# BioKD: Selective Physiology-to-Video Knowledge Distillation via Reliability Gate for Emotion Recognition

Bojing Hou, Ruohao Li, Yitong Zhu, Hongjun Liu, Luwen Yu, and Yuyang Wang

arXiv:2608.06023v1 [cs.LG] 6 Aug 2026

***Abstract*—To address the limitations of video-based emotion recognition under ambiguous or socially masked behavioral cues, as well as the poor deployability of physiological signals, this paper proposes a reliability-aware physiology-to-video knowledge distillation framework, termed BioKD. The proposed framework leverages physiological signals as privileged information during training to guide a video-based student model in learning deep affective representations, while relying solely on non-intrusive video inputs at inference time. To cope with the high noise and instability of physiological teacher supervision caused by inter-subject variability, signal artifacts, and temporal inconsistency, BioKD incorporates a sample-wise reliability-aware gating mechanism together with a progressive distillation strategy. By adaptively regulating the strength of knowledge transfer, the framework suppresses negative transfer induced by unreliable physiological supervision and enables more stable cross-modal distillation. Experiments on DEAP and AMIGOS show that BioKD consistently outperforms representative baselines under both trial-wise and subject-wise evaluation protocols for valence and arousal recognition. For example, BioKD achieves 68.01% on DEAP (trial-wise arousal) and 65.29% under the more challenging subject-wise setting, demonstrating improved performance under a subject-independent evaluation setting. Further analyses show that BioKD effectively mitigates overconfident teacher errors and outperforms an entropy-only weighting strategy, confirming the importance of explicitly modeling supervision reliability. In addition, BioKD introduces no additional inference-time overhead relative to the same video student architecture and removes the need for physiological sensing and multimodal synchronization.**

***Index Terms*—Affective computing, cross-modal learning, knowledge distillation, physiological signals, video-based emotion recognition.**

## I. Introduction

Emotion recognition is a core problem in affective computing, with broad applications in human–computer interaction, mental health assessment, and intelligent content understanding [1]–[3]. In recent years, substantial progress has been achieved by learning emotion representations from observable behavioral cues (e.g., facial expressions and body movements) benefitting from their accessible data acquisition and efficient deployment [4]–[8]. However, these approaches implicitly assume that observable behaviors can faithfully reflect an individual's internal affective state [9], an assumption frequently violated by emotion regulation and social masking, leading to systematic discrepancies between external behaviors and genuine emotional experiences [10].

To bridge the gap between observable behavioral cues and individuals' underlying internal affective states, physiological signals, such as electroencephalography (EEG), electrodermal activity (EDA), and blood volume pulse (BVP), have been widely explored as complementary sources of affective information, as they capture central and autonomic nervous system activity and provide internal affective cues that complement observable behavior [11]–[14]. Despite their high affective informativeness, physiological signals are inherently noisy, temporally misaligned with emotion labels, and costly or intrusive to acquire, which severely limits their reliability and feasibility as inference-time inputs [15]–[17]. Collectively, these factors render physiological signals a rich yet unreliable modality when used directly as inference-time inputs. Accordingly, how physiological signals should be effectively utilized in emotion recognition models warrants careful reconsideration. Conventional multimodal fusion approaches typically incorporate both physiological and behavioral signals at inference time, inevitably inheriting the aforementioned limitations of physiological inputs [18]–[20]. Recent studies have explored using information-rich but deployment-unfriendly modalities exclusively during training to improve the robustness of primary modalities [21]–[25]. Inspired by this paradigm, we treat physiological signals as training-only supervision, while relying solely on video at inference time to ensure accessibility and deployability.

Knowledge distillation (KD) provides a natural mechanism to realize this paradigm, enabling the transfer of supervision from a physiology-based teacher to a video-based student model [26]. However, when the teacher model itself is built upon unreliable physiological signals, the quality and reliability of its supervisory signals can vary substantially across samples due to noise, instability, and pronounced inter-individual variability. Under such conditions, applying naïve knowledge distillation in a uniform manner may not only fail to improve student performance, but can even introduce misleading supervision and cause negative transfer [27]–[29].

Consequently, the main challenge addressed in this work is how to safely distill affective knowledge from unreliable physiological teachers into deployable video-based models without introducing negative transfer. To address these challenges, we

Bojing Hou, Ruohao Li, Yitong Zhu, Luwen Yu, and Yuyang Wang are with The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China (e-mail: {bhou870, rli777, yzhu162}@connect.hkust-gz.edu.cn; luwenyu@hkust-gz.edu.cn; yuyangwang@hkust-gz.edu.cn).

Hongjun Liu is with New York University, New York, NY, USA (e-mail: janne150.lhj@gmail.com).

Corresponding author: Yuyang Wang

propose **BioKD**, a selective physiology-to-video knowledge distillation framework that adopts a teacher–student paradigm, where a physiology-only teacher provides complementary affective supervision during training. The core of BioKD is a *sample-wise reliability gating mechanism* that models variations in the reliability of physiological teacher supervision and selectively regulates knowledge transfer, distilling informative cues when predictions are stable while suppressing unreliable signals to mitigate negative transfer. Furthermore, because the video student has not yet learned stable representations during early training, BioKD incorporates a *progressive distillation strategy* that gradually introduces and strengthens teacher supervision over training stages, enabling reliability modeling to take effect in a smoother and more stable manner. The overall framework consists of physiological teacher modeling, sample-wise reliability assessment, and selective distillation training, with the training workflow and module interactions illustrated in Fig. 2.

Extensive experiments on DEAP and AMIGOS show that BioKD consistently outperforms representative distillation baselines, demonstrating consistent improvements under both trial-wise and subject-wise evaluation protocols, highlighting the practical effectiveness of reliability-aware physiology-to-video distillation. Overall, the main contributions of this work are summarized as follows:

- We propose **BioKD**, a selective physiology-to-video knowledge distillation framework that repositions physiological signals as a *training-only knowledge source*, improving video-based emotion recognition performance without modifying inference-time inputs or deployment conditions.
- We design a *sample-wise reliability gating mechanism* that explicitly models the reliability variation of physiological teacher supervision, enabling selective knowledge transfer and mitigating negative transfer caused by noisy and unstable physiological signals.
- We systematically integrate a *progressive distillation strategy* into the proposed framework to stabilize optimization when distilling from unreliable physiological teachers, facilitating more effective reliability-aware knowledge transfer.

## II. Related Works

### A. Knowledge Distillation and Cross-Modal Distillation

Knowledge distillation (KD) is a widely studied paradigm for transferring knowledge from high-capacity teacher models to lightweight students, enabling competitive performance with reduced model complexity and inference cost across various domains [26], [30]–[33].

Existing KD methods mainly differ in the form of knowledge transferred from the teacher to the student. Logit-based distillation aligns output probability distributions to convey inter-class relationships [26], while feature-based distillation aligns intermediate representations to transfer richer structural information [34]. Relational distillation, multi-layer distillation, and teacher-assistant strategies have further been proposed to transfer sample relationships, integrate supervision across representation levels, and reduce teacher–student capacity gaps [27], [35], [36]. Despite their success, most KD methods assume that teacher supervision is consistently reliable and therefore apply distillation objectives uniformly. This assumption overlooks cases in which teacher predictions or representations are informative for some samples but inaccurate or misleading for others, potentially introducing negative transfer.

With the development of multimodal learning, KD has been extended to cross-modal settings, where information-rich but deployment-unfriendly modalities supervise unimodal or lightweight student models [37]–[39]. This paradigm reduces inference-time sensing and computational costs while preserving complementary information available during training. Compared with same-modality distillation, cross-modal KD must additionally address substantial differences in input structure, temporal characteristics, and representation distributions between teacher and student modalities. Consequently, direct feature alignment may be unreliable when heterogeneous modalities encode different aspects of the same underlying semantic state.

