component analysis (RPCA) [32] first formalized this concept by representing data as the sum of a low-rank matrix and a sparse matrix, offering a theoretically grounded decomposition for separating global trends from outliers. Under incoherence conditions, this decomposition ensures identifiability and robustness to noise [33], which makes it a natural choice for modeling signals with stable backgrounds and bursty events. Beyond classical RPCA, subsequent works in sparse representation and structured regularization further extended low-rank and sparsity priors to dynamic settings, where temporal continuity and spatial locality can be jointly exploited for more reliable decomposition.

This principle has been widely applied to video representation learning, motion segmentation, and dynamic scene analysis, showing strong generalization and stability [34]. In these applications, low-rank structures often correspond to slowly changing scene content or consistent motion patterns, while sparse components capture salient events, local motion discontinuities, or rare but informative observations. Recent studies extended this idea to deep models via low-rank attention, low-rank factorization, and sparse graph modeling, which reduce redundancy while preserving expressive power [35].

From a theoretical viewpoint, the low-rank sparse principle naturally aligns with the generative nature of affective signals. Psychostatistical and affect theories suggest that emotional expressions consist of a stable affective baseline and sparse transient bursts [22], corresponding to the low-rank and sparse components, respectively. Affective signals also exhibit temporal sparsity and spatial locality, since emotions are often activated only in a few key frames or localized facial regions [36]. Therefore, low-rank sparse modeling provides an interpretable mechanism to preserve global affective context (stability) while amplifying discriminative short-lived cues (reactivity), rather than forcing a single latent space to explain both. In addition, its inherent robustness can mitigate common nuisances in real-world videos, such as background clutter, illumination variations, and subject-dependent expression styles [34].

Overall, the low-rank sparse principle offers a theoretically grounded and structurally interpretable foundation for VAC, motivating our *LSEF* to explicitly disentangle, propagate, and integrate stable and transient affective dynamics in a unified and robust manner.

III. METHODOLOGY

We propose the *LSEF*, a unified and interpretable framework for video-based affective representation learning. As shown in Fig. 2, *LSEF* extends X3D [37] with a hierarchical pathway from stability encoding to dynamic decoupling, consistency integration, and rank aware optimization. *SEM* first extracts low-rank affective bases that capture invariant emotional structures and suppress noise, forming a coherent global field. Then, *DDM* disentangles sparse temporal perturbations from the stable component, enabling independent propagation of transient affective cues. Next, *CIM* fuses stable and dynamic representations through multiscale consistency alignment, achieving low-rank sparse equilibrium between global stability and local sensitivity. Finally, *RAO* adaptively regulates gradient perturbations according to the rank sparsity balance, which ensures stable and generalizable optimization. Totally, these modules form a theoretically consistent and practically robust framework to jointly modeling steady affective states and different fine-grained temporal fluctuations across various heterogeneous video conditions.

A. Stability Encoding Module

As shown in Fig. 3, the goal of *SEM* is to construct a structurally consistent emotional base by emphasizing low-rank temporal trends while regularizing transient variations, thereby providing a clean foundation for subsequent dynamic decoupling. This module is designed to (i) stabilize affective trajectories by extracting slow-varying components that encode global emotional context, and (ii) preserve informative short-term dynamics without amplifying noise-like high-frequency artifacts. Given the input video feature $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{B\times C\times T\times H\times W}$,

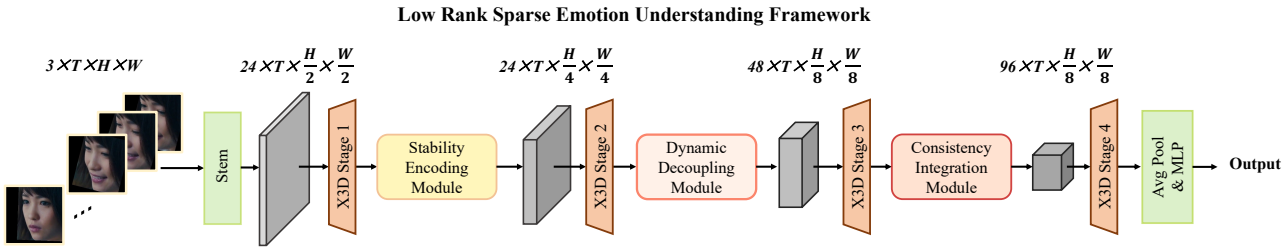


Fig. 2: An overview of the proposed Low-Rank Sparse Emotion Understanding Framework (*LSEF*).

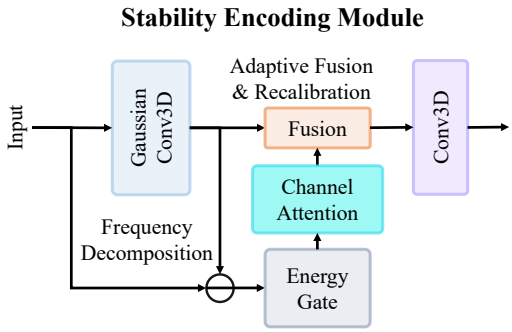


Fig. 3: The Stability Encoding Module.

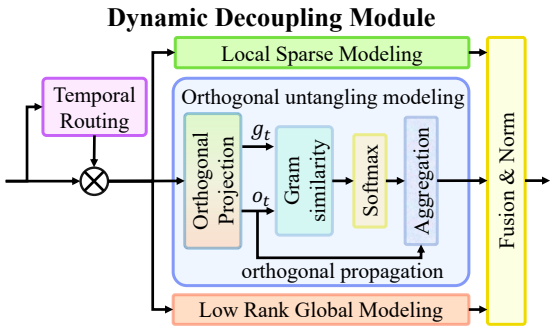


Fig. 4: The Dynamic Decoupling Module.

where B is batch size, C is channel dimension, and (T, H, W) denote temporal length and spatial resolution, we decompose $\boldsymbol{\mathcal{X}}$ into two complementary components:

$$\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}_{\text{low}} + \boldsymbol{\mathcal{X}}_{\text{high}}, \tag{1}$$

where $\boldsymbol{\mathcal{X}}_{\text{low}}$ denotes the low-rank structural component capturing stable emotional trends (long-term tone), and $\boldsymbol{\mathcal{X}}_{\text{high}}$ represents sparse transient perturbations corresponding to short-lived expressive changes. This decomposition provides an explicit separation between stability-oriented and reactivity-oriented signals, making the subsequent modeling stages more controllable and interpretable.

To realize this separation in a differentiable and efficient manner, we perform Gaussian-based frequency decomposition:

$$\boldsymbol{\mathcal{X}}_{\text{low}} = \mathcal{G}_k * \boldsymbol{\mathcal{X}}, \quad \boldsymbol{\mathcal{X}}_{\text{high}} = \boldsymbol{\mathcal{X}} - \boldsymbol{\mathcal{X}}_{\text{low}}, \tag{2}$$

where $\mathcal{G}_k$ is a 3D Gaussian kernel with window size k and $*$ denotes 3D convolution. As a principled low-pass operator, $\mathcal{G}_k$ extracts temporally smooth and spatially coherent tendencies, which align with low-rank affective bases under the assumption that stable emotional context varies slowly and exhibits strong redundancy across frames. The residual branch $\boldsymbol{\mathcal{X}}_{\text{high}}$ captures high-frequency variations, including both meaningful transient affect cues (e.g., brief muscle activations) and nuisance perturbations (e.g., illumination flicker, camera jitter, compression artifacts). Therefore, instead of naively suppressing $\boldsymbol{\mathcal{X}}_{\text{high}}$, we introduce a selective refinement mechanism to retain informative transients while attenuating noise-like fluctuations.