In affective computing, existing cross-modal distillation methods typically align predictions or intermediate features from multimodal or physiological teachers with unimodal students [40]–[42]. However, they largely focus on what knowledge should be transferred or how heterogeneous representations should be aligned, while paying less attention to whether teacher supervision is trustworthy for each individual sample. This limitation is particularly important when the teacher itself relies on noisy and highly variable physiological measurements.

### B. Reliability Challenges in Physiological Knowledge Distillation

When teacher models are built upon physiological signals or other high-noise modalities, their predictions may exhibit substantial sample-level variability [43], [44]. Physiological signals are susceptible to acquisition noise, temporal misalignment, sensor artifacts, and inter-subject variability, which can produce unstable or even misleading teacher predictions, including overconfident errors [45], [46]. Therefore, the average predictive performance of a physiological teacher does not imply that its supervision is equally reliable for every sample. A teacher may provide useful complementary information for one sample while producing confident but incorrect guidance for another.

Under such conditions, naïve distillation may propagate noisy supervision into the student and cause negative transfer [47], [48]. Uniform distillation is particularly problematic because it assigns similar importance to reliable and unreliable teacher outputs, allowing erroneous supervision to influence student optimization. Although uncertainty modeling, confidence-based weighting, and robust loss functions have been explored to mitigate distillation noise [49], [50], these approaches typically treat prediction uncertainty as a proxy for supervision quality. However, confidence and correctness are

not equivalent: a low-confidence prediction may still contain useful information, whereas an overconfident prediction may remain incorrect.

Under non-stationary physiological noise, this limitation becomes more pronounced because teacher reliability may vary across samples and training stages. Global distillation coefficients or coarse-grained weighting strategies cannot adequately capture these dynamic variations, while decisions based solely on instantaneous confidence may respond strongly to transient fluctuations. Reliable physiology-to-video distillation therefore requires a sample-wise mechanism that considers not only the teacher's current prediction strength but also the historical consistency of teacher–student representations and predictions.

Motivated by these limitations, we study a practical cross-modal distillation setting in which the teacher relies on informative yet potentially unreliable physiological signals. Unlike prior approaches that assume uniformly reliable supervision, we explicitly model sample-wise reliability and selectively regulate the strength of knowledge transfer. By treating physiological signals as training-only privileged information and combining reliability-aware gating with progressive distillation, BioKD aims to preserve useful internal affective cues while suppressing unreliable physiological guidance, enabling more robust cross-modal affective knowledge distillation and video-only inference.

## III. Methodology

### A. The Core Challenge: Reliability Gap in Physiological Teachers

*a) Failure of Conventional Assumptions.:* Given a multimodal dataset $\mathcal{D} = \{(x_i^v, x_i^p, y_i)\}_{i=1}^N$, standard knowledge distillation typically assumes that the prediction outputs of the teacher model provide reliable supervision for the student. However, physiological signals are susceptible to non-stationary noise, including motion artifacts, unstable sensor contact, and inter-subject variability.

Let $\mathbf{s}_i^T = \mathcal{T}(x_i^p)$ denote the prediction logits produced by the physiological teacher. We use the following output-level abstraction:

$$\mathbf{s}_i^T = \mathbf{s}_i^\star + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\mathbf{s}_i^\star$ denotes the latent clean supervision that would be obtained under an ideal acquisition condition, and $\boldsymbol{\epsilon}_i$ summarizes sample-dependent distortions propagated from physiological artifacts to the teacher output. Equation (1) does not assume that noise is physically added to the logits; it provides an output-level abstraction of the unreliable teacher supervision observed by the student.

*b) Overconfident Errors.:* To validate the hypothesis that physiological noise $\epsilon_i$ distorts the teacher's outputs (as formulated in Eq. 1), we analyze the relationship between the teacher's prediction confidence ($\mathrm{conf}_i^T$) and its actual correctness on the validation set. As shown in Fig. 1, the physiological teacher exhibits a clear **confidence–accuracy mismatch**: even in the high-confidence regime (e.g., $\mathrm{conf}_i^T > 0.85$), a noticeable number of samples are still misclassified (highlighted by the red dashed circle). This indicates that high teacher confidence does not necessarily imply trustworthy supervision [45].

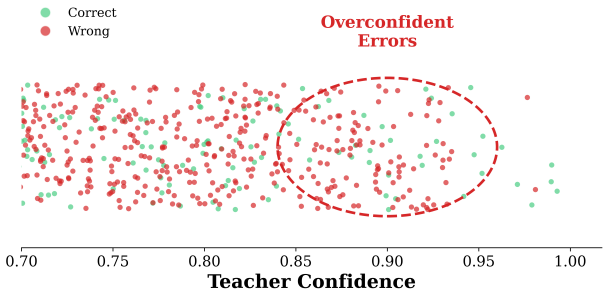


Fig. 1. **Confidence–accuracy reliability analysis of the physiological teacher, highlighting overconfident errors.**

Beyond overconfidence, the reliability of physiological supervision may fluctuate under non-stationary conditions. To address this issue, our framework introduces temporal reliability accumulation, as detailed in Section III-C.

*c) Risk of Negative Transfer.:* These observations indicate a potential risk: ignoring sample-wise reliability variations and directly applying a naïve distillation loss may encourage the student to imitate distorted or misleading teacher supervision, as summarized by the output-level distortion term $\boldsymbol{\epsilon}_i$ in Eq. (1). Because the reliability of physiological teacher outputs varies across samples and acquisition conditions, uniform distillation may propagate teacher errors to the student and result in negative transfer. This finding motivates a mechanism capable of tracking the historical behavior of teacher–student supervision and providing a stabilized reliability estimate for selective knowledge transfer.

### B. Overview of the BioKD Framework

As outlined in Fig. 2, BioKD performs physiology-to-video knowledge distillation in a reliability-aware, progressive manner. During training, a pre-trained physiological teacher provides complementary affective supervision, while a selective reliability gate regulates knowledge transfer based on the estimated reliability of teacher predictions. To stabilize training, a progressive distillation schedule gradually introduces gated supervision.

### C. Selective Reliability Gate

To address reliability fluctuations in physiological teacher supervision, we design the **Selective Reliability Gate** as a closed-loop system that integrates *instantaneous metric modeling*, *temporal aggregation*, and *monotonic gating*. Its objective is to convert noisy alignment signals into robust historical reliability estimates and generate sample-wise, learnable distillation weights for controlled cross-modal feature transfer (as shown in Fig. 2(b)).

*1) From Cross-Modal Alignment to Instantaneous Metrics.:* Due to the manifold discrepancy between the visual representation $\mathbf{z}_i^S$ extracted by the student and the physiological representation $\mathbf{z}_i^T$ extracted by the teacher [51], direct comparison in their original feature spaces may be unreliable. To enable