Specifically, we apply energy modulation and channel attention:

$$\tilde{\boldsymbol{\mathcal{X}}}_{\text{high}} = \sigma(\text{CA}(\boldsymbol{\mathcal{X}}_{\text{high}})) \odot \mathcal{E}(\boldsymbol{\mathcal{X}}_{\text{high}}), \tag{3}$$

where $\mathcal{E}(\cdot)$ denotes depthwise energy gating, $\text{CA}(\cdot)$ is channel attention, $\sigma(\cdot)$ is the sigmoid function, and $\odot$ denotes element-wise multiplication. Intuitively, $\mathcal{E}(\cdot)$ performs a lightweight dynamic response filtering: channels/locations with weak or noisy activations are suppressed, while salient responses are preserved to maintain sensitivity to transient affective bursts. Meanwhile, $\text{CA}(\cdot)$ reweights channels according to their affect relevance, encouraging the refined high-frequency branch to focus on semantically meaningful dynamics rather than background variations.

Finally, the low- and high-frequency representations are adaptively fused:

$$\boldsymbol{\mathcal{X}}_{\text{fused}} = \lambda \cdot \boldsymbol{\mathcal{X}}_{\text{low}} + (1 - \lambda) \cdot \tilde{\boldsymbol{\mathcal{X}}}_{\text{high}}, \tag{4}$$

where $\lambda \in (0, 1)$ is a learnable fusion coefficient that controls the balance between low-rank stability and sparse sensitivity. This adaptive fusion is crucial because the relative importance of stable context and transient cues varies across videos, emotion categories, and affect intensity: for slowly evolving affect, a larger λ stabilizes trajectories; for highly dynamic segments, $(1 - \lambda)$ allows transient cues to contribute more effectively.

Through frequency-aware decomposition and selective refinement, *SEM* produces a structurally coherent emotional base that preserves essential affective dynamics while suppressing disruptive noise, providing a reliable foundation for subsequent dynamic decoupling and cross-scale consistency modeling.

B. Dynamic Decoupling Module

Building upon the stable emotional bases provided by the *SEM*, the *DDM* further processes the feature tensor to achieve explicit separation between sparse temporal perturbations and

stable components. As shown in Fig. 4, the key idea is to construct a decoupled propagation pathway: stable bases should remain globally coherent, while transient cues should be selectively routed and propagated without contaminating the stable manifold. Concretely, *DDM* consists of a temporal routing gate that filters and reweights temporal evidence under an information bottleneck principle, and a spatial relational interaction block that enforces approximate orthogonality between propagation subspaces to reduce cross-contamination.

We first construct a temporal routing gating mechanism that achieves adaptive temporal weighting through information bottleneck principles:

$$\boldsymbol{\mathcal{G}}_t = \sigma\left(\mathbf{W}_2 \cdot \delta\left(\mathbf{W}_1 \cdot \mathcal{P}(\boldsymbol{\mathcal{X}})\right)\right), \tag{5}$$

where $\mathcal{P}(\cdot)$ denotes spatial average pooling, which compresses spatial redundancy and yields a compact temporal descriptor that reflects how affective evidence evolves over time. $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{T\times T}$ are fully-connected projection matrices, δ is ReLU, and σ is sigmoid. The gate $\boldsymbol{\mathcal{G}}_t \in (0,1)^{B\times 1\times T\times 1\times 1}$ (after proper reshaping/broadcasting) acts as a soft temporal selector: it suppresses uninformative or noisy frames while emphasizing key moments that carry discriminative transient affective evidence. The temporally routed representation is obtained via Hadamard product: $\tilde{\boldsymbol{\mathcal{X}}} = \boldsymbol{\mathcal{X}} \odot \boldsymbol{\mathcal{G}}_t$. Importantly, this routing is *content-adaptive*: rather than assuming uniform relevance across time, it learns to allocate representation capacity to short-lived affective bursts, which are common in in-the-wild videos and are critical for dynamic affect understanding.

To further decorrelate local semantic patterns from dynamic responses, *DDM* constructs a spatial graph interaction system based on manifold orthogonalization:

$$\hat{\boldsymbol{\mathcal{G}}} = \mathcal{L}_2(\Phi_g(\tilde{\boldsymbol{\mathcal{X}}})), \quad \hat{\boldsymbol{\mathcal{O}}} = \mathcal{L}_2(\Phi_o(\tilde{\boldsymbol{\mathcal{X}}})), \tag{6}$$

$$\boldsymbol{\mathcal{X}}_{\text{graph}} = \hat{\boldsymbol{\mathcal{O}}} \otimes \text{Softmax}\left(\hat{\boldsymbol{\mathcal{G}}}^{\top}\hat{\boldsymbol{\mathcal{O}}}\right), \tag{7}$$

where Φ_g, Φ_o are independent convolutional projection operators and $\mathcal{L}_2$ denotes L2 normalization. Let $N = H \times W$ be the number of spatial positions; the similarity $\hat{\boldsymbol{\mathcal{G}}}^{\top}\hat{\boldsymbol{\mathcal{O}}} \in \mathbb{R}^{N\times N}$ builds a relational map between spatial nodes, and the Softmax normalizes it into propagation weights. Due to L2 normalization, the inner-product similarity effectively measures angular consistency, which encourages stable and transient channels to occupy different directions in the representation manifold. This design enforces approximate orthogonality between graph channels, thereby reducing redundancy and preventing transient activations from being diffusely propagated into the stable pathway (and vice versa). From a geometric standpoint, this yields a more disentangled propagation behavior on the Riemannian manifold induced by normalized features, improving interpretability and robustness.

Finally, *DDM* integrates three complementary feature subspaces through tensor concatenation and convolution:

$$\boldsymbol{\mathcal{X}}_{\text{out}} = \Psi_{1\times 1}\left([\boldsymbol{\mathcal{X}}_{\text{local}}, \boldsymbol{\mathcal{X}}_{\text{graph}}, \boldsymbol{\mathcal{X}}_{\text{global}}]\right), \tag{8}$$

where $\boldsymbol{\mathcal{X}}_{\text{local}} = \Theta_{\text{local}}(\tilde{\boldsymbol{\mathcal{X}}})$ represents local dynamic features preserving fine-grained, spatially localized affect cues, and $\boldsymbol{\mathcal{X}}_{\text{global}} = \Theta_{\text{global}}(\tilde{\boldsymbol{\mathcal{X}}})$ denotes global temporal context features capturing long-range affective dependencies. $\Psi_{1\times 1}$ adaptively fuses them and restores channel capacity. This multi-subspace design provides complementary signals: local captures micro-variations, graph captures relational propagation under orthogonalized constraints, and global maintains context consistency. Through temporal routing, manifold orthogonalization, and multi-subspace fusion, *DDM* progressively decouples stable trends from transient dynamic variations, ensuring independent and interpretable propagation pathways while producing structurally coherent dynamic features for subsequent multi-scale consistency integration.

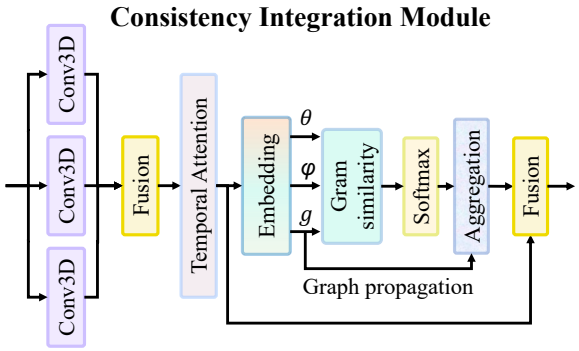


Fig. 5: The Consistency Integration Module.

C. Consistency Integration Module

As shown in Fig. 5, *CIM* aims to achieve cross-scale consistency alignment between the low-rank stable field and sparse dynamic representations based on the decoupling from preceding modules, constructing hierarchical and structurally balanced spatiotemporal affective representations. Therefore, *CIM* explicitly coordinates multi-scale structural cues, temporal consistency, and relational dependency aggregation to maintain a low-rank sparse organization throughout depth.