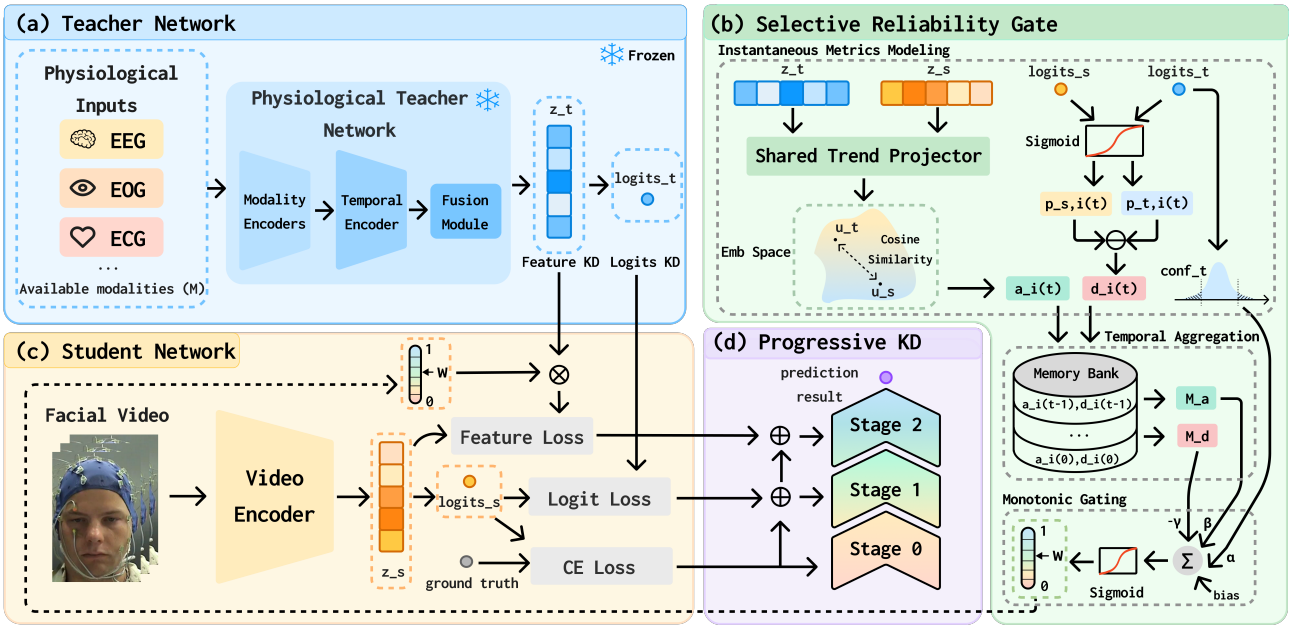


Fig. 2. **Overview of the BioKD framework.** The framework performs physiology-to-video knowledge distillation via reliability-aware gating. (a) A pre-trained physiological teacher provides expert feature supervision. (b) A selective reliability gate dynamically generates the distillation weight $w$ using historical alignment information. (c) A video student learns emotion representations under teacher guidance. (d) A progressive distillation schedule gradually transitions from basic classification to gated feature alignment.

cross-modal comparison, we introduce a shared trend projector $\phi(\cdot)$ that maps both representations into a common semantic space using the same parameters:

$$\mathbf{u}_i^S = \phi(\mathbf{z}_i^S), \qquad \mathbf{u}_i^T = \phi(\mathbf{z}_i^T). \tag{2}$$

Here, $\mathbf{u}_i^S$ and $\mathbf{u}_i^T$ denote the projected student and teacher representations, respectively. Parameter sharing provides a consistent metric space for measuring cross-modal feature agreement.

Within this shared space, we define the **Feature Agreement** $a_i^{(t)}$ to quantify the instantaneous semantic alignment between the teacher and the student:

$$a_i^{(t)} = \frac{1}{2}\left(\frac{\mathbf{u}_i^S \cdot \mathbf{u}_i^T}{\|\mathbf{u}_i^S\|_2 \, \|\mathbf{u}_i^T\|_2} + 1\right), \tag{3}$$

where the output is rescaled to $[0, 1]$ for numerical stability. This metric captures whether the teacher and student representations exhibit a consistent semantic trend at the current training step.

To capture teacher–student conflict at the decision level, we first convert the single-logit outputs into positive-class probabilities:

$$p_i^S = \sigma(s_i^S), \qquad p_i^T = \sigma(s_i^T), \tag{4}$$

where $s_i^S$ and $s_i^T$ denote the student and teacher logits, respectively. The instantaneous prediction disagreement is then defined as

$$d_i^{(t)} = \left|p_i^S - p_i^T\right|. \tag{5}$$

Because $p_i^T$ alone represents the positive-class probability rather than prediction confidence, the teacher confidence is defined symmetrically as

$$\mathrm{conf}_i^T = \max\left(p_i^T, 1 - p_i^T\right). \tag{6}$$

This definition assigns high confidence to decisive predictions of either class while remaining independent of the predicted class identity.

Unlike $a_i^{(t)}$ and $d_i^{(t)}$, this signal does not describe the teacher–student relationship, but instead reflects the teacher's own assessment of sample saliency or discriminability at the current training step. Together, $\{a_i^{(t)}, d_i^{(t)}, \mathrm{conf}_i^T\}$ constitute three complementary instantaneous reliability cues observable at each iteration, where $a_i^{(t)}$ and $d_i^{(t)}$ describe relational quality, and $\mathrm{conf}_i^T$ captures sample saliency.

*2) Time-Scale Separation and Consistency Memory.:* Although these instantaneous metrics are informative, physiological artifacts introduce short-term, non-stationary fluctuations, making direct use of per-step signals unreliable for distillation.

More fundamentally, these signals differ in their temporal characteristics. To validate this observation, we conduct a pilot temporal analysis on the validation set. We find that while the instantaneous agreement $a_i^{(t)}$ and disagreement $d_i^{(t)}$ exhibit noticeable fluctuations, exponential smoothing effectively suppresses high-frequency variations. Specifically, the standard deviation of temporal differences is reduced from $0.0356$ to $0.0049$ for $a_i^{(t)}$ and from $0.0662$ to $0.0070$ for $d_i^{(t)}$ after smoothing. In contrast, the teacher confidence $\mathrm{conf}_i^T$ remains relatively stable at the instantaneous level ($\sigma_\Delta = 0.0373$) and is thus treated as a per-iteration saliency cue.

Motivated by this observation, we perform an explicit **time-scale separation**. Only the structurally meaningful signals $a_i^{(t)}$ and $d_i^{(t)}$ are treated as *trustworthiness cues* and accumulated over time using a *Consistency Memory Bank*, while $\mathrm{conf}_i^T$ is retained as an instantaneous *saliency cue* and directly fed into the gating function. Specifically, for each training sample $i$, the agreement and disagreement memories are updated

independently:

$$\begin{aligned}\mathcal{M}_{\text{agree},i}^{(t)} &= m\mathcal{M}_{\text{agree},i}^{(t-1)} + (1-m)a_i^{(t)},\\ \mathcal{M}_{\text{disag},i}^{(t)} &= m\mathcal{M}_{\text{disag},i}^{(t-1)} + (1-m)d_i^{(t)},\end{aligned} \tag{7}$$

where $t$ denotes the training iteration and $m \in [0, 1)$ is the memory momentum. These updates yield smoothed historical estimates of feature agreement and prediction disagreement without concatenating the two reliability cues. The memory thus yields robust historical estimates of feature agreement $\mathcal{M}_{\text{agree},i}$ and prediction disagreement $\mathcal{M}_{\text{disag},i}$. The consistency memory is maintained exclusively for samples in the training split and is updated only during student optimization. Neither the memory bank nor the reliability gate is required during validation or test inference, and both are discarded after training. Consequently, validation and test samples are never used to update the historical reliability statistics or the parameters of the student model.

*3) Monotonic Trend Gating and Feedback Regulation.:* Given the stabilized historical statistics, the gating module combines the teacher confidence, historical feature agreement, and historical prediction disagreement into an unconstrained reliability score:

$$\begin{aligned}r_i = b &+ \text{Softplus}(\alpha)\,\text{conf}_i^T + \text{Softplus}(\beta)\mathcal{M}_{\text{agree},i}\\ &- \text{Softplus}(\gamma)\mathcal{M}_{\text{disag},i}.\end{aligned} \tag{8}$$

The final sample-wise reliability weight is obtained through

$$w_i = \max\left\{\epsilon,\, \sigma(r_i)\right\}, \tag{9}$$

where $\epsilon$ denotes the minimum gate weight. The Softplus parameterization constrains the scaling coefficients to be non-negative, such that teacher confidence and historical agreement can only increase the gate weight, whereas persistent disagreement can only decrease it.

This formulation enforces an explicit **monotonicity constraint**: both sample saliency and historical agreement can only increase the distillation weight, whereas persistent disagreement always suppresses it. In particular, the disagreement term is designed to identify and mitigate *high-confidence traps*—cases where the teacher exhibits strong confidence but remains consistently misaligned with the student—thereby preventing the propagation of unreliable gradients.

The resulting weight $w_i$ is fed back along the dashed path shown in Fig. 2(b) to modulate the feature distillation loss. For samples with stable and trustworthy teacher supervision, the gate opens to encourage deep feature alignment; for samples affected by physiological noise or persistent conflicts, the gate attenuates gradient flow.

To prevent the overall gate activation from drifting toward a degenerate extreme, we constrain its mini-batch mean using a mean-anchoring regularizer:

$$\mathcal{L}_{\text{gate}} = \left(\frac{1}{B}\sum_{i=1}^{B} w_i - \rho\right)^2, \tag{10}$$

where $\rho = 0.4$ denotes the target mean gate activation. This regularizer controls the global activation level while allowing the monotonic gate to retain sample-specific variations induced by the three reliability cues.

Together, the monotonic cue structure and mean-anchoring constraint enable sample-wise regulation of cross-modal feature transfer while preventing global gate collapse.

### D. Progressive Distillation Strategy

Although the selective reliability gate can dynamically modulate the strength of cross-modal feature distillation, introducing deep feature alignment at the early stage of training may still expose the student to unstable physiological supervision. In physiology-to-video distillation, the student has not yet learned stable facial representations at the beginning of training, and enforcing complex distillation constraints too early may lead to negative transfer.

To address this issue, BioKD adopts a **progressive distillation strategy** that organizes training into three consecutive stages, as illustrated in Fig. 2(d), allowing knowledge transfer to proceed in a coarse-to-fine manner.

*a) Stage 0: Warm-up ($e \le E_0$).:* The student model is trained only with ground-truth supervision, without any teacher guidance, to establish a stable visual representation baseline.

*b) Stage 1: Logits-level Alignment ($E_0 < e \le E_1$).:* The teacher's logits are introduced as soft supervision to perform semantic-level distillation. This stage leverages the cross-modal robustness of logits to guide the student toward learning inter-class semantic relations. Meanwhile, the memory bank in the reliability gate starts tracking teacher–student consistency patterns in the background.

*c) Stage 2: Reliability-driven Feature Refinement ($e > E_1$).:* Feature-level distillation is activated under the control of the reliability gate. The sample-wise distillation weight $w_i$ dynamically adjusts the strength of feature alignment, ensuring that deep alignment is emphasized only when the teacher is considered reliable.

*d) Distillation Objectives.:* Let $s_i^S$ and $s_i^T$ denote the single-logit outputs of the student and teacher, respectively, and let $y_i \in \{0, 1\}$ denote the binary emotion label. The task loss is implemented using binary cross-entropy with logits:

$$\mathcal{L}_{\text{task}} = \frac{1}{B}\sum_{i=1}^{B} \text{BCELogits}(s_i^S, y_i), \tag{11}$$

where $B$ denotes the mini-batch size.

For logit-level distillation, we first construct label-conditioned margins:

$$m_i^S = (1-2y_i)s_i^S, \qquad m_i^T = (1-2y_i)s_i^T. \tag{12}$$

Their temperature-scaled probabilities are defined as

$$\begin{aligned}\widetilde{p}_i^S &= \text{clip}\big(\sigma(m_i^S/\tau),\, \delta,\, 1-\delta\big),\\ \widetilde{p}_i^T &= \text{clip}\big(\sigma(m_i^T/\tau),\, \delta,\, 1-\delta\big).\end{aligned} \tag{13}$$

where $\tau$ is the distillation temperature, $\delta = 10^{-6}$ ensures numerical stability, and $\sigma(\cdot)$ denotes the Sigmoid function. The logit-level distillation loss is then

$$\mathcal{L}_{\text{logit}} = \frac{\tau^2}{B}\sum_{i=1}^{B} \text{BCE}\left(\widetilde{p}_i^S, \text{sg}(\widetilde{p}_i^T)\right), \tag{14}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation, and the second argument of $\mathrm{BCE}(\cdot,\cdot)$ is treated as the target.

The sample-wise feature discrepancy is computed in the shared projection space:

$$\ell_{\mathrm{feat}}^{(i)} = \frac{1}{d}\left\|\mathbf{u}_i^S - \mathbf{u}_i^T\right\|_2^2, \tag{15}$$

where $d$ denotes the feature dimensionality. The reliability-weighted feature distillation loss is defined as

$$\mathcal{L}_{\mathrm{feat}} = \frac{\sum_{i=1}^{B} \mathrm{sg}(w_i)\,\ell_{\mathrm{feat}}^{(i)}}{\sum_{i=1}^{B} w_i + \delta}. \tag{16}$$

Here, $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. It is applied to the gate weights in the numerator, preventing the gate from trivially suppressing samples with large feature discrepancies to reduce the distillation loss. The normalization denominator remains differentiable and does not induce a downward collapse of the gate weights. Meanwhile, the mean-anchoring regularizer in Eq. (10) constrains the average gate activation around $\rho$, preventing the gate from drifting toward either uniformly small or uniformly large values.

*e) Progressive Optimization Objective.:* The overall training objective evolves with the training stage and is defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{task}} &+ \mathbb{I}(e > E_0)\lambda_l \mathcal{L}_{\mathrm{logit}} \\ &+ \mathbb{I}(e > E_1)\left[\lambda_f \mathcal{L}_{\mathrm{feat}} + \lambda_g \mathcal{L}_{\mathrm{gate}}\right],\end{aligned} \tag{17}$$

where $e$ denotes the current epoch, $E_0$ and $E_1$ are the stage-transition epochs, and $\mathbb{I}(\cdot)$ is the indicator function. The coefficients $\lambda_l$, $\lambda_f$, and $\lambda_g$ control the logit distillation, feature distillation, and gate regularization terms, respectively.

## IV. Experiments and Results

### A. Experimental Setup

We evaluate BioKD on two widely used multimodal emotion recognition benchmarks, DEAP [11] and AMIGOS [52]. DEAP contains physiological signals and facial videos from 32 subjects, while AMIGOS provides multimodal recordings from 40 subjects; the short-video subset is used in our experiments.

Valence and arousal labels are binarized into high and low classes using a threshold of 5. We consider two data partition protocols. Under the **trial-wise** protocol, complete trials are assigned to mutually exclusive training, validation, and test sets with a ratio of 8:1:1. Under the **subject-wise** protocol, subjects are assigned to mutually exclusive splits using the same ratio, ensuring no subject overlap across the three sets.

To prevent temporal leakage, dataset partitioning is performed before window extraction. After the splits are determined, physiological signals and video streams are independently segmented within each split using a sliding window of 4 seconds with a stride of 2 seconds. Consequently, overlapping or adjacent windows from the same trial never appear in different splits.

Under both partition protocols, all experiments are repeated over 3 random seeds to assess robustness and result stability. We report accuracy and F1-score as mean±standard deviation over three random seeds.

TABLE I
Architectural configuration of the physiological teacher and video student. The teacher is used only during training.

| Branch | Component | Output | Status |
|---|---|---|---|
| Teacher | Modality encoders | $M \times d_m$ | Frozen |
| Teacher | Temporal encoder | $M \times d_t$ | Frozen |
| Teacher | MoE fusion | 128 | Frozen |
| Teacher | Classifier | 1 | Frozen |
| Student | MobileNetV2 | $T \times d_v$ | Trainable |
| Student | GRU | $d_s$ | Trainable |
| Shared | Trend projector | 128 | Training only |
| Student | Classifier | 1 | Trainable |

To examine robustness under different levels of privileged information, we evaluate multiple teacher modality configurations. On DEAP, teacher models range from EEG-only to full multimodal physiological settings. On AMIGOS, teacher models are likewise expanded from EEG-only to full multimodal physiological settings. All experiments are conducted for both valence and arousal tasks under the same protocol.