First, we construct a family of multi-scale depthwise separable operators $\{\Psi_{s_i}\}_{i=1}^{K}$ that capture structural dependencies from fine to coarse granularity through variable scale spatial kernels:

$$\boldsymbol{\mathcal{X}}_{\text{ms}} = \Phi_{\text{fuse}}\left(\bigoplus_{i=1}^{K} \Psi_{s_i}(\boldsymbol{\mathcal{X}})\right), \tag{9}$$

where $\oplus$ denotes channel concatenation and Φ_{fuse} is a fusion convolution. Depthwise separable operators provide an efficient way to enrich receptive fields while controlling computation, which is also favorable for deployment-oriented settings. The multi-scale family $\{\Psi_{s_i}\}$ encourages the representation to preserve both local discriminative patterns (small kernels) and global affective layouts (large kernels), thereby improving the structural completeness of affective cues.

To enhance cross-frame emotional consistency, *CIM* introduces a temporal attention recalibration mechanism based on global affective consistency:

$$\boldsymbol{\mathcal{X}}_{\text{ta}} = \boldsymbol{\mathcal{X}}_{\text{ms}} \odot \sigma\left(\Gamma_{\text{temp}}(\boldsymbol{\mathcal{X}}_{\text{ms}})\right), \tag{10}$$

where $\Gamma_{\text{temp}}(\cdot)$ is a composite operator of temporal pooling and sequential attention mapping. This mechanism acts as a

consistency-aware temporal filter: it strengthens temporally stable affect segments that should dominate the emotional bases, while still allowing sharp transient variations to pass when they are consistently supported by the multi-scale representation. In practice, this helps reduce jitter in predicted trajectories and prevents unstable bursts caused by nuisance perturbations.

Finally, we construct a non-local graph relational modeling operator that achieves cross-scale dependency aggregation through feature projection and relational weighting:

$$\boldsymbol{\mathcal{A}} = \text{Softmax}\left(\frac{\Theta(\boldsymbol{\mathcal{X}}_{\text{ta}})^{\top}\Phi(\boldsymbol{\mathcal{X}}_{\text{ta}})}{\sqrt{C}}\right), \tag{11}$$

$$\boldsymbol{\mathcal{X}}_{\text{out}} = \Gamma_{\text{graph}}(\boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{X}}_{\text{ta}}), \tag{12}$$

where Θ and Φ are projection operators and Γ_{graph} is the graph aggregation operator. The affinity matrix $\boldsymbol{\mathcal{A}}$ establishes long-range relational consistency among spatiotemporal nodes, allowing distant but semantically consistent affect evidence to reinforce each other. This is particularly beneficial when transient cues are spatially localized or temporally short-lived: relational aggregation helps recover coherent semantics by linking them to supportive contexts across space and time. By jointly integrating multi-scale structural cues, global temporal consistency, and relational graph modeling, *CIM* achieves a representational equilibrium between low-rank stability and sparse dynamic sensitivity, yielding a unified affective embedding that is globally coherent yet locally responsive.

D. Rank Aware Optimization

To achieve stable yet adaptive optimization under heterogeneous affective representations, we propose the *RAO*, which constructs a structure-aware perturbation mechanism for robust gradient-based learning. Unlike homogeneous perturbation strategies that use a uniform perturbation radius for all parameters [38], *RAO* modulates the perturbation intensity in a parameter-group-aware manner by explicitly characterizing the structural complexity and sparsity sensitivity of each weight tensor. The underlying intuition is that different parameter groups exhibit different optimization sensitivities under heterogeneous affective dynamics; therefore, imposing the same perturbation magnitude on all tensors may be suboptimal and may introduce unnecessary directional instability.

For each parameter tensor $\boldsymbol{\mathcal{W}}$, we first reshape it into a matrix $\boldsymbol{\mathcal{W}}_m \in \mathbb{R}^{m\times n}$. For convolutional kernels $\boldsymbol{\mathcal{W}} \in \mathbb{R}^{C_{\text{out}}\times C_{\text{in}}\times k_t\times k_h\times k_w}$, we use the matricization $\boldsymbol{\mathcal{W}}_m \in \mathbb{R}^{C_{\text{out}}\times (C_{\text{in}}k_tk_hk_w)}$. Based on $\boldsymbol{\mathcal{W}}_m$, we define a rank-aware structural ratio

$$\rho_r = \frac{\left\|\mathbf{1}\big(\sigma(\boldsymbol{\mathcal{W}}_m) > \tau_r\big)\right\|_1}{\min(m,n)}, \tag{13}$$

where $\sigma(\boldsymbol{\mathcal{W}}_m)$ denotes the singular value vector of $\boldsymbol{\mathcal{W}}_m$, and τ_r is a threshold for identifying effective singular components. To reduce computational overhead, the dominant singular values can be approximated using randomized SVD. We further define a sparsity-sensitive ratio

$$\rho_s = \frac{\left\|\mathbf{1}\big(|\boldsymbol{\mathcal{W}}| < \tau_s\big)\right\|_1}{N(\boldsymbol{\mathcal{W}})}, \tag{14}$$

where $N(\boldsymbol{\mathcal{W}})$ denotes the number of elements in $\boldsymbol{\mathcal{W}}$, and τ_s is the near-zero threshold.

Intuitively, a larger ρ_r indicates that the current tensor retains richer effective structural capacity, for which a moderately larger perturbation can be tolerated to explore flatter local neighborhoods. In contrast, a larger ρ_s implies that the tensor contains more near-zero and potentially brittle entries, for which overly aggressive perturbations may harm optimization stability. Accordingly, for the i-th parameter tensor, we define the structure-aware perturbation radius as

$$\rho_{\text{dyn}}^{(i)} = \rho_{\text{base}}^{(i)}\left(1 + \alpha\rho_r^{(i)} - \beta\rho_s^{(i)}\right), \tag{15}$$

where $\rho_{\text{base}}^{(i)}$ is the base perturbation coefficient for the i-th parameter group, and α, β control the balance between rank-aware perturbation amplification and sparsity-aware perturbation suppression.

Subsequently, *RAO* performs a dual-phase perturb-and-update procedure, as summarized in Alg. 1. Given the clean

Algorithm 1: Rank Aware Optimization

Input: Parameter set $\boldsymbol{\Theta} = \{\boldsymbol{W}^{(i)}\}_{i=1}^{M}$, base optimizer $\mathcal{O}$, base radii $\{\rho_{\text{base}}^{(i)}\}_{i=1}^{M}$, coefficients α, β, γ
Output: Updated parameters $\boldsymbol{\Theta}$