*a) Network Architectures:* We adopt an asymmetric teacher–student architecture, whose detailed configuration is summarized in Table I. The physiological teacher takes synchronized physiological windows as input and employs modality-specific encoders, temporal modeling, and mixture-of-experts fusion to obtain an affective representation $\mathbf{z}_i^T$ and the corresponding prediction logits. The teacher is initialized with pre-trained weights and remains frozen throughout distillation. The video student receives only facial video frames without audio input, since speech may be absent during passive viewing and recorded audio may be dominated by stimulus or environmental sound. MobileNetV2 [53] extracts frame-level visual features, which are subsequently aggregated by a GRU to obtain the video representation $\mathbf{z}_i^S$. A shared projection head $\phi(\cdot)$ maps both teacher and student representations into a common 128-dimensional space for feature-level distillation. The projection head is used only during training and is discarded during inference. The student is trained from scratch and is the only network retained during inference. Unless otherwise specified, MobileNetV2–GRU is used as the default student architecture; VideoMAE is evaluated separately in Section IV-D.

*b) Implementation Details:* All models are implemented in PyTorch and trained on an RTX A6000 GPU. We use Adam with a learning rate of $3 \times 10^{-4}$ and a batch size of 16. The memory momentum is set to $m = 0.92$, and the minimum gate weight is set to $\epsilon = 0.01$. We adopt a three-stage progressive training schedule over 10 epochs: (i) task-only learning in epochs 1–2, (ii) logit-level distillation in epochs 3–5, and (iii) full reliability-aware feature distillation in epochs 6–10.

*c) Baselines:* To validate the effectiveness of the proposed BioKD framework, we reproduce several representative knowledge distillation methods as baselines under a cross-modal experimental setup, including KD [26], Fitnets [34], VID [54], RKD [35], EmotionKD [40], C2KD [55] and GRCGD [56], as shown in Table III and Table II. In addition, a

TABLE II
QUANTITATIVE COMPARISON UNDER THE TRIAL-WISE PROTOCOL ON DEAP AND AMIGOS. ENTRIES REPORT ACCURACY AND F1-SCORE AS MEAN±STANDARD DEVIATION (%). BOLD INDICATES THE BEST RESULT AMONG VIDEO-BASED STUDENT MODELS.

| Method | DEAP | | | | AMIGOS | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | | F1-Score | | Accuracy | | F1-Score | |
| | Valence | Arousal | Valence | Arousal | Valence | Arousal | Valence | Arousal |
| *Baselines* | | | | | | | | |
| Physiological Teacher (EEG) | 61.48±4.51 | 64.49±5.57 | 62.84±6.21 | 64.87±8.06 | 54.33±6.34 | 65.66±7.18 | 59.86±6.71 | 64.51±7.65 |
| Video Student (No KD) | 70.87±3.99 | 65.14±5.03 | 72.69±5.24 | 72.14±6.94 | 55.42±5.71 | 60.63±6.72 | 58.28±6.03 | 63.55±7.15 |
| *Distillation Methods* | | | | | | | | |
| KD [26] | 68.38±5.31 | 53.10±7.11 | 69.21±5.67 | 56.82±7.68 | 54.07±6.02 | 61.08±7.34 | 56.11±6.28 | 62.40±7.82 |
| FitNets [34] | 72.98±6.13 | 51.96±8.27 | 74.18±6.48 | 57.91±8.81 | 55.65±6.44 | 59.33±8.54 | 58.76±6.72 | 60.07±9.06 |
| VID [54] | 71.98±5.62 | 55.80±7.62 | 71.62±5.97 | 58.06±8.18 | 57.91±5.89 | 60.32±7.85 | 59.24±6.14 | 64.01±8.34 |
| RKD [35] | 65.44±6.57 | 64.69±9.12 | 67.86±6.91 | 66.02±9.68 | 53.11±6.83 | 53.03±9.34 | 56.90±6.96 | 56.97±9.81 |
| EmotionKD [40] | 71.55±4.78 | 55.60±6.52 | 72.36±5.12 | 60.36±7.01 | 57.48±5.02 | 61.50±6.73 | 60.04±5.38 | 63.13±7.26 |
| C2KD [55] | 69.55±4.61 | 54.11±6.38 | 71.84±4.94 | 55.42±6.84 | 51.90±5.15 | 61.11±6.57 | 54.62±5.44 | 64.86±7.03 |
| GRCGD [56] | 72.53±5.08 | 53.45±6.91 | 73.46±5.46 | 56.79±7.43 | 57.46±5.34 | 61.31±7.12 | 61.02±5.71 | 62.08±7.58 |
| **BioKD (EEG → Video) (Ours)** | **74.00±3.16** | **68.01±5.18** | **76.74±3.32** | **74.60±4.86** | **59.32±3.58** | **64.63±4.51** | **63.85±3.96** | **69.10±5.03** |

TABLE III
QUANTITATIVE COMPARISON UNDER THE SUBJECT-WISE PROTOCOL ON DEAP AND AMIGOS. ENTRIES REPORT ACCURACY AND F1-SCORE AS MEAN±STANDARD DEVIATION (%). BOLD INDICATES THE BEST RESULT AMONG VIDEO-BASED STUDENT MODELS.

| Method | DEAP | | | | AMIGOS | | | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | | F1-Score | | Accuracy | | F1-Score | |
| | Valence | Arousal | Valence | Arousal | Valence | Arousal | Valence | Arousal |
| *Baselines* | | | | | | | | |
| Physiological Teacher (EEG) | 51.94±8.21 | 59.06±10.42 | 54.54±8.77 | 65.01±11.36 | 55.21±6.18 | 66.98±8.85 | 60.30±6.92 | 71.88±9.47 |
| Video Student (No KD) | 50.74±7.11 | 56.43±9.24 | 53.01±7.84 | 62.58±10.07 | 51.20±6.52 | 55.71±8.31 | 55.12±7.08 | 59.36±8.74 |
| *Distillation Methods* | | | | | | | | |
| KD [26] | 53.34±8.03 | 56.28±9.68 | 57.16±8.41 | 61.94±10.22 | 54.86±6.76 | 55.36±7.69 | 58.31±7.25 | 59.02±8.69 |
| FitNets [34] | 51.24±9.27 | 52.89±11.48 | 54.88±9.81 | 58.47±12.06 | 44.07±8.36 | 58.73±9.72 | 49.35±8.92 | 61.24±10.41 |
| VID [54] | 54.09±8.62 | 51.08±10.73 | 56.72±9.14 | 57.36±11.39 | 56.70±7.42 | 58.19±9.18 | 60.18±8.03 | 62.73±9.86 |
| RKD [35] | 50.10±9.64 | 57.13±12.21 | 53.42±9.93 | 63.08±12.68 | 50.33±8.71 | 57.48±10.34 | 54.69±8.88 | 61.81±10.82 |
| EmotionKD [40] | 57.74±7.28 | 61.07±9.85 | 61.29±7.91 | 66.84±10.36 | 47.20±7.13 | 59.59±8.64 | 51.68±7.67 | 62.16±9.21 |
| C2KD [55] | 54.27±8.11 | 59.30±10.06 | 58.65±8.55 | 62.11±10.69 | 51.42±6.89 | 58.54±8.97 | 55.06±6.88 | 62.82±9.54 |
| GRCGD [56] | 54.07±7.83 | 60.23±10.51 | 56.31±8.28 | 65.42±11.08 | 50.59±7.54 | 57.26±8.76 | 54.92±7.96 | 60.14±9.33 |
| **BioKD (EEG → Video) (Ours)** | **59.13±7.42** | **65.29±9.16** | **70.69±8.08** | **76.91±9.73** | **57.05±6.63** | **60.95±7.82** | **60.65±7.14** | **63.55±8.31** |

Student Only model trained without any teacher supervision is included as a baseline. Unless otherwise specified, the teacher model used for distillation is trained on EEG signals only. The impact of different teacher modality configurations is studied separately in Table XI.

For the calibration analysis, we additionally include Similarity-Preserving Knowledge Distillation (SP) [57] and Neuron Selectivity Transfer (NST) [58].

### *B. Comparison Study*

Tables II and III report the experimental results on DEAP and AMIGOS under the trial-wise and subject-wise protocols, respectively. Both Accuracy and F1-score are reported as mean±standard deviation, and the EEG-based physiological teacher is used for all distillation methods. Overall, BioKD achieves the highest mean classification accuracy among the evaluated video-based student models across all datasets, emotion dimensions, and evaluation protocols. Notably, under the more challenging subject-wise protocol, BioKD continues to provide clear mean improvements, indicating improved robustness under subject-independent evaluation.

Compared with the best competing methods under each setting, BioKD maintains consistent gains in mean accuracy. For example, on DEAP trial-wise arousal, BioKD improves the mean accuracy from RKD's 64.69% to **68.01%**. On AMIGOS trial-wise arousal, BioKD achieves **64.63%**, compared with 61.50% for EmotionKD. Under the more challenging DEAP subject-wise arousal setting, BioKD reaches **65.29%**, compared with 61.07% for EmotionKD. In addition, BioKD consistently improves over the Video Student (No KD) baseline under both evaluation protocols, indicating that physiological teacher supervision can enhance video-based emotion recognition.

Examining the behavior of different distillation methods reveals that traditional knowledge distillation techniques exhibit varying levels of effectiveness in the considered **physiology-to-video distillation** setting. Standard KD and some feature-alignment-based methods even lead to performance degradation in certain cases, suggesting that when the teacher model is built upon noisy physiological signals with substantial inter-subject variability, applying uniform distillation constraints may propagate unreliable supervision to the student model. In contrast, methods such as FitNets, VID, EmotionKD, C2KD, and GRCGD achieve relatively better performance in some tasks, indicating that feature transfer, sample selection, or relational modeling can partially alleviate cross-modal discrepancies. However, these approaches generally do not explicitly

model the **sample-wise reliability** of the teacher supervision, causing their improvements to depend heavily on specific datasets, tasks, or evaluation protocols. In comparison, BioKD introduces a reliability-aware gating mechanism to adaptively regulate the distillation strength at the sample level, enabling more stable and effective cross-modal knowledge transfer across datasets and evaluation settings.

### C. Ablation Study

Table IV presents the ablation results for the DEAP arousal task under the subject-wise protocol, where all variants are distilled from the same EEG-based teacher for a fair comparison. The student-only baseline already provides a reasonable starting point, while naïve logits distillation leads to a performance drop, confirming that directly transferring supervision from noisy physiological teachers can be harmful.

Compared with logits KD, naïve feature distillation improves performance, suggesting that feature-level guidance provides a more informative and transferable supervision signal in the considered cross-modal setting. Introducing the reliability-aware gating mechanism further improves the results, indicating the importance of selectively filtering teacher supervision according to sample-wise reliability. Removing the memory component causes a noticeable degradation, showing that temporal reliability accumulation helps stabilize the gating behavior across samples. Likewise, removing the progressive training strategy leads to a clear performance drop, demonstrating that gradually introducing teacher supervision is important for stable optimization.

The component-wise gate ablations further show that different reliability cues contribute differently to the final performance. In particular, removing the agreement reward or the confidence term causes larger degradation, suggesting that these two factors play a critical role in identifying trustworthy supervision. Removing the disagreement penalty also reduces performance, although to a relatively smaller extent. Overall, the full BioKD model achieves the best result, validating the effectiveness and complementarity of all components in the proposed framework.

Beyond the component-wise ablations, we further examine two implementation choices in BioKD: the distance function used to measure prediction disagreement and the objective to which the reliability gate is applied. The results are summarized in Table V. As shown in Table V, the default L1 disagreement measure outperforms the L2 and cosine alternatives, suggesting that absolute probability differences provide a more stable decision-level conflict signal in the considered cross-modal setting. Applying the gate only to feature-level distillation achieves the best performance, whereas applying it to logit-level distillation or both objectives leads to lower results. This finding indicates that deep cross-modal feature alignment is particularly sensitive to unreliable physiological supervision, while logit-level alignment is comparatively less dependent on sample-wise gating.

### D. Generalization across Student Backbones

To examine whether the effectiveness of BioKD depends on the relatively lightweight MobileNetV2–GRU student, we further instantiate the video student using VideoMAE. For each backbone, we compare student-only training, standard knowledge distillation, and BioKD under the same data splits and evaluation protocol. As shown in Table VI, BioKD consistently improves both student architectures, indicating that its benefit arises from reliability-aware knowledge transfer rather than a particular video backbone.

TABLE IV
ABLATION STUDY OF BIOKD COMPONENTS ON DEAP (AROUSAL) UNDER THE SUBJECT-WISE PROTOCOL. ALL VARIANTS ARE DISTILLED FROM THE SAME EEG-BASED PHYSIOLOGICAL TEACHER.

| Variant | Stage | $\mathcal{L}_{KD}$ | $\mathcal{L}_{Feat}$ | Gate | Mem | Acc |
|---|---|---|---|---|---|---|
| Student Baseline | | | | | | 56.43 |
| Logits KD | ✓ | ✓ | | | | 50.28 |
| Naïve Feature KD | ✓ | ✓ | ✓ | | | 52.91 |
| w/o Memory | ✓ | ✓ | ✓ | ✓ | | 63.43 |
| w/o Stage | | ✓ | ✓ | ✓ | ✓ | 62.36 |
| w/o Agreement Reward | ✓ | ✓ | ✓ | ✓ | ✓ | 56.88 |
| w/o Disagreement Penalty | ✓ | ✓ | ✓ | ✓ | ✓ | 62.59 |
| w/o Confidence Term | ✓ | ✓ | ✓ | ✓ | ✓ | 57.55 |
| **BioKD (Full)** | ✓ | ✓ | ✓ | ✓ | ✓ | **65.29** |

TABLE V
ABLATION OF DISAGREEMENT METRICS AND GATE PLACEMENT ON DEAP (AROUSAL) UNDER THE SUBJECT-WISE PROTOCOL.

| Design | Variant | Acc. (%) | F1-Score (%) |
|---|---|---|---|
| Disagreement | L1 (default) | **65.29** | **76.91** |
| | L2 | 52.71 | 63.54 |
| | Cosine | 61.90 | 72.47 |
| Gate placement | No gate | 54.16 | 66.61 |
| | Logit only | 62.99 | 73.49 |
| | Feature only | **65.29** | **76.91** |
| | Both | 60.85 | 71.80 |

### E. In-depth Analysis of Distillation Mechanism

*1) Gate Reliability across Correctness Quadrants:* We evaluate whether the learned gate reflects teacher correctness and potential per-sample transfer utility rather than merely teacher–student agreement. Since the consistency memory is defined for training samples, this diagnostic analysis is con-

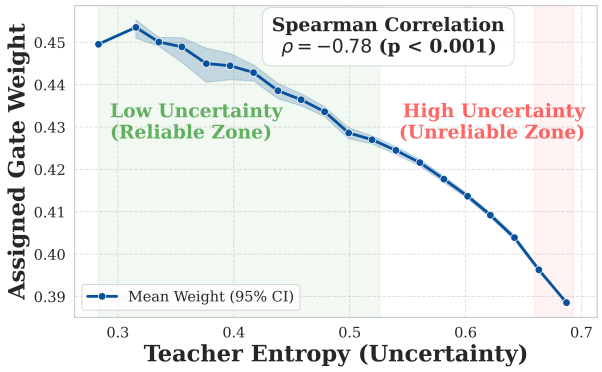


Fig. 3. Relationship between teacher prediction entropy and the learned gate weight. The curve and shaded region denote the mean and 95% confidence interval, respectively.