Compute clean loss $\mathcal{L}_{\text{clean}}(\boldsymbol{\Theta})$ and gradients $\{\nabla_{\boldsymbol{W}^{(i)}}\mathcal{L}_{\text{clean}}\}_{i=1}^{M}$;
for $i = 1$ **to** M **do**
  Estimate structural statistics: $\rho_r^{(i)}, \rho_s^{(i)} \leftarrow \Gamma_{\text{struct}}(\boldsymbol{W}^{(i)})$;
  Compute dynamic perturbation radius:
  $$\rho_{\text{dyn}}^{(i)} \leftarrow \rho_{\text{base}}^{(i)}\left(1 + \alpha\rho_r^{(i)} - \beta\rho_s^{(i)}\right)$$
  Normalize gradient:
  $$\boldsymbol{g}^{(i)} \leftarrow \frac{\nabla_{\boldsymbol{W}^{(i)}}\mathcal{L}_{\text{clean}}}{\|\nabla_{\boldsymbol{W}^{(i)}}\mathcal{L}_{\text{clean}}\|_F + \varepsilon}$$
  Construct perturbation:
  $$\boldsymbol{e}^{(i)} \leftarrow \rho_{\text{dyn}}^{(i)} \cdot \boldsymbol{g}^{(i)}$$
Apply joint perturbation:
$$\tilde{\boldsymbol{W}}^{(i)} \leftarrow \boldsymbol{W}^{(i)} + \boldsymbol{e}^{(i)}, \quad i = 1, \ldots, M$$
Compute perturbed loss $\mathcal{L}_{\text{pert}}(\tilde{\boldsymbol{\Theta}})$ and corresponding gradients;
Update parameters with the base optimizer:
$$\boldsymbol{\Theta} \leftarrow \mathcal{O}\big(\boldsymbol{\Theta}, \nabla\mathcal{L}_{\text{pert}}(\tilde{\boldsymbol{\Theta}})\big)$$
Compute perturbation-response factor:
$$\delta \leftarrow \frac{|\mathcal{L}_{\text{pert}} - \mathcal{L}_{\text{clean}}|}{\mathcal{L}_{\text{clean}} + \varepsilon}$$
Optionally adapt the base perturbation radii:
$$\rho_{\text{base}}^{(i)} \leftarrow \rho_{\text{base}}^{(i)}(1 + \gamma\delta), \quad i = 1, \ldots, M$$
return $\boldsymbol{\Theta}$;

loss $\mathcal{L}_{\text{clean}}$, we first compute the gradient with respect to the current parameters and construct a structure-aware perturbation for each parameter tensor. All perturbations are then jointly applied to obtain a perturbed parameter set, on which a perturbed loss $\mathcal{L}_{\text{pert}}$ is evaluated. The final parameter update is performed by the base optimizer using the gradients induced by $\mathcal{L}_{\text{pert}}$. Then, a perturbation-response factor can be introduced as

$$\delta = \frac{|\mathcal{L}_{\text{pert}} - \mathcal{L}_{\text{clean}}|}{\mathcal{L}_{\text{clean}} + \varepsilon}, \tag{16}$$

which measures the relative loss variation under the imposed perturbation and can be used to adapt the perturbation scale across iterations.

In this way, *RAO* integrates structure-aware perturbation with adaptive feedback, so that tensors with richer structural capacity are allowed to explore flatter neighborhoods more actively, while sparsity-sensitive tensors are updated more conservatively. This balancing strategy is particularly suitable for VAC, where heterogeneous affective dynamics may induce non-uniform curvature and optimization sensitivity across parameter groups. Through *RAO*, *LSEF* encourages optimization trajectories that are more compatible with the low-rank sparse structural prior, thereby improving stability and generalization for both categorical and dimensional VAC tasks.

IV. Experiments

A. Experimental Setup

1) Datasets: We conduct experiments on three representative in-the-wild video affective computing benchmarks, including two categorical datasets DFEW [39] and FERV39k [48] and one dimensional dataset VEATIC [12].

DFEW and FERV39k are large-scale categorical datasets collected from real-world movies and diverse scenes, covering seven basic emotion categories, happy, sad, neutral, angry, surprised, disgusted, and fearful. DFEW contains over 16000 clips from more than 1500 films with significant variations in illumination and pose, while FERV39k extends to 38935 clips across four major scenes and 22 subdomains, annotated by 30 professional raters to ensure label reliability. For continuous emotion regression, we adopt VEATIC, a video-based valence–arousal dataset that captures fine-grained temporal variations of affect under unconstrained conditions, enabling evaluation of our framework's generalization from categorical recognition to continuous affect prediction.

2) Metrics: To maintain consistency with prior works, we evaluate categorical emotion recognition using weighted average recall (WAR) and unweighted average recall (UAR), and assess continuous emotion regression with root mean square error (RMSE) and Pearson correlation coefficient (PCC). WAR measures overall recognition accuracy weighted by class frequency, reflecting real world effectiveness under imbalanced emotion distributions, while UAR equally averages recall across all categories to assess balanced performance.

For continuous valence–arousal estimation, RMSE quantifies the deviation between predicted and ground-truth trajectories, where lower values indicate more precise modeling of fine-grained emotional dynamics. In addition, PCC measures the linear correlation between predicted and ground-truth trajectories for each dimension (valence/arousal), complementing RMSE by characterizing temporal trend consistency rather than absolute magnitude error; higher PCC indicates stronger agreement in affective dynamics.

3) Implementation Details: Our entire framework is implemented using PyTorch-GPU and trained on 4 NVIDIA RTX A6000 GPUs. In our experiment, all images are resized to 112×112. The model is trained using AdamW as the base optimizer in combination with a cosine scheduler. The learning rate is set to 1e-4, with a minimum learning rate of 5e-6. For discrete datasets (DFEW and FERV39k), we uniformly sample T=16 frames per clip and train for 100 epochs; for the continuous dataset (VEATIC), we use T=5 and train for 20 epochs. For fair comparison, we adopt the same training configuration across datasets and evaluate all methods under the official splits and metrics.

The *RAO* is used as a perturbation wrapper on top of AdamW. We set $(\alpha, \beta, \gamma) = (0.2, 0.15, 0.1)$ and use $(\tau_r, \tau_s) = (10^{-3}, 10^{-3})$ for rank/sparsity sensitivity computation. The rank ratio is estimated by kernel matricization with a lightweight randomized SVD of rank r=32, and the structural sensitivities are refreshed every K=100 training iterations and cached between refreshes.

B. Comparison with the State-of-the-art Methods

1) Results on Discrete VAC Tasks: We evaluate *LSEF* on two large-scale discrete video affective categorization benchmarks, DFEW and FERV39k, under the official split protocols. As shown in Tab. I and Tab. II, *LSEF* achieves consistent SOTA performance on both datasets across WAR and UAR metrics. Compared with recent transformer-based and hybrid spatiotemporal models, *LSEF* delivers not only stronger overall recognition accuracy (WAR) but also more balanced class-wise performance (UAR), indicating that the proposed hierarchical low-rank sparse modeling alleviates the over-dominance of majority classes and improves robustness under class imbalance. In particular, the gains on UAR suggest that *LSEF* better preserves discriminative cues for minority or visually subtle emotion categories, which are often overwhelmed by dominant coarse affect patterns when representations are learned in a fully entangled manner. This effect is consistent with our motivation: stable emotional bases provide a coherent global context, while sparse transient fluctuations capture short-lived, discriminative bursts; disentangling them prevents transient cues from being washed out by global averaging and prevents stable trends from being corrupted by noisy perturbations.

Overall, these results verify that explicitly modeling affective dynamics as a hierarchical low-rank sparse composition yields more transferable and reliable video emotion understanding than purely monolithic feature learning.

2) Results on Continuous VAC Tasks: We further evaluate *LSEF* on the continuous video affective regression benchmark VEATIC, where frame-level valence–arousal trajectories are estimated. As presented in Tab. III, *LSEF* achieves the lowest RMSE on both valence and arousal dimensions, surpassing the compared continuous VAC baselines. These improvements are

TABLE I: Comparison (%) of our *LSEF* with the SOTA methods on DFEW 5-fd (**Bold**: Best, Underline: Second best).