TABLE VI
GENERALIZATION OF BIOKD ACROSS DIFFERENT VIDEO STUDENT BACKBONES ON DEAP (AROUSAL) UNDER THE SUBJECT-WISE PROTOCOL.

| Backbone | Training | Acc | F1-Score |
|---|---|---|---|
| MobileNetV2–GRU | Video Student (No KD) | 56.43 | 62.58 |
| | Standard KD | 56.28 | 61.94 |
| | BioKD | **65.29** | **76.91** |
| VideoMAE | Video Student (No KD) | 59.63 | 61.35 |
| | Standard KD | 63.07 | 64.66 |
| | BioKD | **66.78** | **70.73** |

TABLE VII
GATE WEIGHTS ACROSS TEACHER–STUDENT CORRECTNESS QUADRANTS. $T$, $S$, AND $Y$ DENOTE THE PHYSIOLOGICAL TEACHER PREDICTION, VIDEO STUDENT (NO KD) PREDICTION, AND GROUND-TRUTH LABEL, RESPECTIVELY.

| Subset | Definition | Samples | Mean gate |
|---|---|---|---|
| Both correct | $T = Y,\ S = Y$ | 8,348 | 0.4445 |
| Teacher helpful | $T = Y,\ S \neq Y$ | 3,761 | 0.4274 |
| Teacher harmful | $T \neq Y,\ S = Y$ | 3,574 | 0.3911 |
| Both wrong | $T \neq Y,\ S \neq Y$ | 5,110 | 0.3984 |

ducted exclusively on the training split. For each sample, we record the gate weights produced during Stage 2 and average them across its training visits. The ground-truth labels are used only for post-hoc analysis and do not directly supervise the gate.

We then partition the samples into four correctness quadrants according to the physiological teacher prediction, the independently trained Video Student (No KD) prediction, and the ground-truth label. This partition distinguishes teacher-helpful cases, where the teacher is correct and the video student is wrong, from teacher-harmful cases, where the teacher is wrong and the video student is correct. Validation and test samples are not used to construct the memory statistics or the correctness quadrants in this analysis.

As shown in Table VII, both-correct samples receive the highest average gate weight, while teacher-harmful samples receive the lowest. Teacher-helpful samples also receive relatively high weights, indicating that BioKD preserves physiological supervision when it has the potential to correct the video student. Aggregating the four quadrants, teacher-correct samples receive higher average gate weights than teacher-wrong samples (0.4392 vs. 0.3954), and the gate distinguishes teacher correctness with an AUROC of 0.7021. These results show that the learned gate captures supervision utility beyond teacher–student agreement alone.

*2) Correction of Overconfident Teacher Errors:* To further examine whether BioKD can mitigate the overconfident errors highlighted in Fig. 1, we collect sample-level predictions from the teacher, Video Student (No KD), standard KD, and BioKD on the same test set. Specifically, we define overconfident wrong teacher samples as those on which the teacher prediction is incorrect and the prediction confidence, computed as the maximum posterior probability, exceeds 0.85. For this subset, we evaluate whether each student model can correct the teacher's error, i.e., whether the student prediction matches the ground-truth label. As shown in Table VIII, among the 76 overconfident wrong teacher samples, BioKD successfully corrects 30 cases (39.47%), outperforming both Video Student (No KD) (25.00%) and standard KD (18.42%). This result suggests that the proposed reliability-aware gating mechanism can effectively suppress misleading teacher supervision and reduce the tendency of the student to inherit overconfident teacher errors.

TABLE VIII
**CORRECTION OF OVERCONFIDENT WRONG TEACHER PREDICTIONS.** RESULTS ARE REPORTED ON DEAP (AROUSAL) UNDER THE SUBJECT-WISE PROTOCOL USING AN EEG-BASED TEACHER.

| Method | Corrected / Total | Correction Rate (%) |
|---|---|---|
| Video Student (No KD) | 19 / 76 | 25.00 |
| KD | 14 / 76 | 18.42 |
| BioKD | **30 / 76** | **39.47** |

TABLE IX
SENSITIVITY TO THE PROGRESSIVE DISTILLATION SCHEDULE ON DEAP (AROUSAL) UNDER THE SUBJECT-WISE PROTOCOL. STAGE ALLOCATION REPORTS THE NUMBERS OF EPOCHS ASSIGNED TO STAGES 0, 1, AND 2, RESPECTIVELY.

| Setting | $E_0$ | $E_1$ | Stage Allocation | Acc. (%) | F1 (%) |
|---|---|---|---|---|---|
| Early | 1 | 4 | 1/3/6 | 64.33 | 75.16 |
| Default | 2 | 5 | 2/3/5 | **65.29** | **76.91** |
| Late | 3 | 6 | 3/3/4 | 64.04 | 76.03 |

*3) Beyond Teacher Confidence:* We first examine how the learned gate relates to teacher prediction uncertainty. This analysis characterizes the behavior of the gate but does not by itself establish supervision reliability, since a confident teacher prediction may still be incorrect. Fig. 3 shows the relationship between teacher prediction entropy and the average gate weight.

As shown in the figure, the gate weight exhibits a decreasing trend as teacher entropy increases, and this relationship is further supported by a negative Spearman correlation of $\rho = -0.78$ with $p < 0.001$. This result indicates that uncertain teacher predictions tend to receive weaker supervision weights. Nevertheless, the correctness-based analyses in Section IV-E1 show that the gate behavior cannot be explained by confidence alone.

To further examine whether the proposed gating mechanism can be reduced to a simple entropy-based weighting strategy, we implement an *entropy-only* baseline, where the distillation weights are computed solely from teacher prediction entropy. Under the subject-wise protocol on DEAP, the entropy-only baseline achieves 55.36% and 59.48% accuracy on the valence and arousal tasks, respectively. While it performs better than uniformly weighted standard KD, it still underperforms BioKD, which reaches 59.13% and 65.29% under the same setting. This indicates that teacher confidence alone is insufficient to fully characterize supervision reliability.

*4) Prediction Calibration:* Beyond classification accuracy, we evaluate whether reliability-aware distillation improves the calibration of the video student. For each sample, prediction

TABLE X
EXPECTED CALIBRATION ERROR ON DEAP AROUSAL UNDER THE SUBJECT-WISE PROTOCOL. LOWER VALUES INDICATE BETTER CALIBRATION.

| Method | ECE ↓ |
|---|---|
| Video Student (No KD) | 0.234 |
| SP [57] | 0.215 |
| FitNets [34] | 0.190 |
| VID [54] | 0.208 |
| NST [58] | 0.242 |
| BioKD | **0.089** |

confidence is defined as the maximum posterior probability, i.e., $\max(p_i^S, 1 - p_i^S)$. Expected calibration error (ECE) is computed using 15 equally spaced confidence bins:

$$\mathrm{ECE} = \sum_{k=1}^{15} \frac{|\mathcal{B}_k|}{N} \left|\mathrm{acc}(\mathcal{B}_k) - \mathrm{conf}(\mathcal{B}_k)\right|, \tag{18}$$

where $\mathcal{B}_k$ denotes the set of predictions assigned to the $k$-th confidence bin, and $N$ is the total number of evaluated samples. Lower ECE indicates better agreement between predictive confidence and empirical correctness. In addition to the main baselines, we include Similarity-Preserving Knowledge Distillation (SP) [57] and Neuron Selectivity Transfer (NST) [58] in this calibration analysis.

As shown in Table X, BioKD achieves the lowest ECE of 0.089, compared with 0.234 for the Video Student (No KD) and 0.190 for the best-performing competing distillation baseline, FitNets. This corresponds to ECE reductions of 62.0% and 53.2%, respectively, indicating that the confidence estimates produced by BioKD are better aligned with empirical correctness. This result suggests that reliability-aware distillation improves not only classification performance but also the calibration of the deployed video student.