Method	years	Accuracy of Each Emotion(%)							Metrics (%)		FLOPs (G)
		Hap.	Sad.	Neu.	Ang.	Sur.	Dis.	Fea.	WAR	UAR	
ResNet18+LSTM [39]	ACM MM'20	78.00	40.65	53.77	56.83	45.00	4.14	21.62	53.08	42.86	7.78
EC-STFL [39]	ACM MM'20	79.18	49.05	57.85	60.98	46.15	2.76	21.51	56.51	45.35	8.32
Former-DFER [17]	ACM MM'21	84.05	62.57	67.52	70.03	56.43	3.45	31.78	65.70	53.69	9.11
STT [18]	arXiv'22	87.36	67.90	64.97	71.27	53.10	3.49	34.04	66.45	54.58	N/A
Logo-Former [40]	ICASSP'23	85.39	66.52	68.94	71.33	54.59	0.00	32.71	66.98	54.21	N/A
NR-DFERNet [41]	arXiv'22	88.47	64.84	70.03	75.09	61.60	0.00	19.43	68.19	54.21	6.33
GCA+IAL [42]	AAAI'23	87.95	67.21	70.10	76.06	62.22	0.00	26.44	69.24	55.71	9.63
M3DFEL [16]	CVPR'23	89.59	68.38	67.88	74.24	59.69	0.00	31.63	69.25	56.10	**1.65**
T-MEP [43]	T-CSVT'24	N/A	N/A	N/A	N/A	N/A	N/A	N/A	68.85	57.16	24.70
LG-DSTF [44]	T-MM'24	N/A	N/A	N/A	N/A	N/A	N/A	N/A	69.82	58.89	N/A
CFAN-SDA [45]	T-CSVT'24	**90.84**	**70.91**	65.72	69.97	57.86	13.10	35.36	69.19	57.70	N/A
RDFER [46]	T-BIOM'25	89.69	69.22	**70.18**	71.47	62.08	0.69	28.71	69.73	56.93	N/A
ST-RDGCN [47]	T-AFFC'25	89.57	69.92	62.17	**79.31**	62.24	**20.69**	30.39	69.37	59.18	30.38
LSEF (Ours)	-	89.77	69.93	66.03	76.86	**67.49**	17.93	**39.19**	**71.71**	**61.12**	2.80

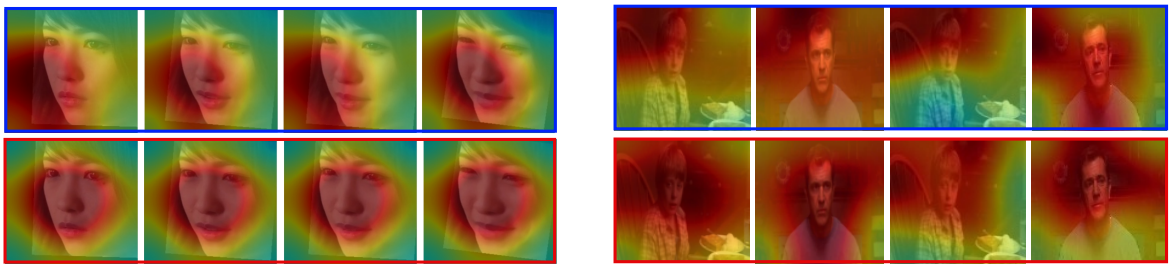


Fig. 6: Visualization of the learned feature maps.

TABLE II: Comparison (%) of our *LSEF* with the SOTA methods on FERV39k.

Method	Metrics (%)	
	WAR	UAR
2C3D [48]	41.77	30.72
ResNet18+LSTM [48]	42.59	30.92
2ResNet18+LSTM [48]	43.20	31.28
VGG13+LSTM [48]	43.37	32.42
2VGG13+LSTM [48]	44.54	32.79
Former-DFER [17]	46.85	37.20
M3DFEL [16]	47.67	35.94
Logo-Former [40]	48.13	38.22
LG-DSTF [44]	48.19	39.84
GCA+IAL [42]	48.54	35.82
RDFER [46]	48.60	36.47
CFAN-SDA [45]	49.48	39.56
ST-RDGCN [47]	49.03	40.48
HDF [20]	50.30	40.49
LSEF (Ours)	**50.73**	**41.26**

particularly notable for continuous VAC because regression is highly sensitive to temporal instability: small errors in modeling transient shifts can accumulate into trajectory drift, while over-smoothing can suppress meaningful affective transitions. In contrast, *LSEF* explicitly separates stable emotional trends (low-rank bases) from short-term affective shifts (sparse fluctuations) and then reconstructs cross-scale coherence through consistency integration, which jointly reduces trajectory noise and preserves salient local changes.

From an affect-dynamics perspective, valence often reflects longer-term mood-related tendencies, while arousal is more reactive to momentary stimuli; the consistent RMSE reduction in both dimensions indicates that *LSEF* can simultaneously maintain global stability and local sensitivity. This is aligned with our framework design: *SEM* stabilizes affective baselines by suppressing noisy high-frequency artifacts, *DDM* prevents transient bursts from contaminating stable representations while preserving discriminative temporal perturbations, and *CIM* coordinates multi-scale aggregation to avoid diluting bases or over-smoothing fluctuations. Overall, *LSEF* provides a unified modeling paradigm capable of delivering robust and interpretable affective understanding across both discrete and continuous VAC scenarios, which is desirable for real-world consumer-facing affect applications requiring stable yet responsive predictions. In addition to RMSE, we report the PCC. The improved overall PCC indicates that *LSEF* not only reduces absolute errors but also better preserves the temporal trend consistency of affect trajectories, mitigating drift while retaining meaningful transitions.

Overall, *LSEF* provides a unified modeling paradigm capa-

TABLE III: Comparison of our *LSEF* with the commonly used Continuous VAC methods on VEATIC.

Method	RMSE			PCC		
	Valence	Arousal	Overall	Valence	Arousal	Overall
EMOTIC [49]	0.3219	0.2645	0.2931	0.6133	0.5906	0.6120
X3D [20]	0.3136	0.2516	0.2826	0.5971	0.5994	0.5983
VEATIC [12]	0.3107	0.2453	0.2780	**0.6250**	0.6012	0.6131
LSEF (Ours)	**0.3094**	**0.2369**	**0.2732**	0.6244	**0.6138**	**0.6191**

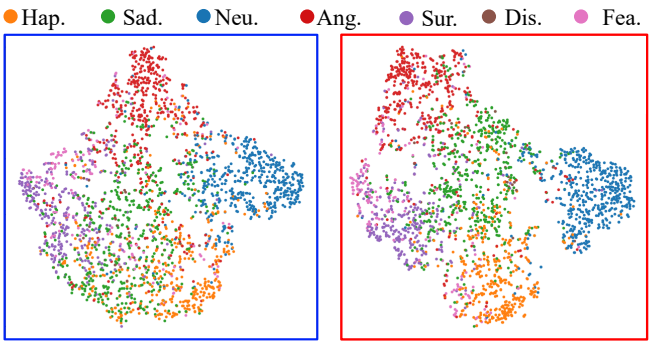


Fig. 7: Illustration of the learned feature distribution.

ble of delivering robust and interpretable affective understanding across both discrete and continuous VAC scenarios, which is desirable for real-world consumer-facing affect applications requiring stable yet responsive predictions.

C. Ablation Studies

1) Effectiveness of Each Module: We conduct ablation studies to assess the contribution of each component in *LSEF*. As shown in Tab. IV, the baseline without any module performs the weakest, highlighting the limitation of learning affective dynamics in a single entangled pathway. Introducing *SEM* significantly improves both discrete and continuous metrics by enhancing low-rank stability and reducing noise-driven fluctuations. This supports our hypothesis that stabilizing long-term emotional bases provides a more coherent contextual field, which benefits both category decision making and continuous trajectory regression.

Adding *DDM* brings further gains by decoupling sparse temporal perturbations and improving the discriminability of transient affect dynamics. Notably, the improvements after *SEM* +*DDM* suggest that stability and reactivity should not be optimized in a coupled manner: when transient fluctuations are routed and orthogonalized, the model becomes better at preserving short-lived discriminative cues without distorting the stable affective baseline. The incorporation of *CIM* provides additional improvements through cross-scale coordination, aligning global stability with local sensitivity. This indicates that hierarchical aggregation, while necessary for semantic abstraction, can introduce structural drift if stable and transient components are not explicitly reconciled across scales; *CIM* helps to maintain a consistent low-rank sparse organization throughout the network depth.

Finally, integrating *RAO* yields the best overall performance by maintaining balanced optimization between smooth low-rank components and sensitive sparse dynamics. This result highlights that the benefit of hierarchical low-rank sparse modeling is not only architectural but also optimization-related: when gradient perturbations are adaptively regulated according to rank/sparsity-sensitive structure, training becomes more stable and generalization improves.

Overall, these observations suggest clear interaction effects among the proposed components: *SEM* first improves temporal coherence by stabilizing the low-frequency emotional base, which creates a more reliable context for subsequent modeling; *DDM* then becomes more effective by isolating and orthogonalizing transient perturbations, reducing stability–sensitivity interference and improving discriminability of short-lived cues; *CIM* further reinforces this benefit by reconciling stable and transient representations across depth to mitigate multi-scale drift introduced by hierarchical aggregation; finally, *RAO* complements the architecture at the optimization level by regulating perturbations according to rank/sparsity-sensitive structure, yielding more stable training and stronger generalization. Together, the modules form a coordinated stability–reactivity pipeline that supports both robust classification and accurate affect regression.

In addition, we summarize the per-module efficiency overhead in Tab. IV. Overall, *SEM*, *DDM*, and *CIM* introduce only lightweight architectural overhead, with modest increases in parameters and computation as modules are progressively enabled. The added cost is dominated by *DDM* due to its relational interaction, while *SEM* and *CIM* remain relatively inexpensive. These results suggest that the accuracy improvements are achieved with a favorable accuracy–efficiency trade-off, and the end-to-end latency remains practically tractable under a consumer electronics-oriented deployment budget.

2) Accuracy and Overhead of RAO: We report the added wall-clock and GPU-memory overhead of *RAO* relative to AdamW and SAM on DFEW under identical settings. Tab. V indicates that perturbation-based training yields consistent accuracy gains over the AdamW baseline, and *RAO* provides the strongest improvement among the compared methods. In terms of cost, *RAO* introduces a moderate increase in wall-clock time compared with AdamW, while the peak GPU memory is comparable to SAM. SAM is generally more time-efficient than *RAO* due to its simpler perturbation rule, but its overall accuracy gain is smaller. These results highlight a clear accuracy–efficiency trade-off that *RAO* attains the best accuracy with small memory overhead at the expense of additional training time.

3) Sensitivity Analysis of Key Hyperparameters: We analyze the sensitivity of *LSEF* to key design choices in *SEM* on DFEW, including the Gaussian kernel size k used for frequency decomposition and the fusion coefficient λ that bal-

TABLE IV: Ablation study of different components in *LSEF* on DFEW (fd5) and VEATIC.

Method	DFEW		VEATIC	Params(M)	FLOPs(G)
	WAR ↑	UAR ↑	RMSE ↓		
base.	68.12	58.21	0.2826	2.98	2.60
+ *SEM*	68.88	59.15	0.2812	2.99	2.64
+ *SEM* + *DDM*	70.72	60.57	0.2787	3.03	2.72
+ *SEM* + *DDM* + *CIM*	71.88	60.86	0.2753	3.09	2.80
+ *SEM* + *DDM* + *CIM* + *RAO* (*LSEF*)	**72.22**	**62.65**	**0.2732**	3.09	2.80

TABLE V: Training overhead comparison of optimizers under identical settings. Time is reported as a relative multiplier normalized by AdamW.

Optimizer	**WAR ↑**	**UAR ↑**	**Time ↓**	**Mem(GB) ↓**
AdamW	71.88	60.86	1.00×	11.51
SAM	72.06	61.43	2.14×	11.55
RAO (ours)	72.22	62.65	2.57×	11.59

TABLE VI: Sensitivity analysis on DFEW.

Factor	**Setting**	**WAR ↑**	**UAR ↑**
SEM kernel size k	$k = 1$	71.84	62.39
	$k = 3$ (default)	72.22	62.65
	$k = 5$	72.09	62.41
Fusion coefficient λ	$\lambda = 0.3$	72.16	62.33
	$\lambda = 0.5$ (default)	72.22	62.65
	$\lambda = 0.7$	72.35	62.57

ances the low-frequency base and high-frequency perturbation-enhanced branch. Note that λ is implemented as a learnable scalar in our *SEM*; thus we evaluate sensitivity to its initialization. As shown in Tab. VI, the performance remains stable within a reasonable range, indicating that *LSEF* is not overly sensitive to these hyperparameters and can maintain consistent accuracy under moderate configuration changes.

D. Visualization.

1) Learned Feature Maps Visualization: To further demonstrate the effectiveness of the proposed *LSEF* (red) compared to the baseline (blue) in modeling both discrete and continuous affective representations, we visualize the learned feature maps on DFEW and VEATIC. As shown in Fig. 6, *LSEF* focuses more precisely on emotion-relevant facial areas such as the eyes and mouth, while effectively suppressing redundant background activations, indicating improved spatial compactness and semantic consistency brought by low-rank structural encoding. Compared with the baseline, *LSEF* exhibits cleaner attention distributions with reduced spurious responses on non-informative regions, which suggests that SEM-driven stabilization helps the model filter out high-frequency nuisance factors (*e.g.*, illumination flicker, slight camera motion, background clutter) that frequently occur in real-world videos. Moreover, *LSEF* produces temporally smoother and spatially coherent activations across consecutive frames, highlighting the model's capability to maintain stable affective dynamics while remaining responsive to short-term expression changes. This behavior is consistent with the low-rank sparse principle: low-rank bases encourage coherent long-term patterns, while sparse transients allow localized bursts to be selectively emphasized rather than diffusely propagated.

In addition, the observed robustness under variations in illumination, pose, and expression intensity provides practical evidence that *LSEF* can maintain stable attention even under capture variability, which is especially relevant to consumer electronics scenarios where cameras, viewpoints, and lighting conditions are highly diverse. Overall, these results confirm that the low-rank and sparse aware mechanism enhances both spatial discriminability and temporal robustness in VAC, providing more reliable and interpretable affective evidence for downstream prediction.

2) t-SNE Visualization: We employ t-SNE [50] to visualize the learned affective embeddings of the baseline (blue) and *LSEF* (red), as shown in Fig. 7. The baseline produces scattered and overlapping clusters, indicating weak structural compactness and limited discriminability among affective states. This suggests that when stable trends and transient cues are entangled, the learned embedding space may encode mixed factors that blur the boundaries between affect categories and degrade regression consistency.

In contrast, *LSEF* generates more distinguishable manifolds, where emotion categories exhibit tighter intra-class cohesion and more distinct inter-class boundaries. The improved clustering pattern implies that *LSEF* reduces representation redundancy and suppresses nuisance variations by organizing affective dynamics into a more structured low-rank (global context) and sparse (local bursts) composition. As a result, samples from the same affect state become more compact even under large appearance diversity, while different affect states are better separated, reflecting stronger dynamic discriminability and structural consistency. These observations qualitatively support our quantitative improvements on both WAR/UAR and RMSE, and further validate that hierarchical low-rank sparse modeling provides a more robust embedding geometry for video-based affective understanding.

V. Discussion

The experimental results validate that explicitly disentangling affective dynamics into a low-rank stable base and sparse transient perturbations provides a fundamental advantage over traditional entangled feature learning. By isolating long-term emotional tones from short-lived reactions, *LSEF* effectively mitigates feature drift and cross-contamination during propagation. This structural decoupling explains its consistent superiority in both discrete categorical recognition and continuous

trajectory estimation under heterogeneous conditions. Despite its robust performance, *LSEF* encounters boundary challenges in extreme in-the-wild scenarios. When affective cues are exceptionally subtle (e.g., micro-expressions) or obscured by severe illumination degradation and heavy occlusions, the mathematical boundary between informative sparse perturbations and environmental noise can blur, potentially mixing vital micro-cues with nuisance artifacts. Moving forward, addressing these limitations requires more adaptive, context-aware routing strategies to salvage subtle perturbations under severe noise. Furthermore, extending this low-rank sparse mechanism to tokenized representations in Vision Transformers, alongside rigorous on-device deployment profiling, will be critical steps toward advancing reliable affective perception for consumer electronics.

VI. Conclusion

We propose *LSEF*, a unified low-rank sparse representation learning framework for VAC. By decomposing affective dynamics into low-rank stable structures and sparse transient variations, *LSEF* achieves a structured representation of both persistent and fluctuating emotional patterns. Through hierarchical modules, *SEM* for stable encoding, *DDM* for dynamic disentanglement, *CIM* for consistency integration, and *RAO* for adaptive optimization, our framework establishes a principled bridge between robustness and plasticity. Extensive experiments on discrete and continuous VAC datasets demonstrate its superior generalization under complex real world conditions. In future, we will extend this paradigm toward multi-modal affective representation learning and efficient deployment for real time emotion analysis.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No. 62302145), Major Scientific and Technological Project of Anhui Provincial Science and Technology Innovation Platform (Grant No. 202305a12020012), and Fundamental Research Funds for the Central Universities (Grant No. JZ2025HGTB0225).

References

[1] J. Li, L. Yang, C. Lv, Y. Chu, and Y. Liu, "Glf-staf: A global-local-facial spatio-temporal attention fusion approach for driver emotion recognition," *IEEE Transactions on Consumer Electronics*, 2025.

[2] W.-Y. Hsu and T.-H. Chiang, "Triple-attribute perceptron facial expression recognition in real-world environments," *IEEE Transactions on Consumer Electronics*, 2025.

[3] D. Li, H. Yang, Z. Song, and Z. Wang, "Msmf-mil: Multi-scale mixed feature-based multiple instance learning for speech emotion recognition," *IEEE Transactions on Consumer Electronics*, 2025.

[4] X. Jin, S. Khan, M. Hosseinzadeh, N. Kumar, and X. Wu, "Mamba-enhanced emotion analysis tinyml models for embedded devices deployment," *IEEE Transactions on Consumer Electronics*, 2025.

[5] M.-C. Su, C.-T. Cheng, M.-C. Chang, and Y.-Z. Hsieh, "A video analytic in-class student concentration monitoring system," *IEEE Transactions on Consumer Electronics*, 2021.

[6] Y. Gu, X. Zhang, H. Yan, J. Huang, Z. Liu, M. Dong, and F. Ren, "Wife: Wifi and vision based unobtrusive emotion recognition via gesture and facial expression," *IEEE Transactions on Affective Computing*, 2023.

[7] M. Li, X. Sun, X. Wang, F. Cui, and X. Yang, "Causalsymptom: Learning causal disentangled representation for depression severity estimation on transcribed clinical interviews," *IEEE Transactions on Affective Computing*, 2025.

[8] Y. Gu, Y. Weng, Y. Wang, M. Wang, G. Zhuang, J. Huang, X. Peng, L. Luo, and F. Ren, "Emotake: Exploring drivers' emotion for takeover behavior prediction," *IEEE Transactions on Affective Computing*, 2024.

[9] X. Zhou, W. Liang, J. She, Z. Yan, and K. I.-K. Wang, "Two-layer federated learning with heterogeneous model aggregation for 6g supported internet of vehicles," *IEEE Transactions on Vehicular Technology*, 2021.

[10] C. Wang, X.-Y. Zhang, P. Guo, H. Wang, C. Mu, and C. Sun, "A lightweight facial expression detection model for elderly depression monitoring on consumer edge devices," *IEEE Transactions on Consumer Electronics*, 2025.

[11] Y. Ji, W. Hu, G. Gan, Z. Long, Y. Zhang, A. Zeng, and H. Zhang, "Smacnet: A unified framework for one-shot talking head synthesis via subtle motion and appearance compensation," *IEEE Transactions on Consumer Electronics*, 2025.

[12] Z. Ren, J. Ortega, Y. Wang, Z. Chen, Y. Guo, S. X. Yu, and D. Whitney, "Veatic: Video-based emotion and affect tracking in context dataset," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[13] X. Ji, Z. Dong, Y. Han, C. S. Lai, G. Zhou, and D. Qi, "Emsn: An energy-efficient memristive sequencer network for human emotion classification in mental health monitoring," *IEEE Transactions on Consumer Electronics*, 2023.

[14] X. Zhang, Y. Lu, H. Yan, J. Huang, Y. Gu, Y. Ji, Z. Liu, and B. Liu, "Resup: Reliable label noise suppression for facial expression recognition," *IEEE Transactions on Affective Computing*, 2025.

[15] M. Khan, J. Ahmad, W. Gueaieb, G. De Masi, F. Karray, and A. El Saddik, "Joint multi-scale multimodal transformer for emotion using consumer devices," *IEEE Transactions on Consumer Electronics*, 2025.

[16] H. Wang, B. Li, S. Wu, S. Shen, F. Liu, S. Ding, and A. Zhou, "Rethinking the learning paradigm for dynamic facial expression recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[17] Z. Zhao and Q. Liu, "Former-dfer: Dynamic facial expression recognition transformer," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021.

[18] F. Ma, B. Sun, and S. Li, "Spatio-temporal transformer for dynamic facial expression recognition in the wild," *ArXiv*, vol. abs/2205.04749, 2022.

[19] R. Wang, J. Huang, J. Zhang, X. Liu, X. Zhang, Z. Liu, P. Zhao, S. Chen, and X. Sun, "Facialpulse: An efficient RNN-based depression detection via temporal facial landmarks," in *ACM Multimedia 2024*, 2024.

[20] F.-Q. Cui, A. Tong, J. Huang, J. Zhang, D. Guo, Z. Liu, and M. Wang, "Learning from heterogeneity: Generalizing dynamic facial expression recognition via distributionally robust optimization," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025.

[21] J. Huang, Y. Feng, F.-Q. Cui, X. Zhang, Z. Liu, X. Liu, J. Liu, F. Zhang, and M. Li, "Identifying who you are no matter what you write through abstracting handwriting style," *IEEE Transactions on Dependable and Secure Computing*, 2026.

[22] R. W. Larson, G. Moneta, M. H. Richards, and S. Wilson, "Continuity, stability, and change in daily emotional experience across adolescence," *Child development*, 2002.

[23] W. Tu, F. Fu, L. Kong, B. Jiang, D. Cobzas, and C. Huang, "Low-rank plus sparse decomposition of fmri data with application to alzheimer's disease," *Frontiers in Neuroscience*, 2022.

[24] U. Bilotti, C. Bisogni, M. De Marsico, and S. Tramonte, "Multimodal emotion recognition via convolutional neural networks: Comparison of different strategies on two multimodal datasets," *Engineering Applications of Artificial Intelligence*, 2024.

[25] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007.

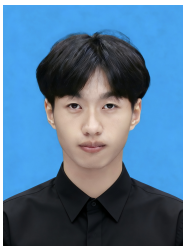
[26] J. Lee, S. Kim, S. Kim, and K. Sohn, "Multi-modal recurrent attention networks for facial expression recognition," *IEEE Transactions on Image Processing*, 2020.

[27] F.-Q. Cui, Z. Lin, X. Rao, A. Tong, S. Li, F. Wang, C. Chen, and B. Liu, "Micacl: Multi-instance category-aware contrastive learning for long-tailed dynamic facial expression recognition," in *2025 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, 2025.

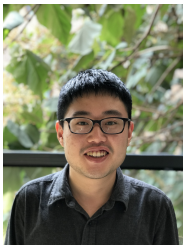
[28] H. Chen, H. Huang, J. Dong, M. Zheng, and D. Shao, "Finecliper: Multi-modal fine-grained clip for dynamic facial expression recognition with adapters," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

[29] H. Wang, X. Mai, Z. Tao, X. Tong, J. Lin, Y. Wang, J. Yu, S. Yan, Z. Zhou, and W. Zhang, "D2sp: Dynamic dual-stage purification framework for dual noise mitigation in vision-based affective recognition," in

2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.
[30] Y. Chen, J. Li, S. Shan, M. Wang, and R. Hong, "From static to dynamic: Adapting landmark-aware image models for facial expression recognition in videos," *IEEE Transactions on Affective Computing*, 2025.
[31] Y. Liang, Z. Wang, F. Liu, M. Liu, and Y. Yao, "Mamba-va: A mamba-based approach for continuous emotion recognition in valence-arousal space," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.
[32] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, 2011.
[33] D. Bertsimas, R. Cory-Wright, and N. A. Johnson, "Sparse plus low rank matrix decomposition: A discrete optimization approach," *Journal of Machine Learning Research*, 2023.
[34] G. Zhang, M. Yu, Y.-J. Liu, G. Zhao, D. Zhang, and W. Zheng, "SparseDGCNN: Recognizing emotion from multichannel EEG signals," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 537–548, 2023.
[35] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *WIREs Data Mining and Knowledge Discovery*, 2024.
[36] K. Lin, X. Wang, Z. Zheng, L. Zhu, and Y. Yang, "Less is more: Sparse sampling for dense reaction predictions," arXiv preprint arXiv:2106.01764, 2021.
[37] C. Feichtenhofer, "X3d: Expanding architectures for efficient video recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
[38] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2021.
[39] X. Jiang, Y. Zong, W. Zheng, C. Tang, W. Xia, C. Lu, and J. Liu, "Dfew: A large-scale database for recognizing dynamic facial expressions in the wild," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.
[40] F. Ma, B. Sun, and S. Li, "Logo-former: Local-global spatio-temporal transformer for dynamic facial expression recognition," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
[41] H. Li, M. Sui, Z. Zhu, and F. zhao, "Nr-dfernet: Noise-robust network for dynamic facial expression recognition," 2022.
[42] H. Li, H. Niu, Z. Zhu, and F. Zhao, "Intensity-aware loss for dynamic facial expression recognition in the wild," *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
[43] X. Zhang, M. Li, S. Lin, H. Xu, and G. Xiao, "Transformer-based multimodal emotional perception for dynamic facial expression recognition in the wild," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[44] Z. Zhang, X. Tian, Y. Zhang, K. Guo, and X. Xu, "Label-guided dynamic spatial-temporal fusion for video-based facial expression recognition," *IEEE Transactions on Multimedia*, 2024.
[45] D. Chen, G. Wen, P. Yang, H. Li, C. Chen, and B. Wang, "Cfan-sda: Coarse-fine aware network with static-dynamic adaptation for facial expression recognition in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[46] F. Liu, H. Wang, and S. Shen, "Robust dynamic facial expression recognition," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2025.
[47] C. Huang, F. Jiang, Z. Han, X. Huang, S. Wang, Y. Zhu, Y. Jiang, and B. Hu, "Modeling fine-grained relations in dynamic space-time graphs for video-based facial expression recognition," *IEEE Transactions on Affective Computing*, 2025.
[48] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang, "Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.
[49] R. Kosti, J. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using emotic dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
[50] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, 2008.



Feng-Qi Cui (Student Member, IEEE) is currently pursuing the Ph.D. degree with the MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition, School of Information Science and Technology, University of Science and Technology of China, Hefei, China. He is also affiliated with Anhui Provincial Key Laboratory of Affective Computing and Advanced Intelligent Machines, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center. His research interests include computer vision and brain-inspired intelligence.



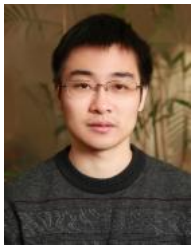
Jinyang Huang (Member, IEEE) is an Associate Professor at the School of Computer Science and Information Engineering, Hefei University of Technology (HFUT). His research interests include Multimodal Perception, Human-computer Interaction, Wireless Security, and Signal Processing. In this area, he has published 60 papers in international peer-reviewed journals and conferences, including ToN, TMC, TIFS, TDSC, MobiCom, IEEE S&P, USENIX Security, Ubicomp, NeurIPS, INFOCOM, and ACM MM. He has served as a TPC member for conferences, including ACM MM, IEEE ICME, and Globecom, and has the honor of becoming ACM MM 2024 Outstanding Reviewers. He is a Guest Editor for the Technical Committee of ICME 25 Special Session. He is the recipient of the Young Scientist of Anhui Computer Federation and IEEE HITC Distinguished PhD Dissertation Award.

Sirui Zhao received the PhD degree with the Department of Computer Science and Technology from University of Science and Technology of China (USTC). He is also a faculty member with the USTC. His research interests include multi-modal analysis, human-computer interaction (HCI) and affect computing. He has published 30+ papers in refereed conferences and journals, including ACM MM, KDD, ICME, IEEE TAFFC, ACM TOMM, etc.

Xinyu Li is currently working toward the PhD degree with the Anhui Provincial Key Laboratory of Affective Computing and Advanced Intelligent Machines, School of Computer Science and Technology, Hefei University of Technology, China. Her research interests include wireless sensing.

Xin Yan received his bachelor's degree in Automation from East China University of Science and Technology. He is currently a Senior algorithm engineer at Beijing XY Robot Technology Co., Ltd. His research interests include automatic multimodal large models, human-computer interaction (HCI), and affective computing.

Ziyu Jia (Member, IEEE) is an Assistant Professor at the Institute of Automation, Chinese Academy of Sciences. His research focuses on time-series analysis methods and their applications in health and medicine, including multimodal affective computing, sleep stage classification, and brain-computer interfaces. He has published over 50 peer-reviewed papers in venues such as IEEE Transactions on Affective Computing, IEEE Transactions on Multimedia, IEEE Transactions on Neural Systems and Rehabilitation Engineering, KDD, and ICLR. Dr. Jia currently serves as an Associate Editor or Editorial Board Member for prestigious journals including IEEE Transactions on Affective Computing and Information Fusion, and he is an Area Chair for major AI and machine learning conferences such as IJCAI and IJCNN. In addition to his academic contributions, Dr. Jia has extensive industry experience, having successfully led multiple R&D projects and secured several patents. He has received numerous honors, including the MSRA StarTrack Award, and the CIE Young Talent Award.



Xiaokang Zhou (M'12, SM'25) is currently an associate professor with the Faculty of Business Data Science, Kansai University, Japan. He received the Ph.D. degree in human sciences from Waseda University, Japan, in 2014. From 2012 to 2015, he was a research associate with the Faculty of Human Sciences, Waseda University, Japan. He was a lecturer/associate professor with the Faculty of Data Science, Shiga University, Japan, from 2016 to 2024. He also works as a visiting researcher in the RIKEN Center for Advanced Intelligence Project (AIP), RIKEN, Japan, since 2017. Dr. Zhou has been engaged in interdisciplinary research in computer science and engineering, information systems, and social and human informatics. His recent research interests include ubiquitous computing, big data, machine learning, behavior and cognitive informatics, cyber-physical-social systems, cyber intelligence, and security. Dr. Zhou is a senior member of the IEEE CS, USA, and a member of the ACM, USA, IPSJ, JSAI, Japan, and CCF, China.