*5) Qualitative Gate Analysis:* To complement the aggregate analyses, we examine representative sample-level cases in which teacher supervision is useful, harmful, or overconfidently incorrect. At the group level, useful-teacher-used cases receive higher average gate weights than harmful-teacher-avoided cases (0.4265 vs. 0.3710). Moreover, overconfident wrong-teacher cases exhibit a high average teacher confidence of 0.7759 but receive a substantially lower average gate weight of 0.3950. Fig. 4 presents representative cases from these groups.

As illustrated in Fig. 4, BioKD assigns relatively high gate weights when physiological teacher supervision provides useful corrective information and suppresses knowledge transfer when the teacher conflicts with an already correct video student. The overconfident wrong-teacher examples further demonstrate that the gate can remain conservative even when the teacher produces a strong but incorrect prediction.

*6) Hyperparameter Sensitivity Analysis:* We evaluate the sensitivity of BioKD to the feature-distillation coefficient $\lambda_f$, the logit-distillation coefficient $\lambda_l$, and the progressive schedule on DEAP (Arousal) under the subject-wise protocol. As shown in Fig. 5, each coefficient is varied independently while the other is fixed at its default value of 1.0. Although $\lambda_f$ and $\lambda_l$

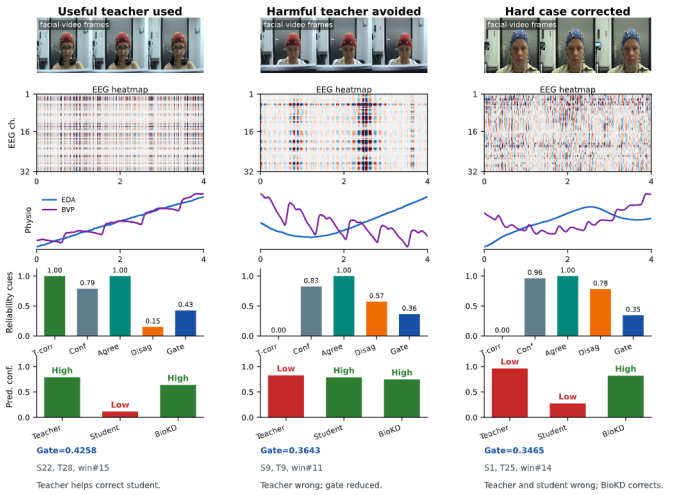


Fig. 4. Representative gate behaviors for helpful, harmful, and overconfident wrong-teacher cases.

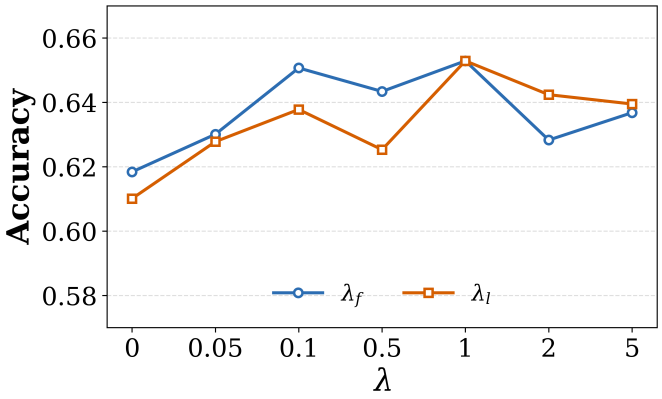


Fig. 5. Sensitivity of BioKD to the feature-level coefficient $\lambda_f$ and logit-level coefficient $\lambda_l$ on DEAP (Arousal) under the subject-wise protocol. When varying one coefficient, the other is fixed at its default value of 1.0.

exhibit different local sensitivity patterns because they regulate knowledge transfer at different representation levels, both achieve the highest accuracy at 1.0. Weakening or removing either objective provides insufficient distillation supervision, whereas an excessively large coefficient may over-constrain the student to match modality-specific features or unreliable teacher predictions. Therefore, the two results obtained at a coefficient value of zero correspond to different training configurations. Table IX further compares different transition timings while keeping the duration of Stage 1 fixed at three epochs. The default schedule $(E_0, E_1) = (2, 5)$ achieves the best overall result, whereas shifting the transitions earlier or later results in only moderate performance changes. These results indicate that BioKD benefits from balanced feature- and logit-level supervision together with a progressive schedule, without relying on a narrowly tuned transition point.

### F. Robustness to Teacher Modality Configurations

The results in Table XI show that BioKD provides stable and effective supervision under different teacher modality configurations. Even when only a uni-modal EEG teacher is used,

TABLE XI
IMPACT OF TEACHER MODALITY CONFIGURATIONS.

| Teacher Configuration | Arousal | Valence |
|---|---|---|
| *DEAP Dataset* | | |
| 1. EEG Only | 65.29 | **59.13** |
| 2. EEG + EMG + EOG | **69.78** | 55.83 |
| 3. EEG + BVP + EDA + TEMP + RESP | 65.23 | 56.36 |
| 4. All Modalities (Full) | 67.34 | 54.48 |
| *AMIGOS Dataset* | | |
| 1. EEG Only | **60.95** | 57.05 |
| 2. EEG + ECG | 58.75 | **61.66** |
| 3. EEG + GSR | 57.79 | 56.39 |
| 4. EEG + GSR + ECG (Full) | 60.27 | 59.45 |

TABLE XII
INFERENCE COST AND DEPLOYMENT CHARACTERISTICS.

| Model | Input | Params (M) | FLOPs (G) | Latency (ms) |
|---|---|---|---|---|
| Physiological Teacher | Physiology | 2.94 | 1.93 | 8.96 |
| Video Student | Video | 2.57 | 3.38 | 4.82 |

the student still achieves competitive performance, indicating that BioKD does not rely on a specific or highly complex teacher design. Meanwhile, the best-performing configuration varies across tasks and datasets, suggesting that additional physiological modalities may offer complementary benefits, while a full-modality teacher is not necessarily the optimal choice.

### G. Inference Cost and Deployment Characteristics

BioKD is designed to reduce the sensing and modality requirements of physiology-assisted emotion recognition rather than to compress the video backbone. The physiological teacher, reliability gate, consistency memory, and shared projection head are used only during training and are discarded afterward. Consequently, the deployed model consists solely of the video student and introduces no additional inference-time modules or computational cost relative to the same student trained without distillation.

As shown in Table XII, the deployed video student contains 2.57M parameters and requires 3.38G FLOPs, with a measured forward-pass latency of 4.82 ms under the reported hardware setting. Although processing video requires more FLOPs than processing one-dimensional physiological signals (3.38G vs. 1.93G), BioKD does not require physiological sensing, teacher execution, or multimodal synchronization during deployment. Therefore, its practical benefit lies in enabling non-intrusive video-only inference without increasing the computational complexity of the underlying video student, rather than in compressing the student backbone.

## V. DISCUSSION

This work revisits a core challenge in affective computing. Video is easy to acquire and deploy, but as an external behavioral signal, it may fail to reflect internal emotional states [1], [9], [10]. On the other hand, physiological signals provide more direct affective information, but they are noisy, highly subject-dependent, and difficult to use in real-world deployment [2], [12], [59]. Therefore, the problem studied here is not simply how to distill knowledge across modalities, but how to exploit useful physiological information during training while avoiding the transfer of unreliable supervision to the student model.

The main finding of this work is that, in physiology-to-video distillation, the overall performance of the teacher does not necessarily reflect the reliability of its supervision at the sample level. Even when the physiological teacher performs well on average, it can still produce high-confidence but incorrect predictions on difficult or noisy samples. This helps explain why uniform distillation may become unstable and even lead to negative transfer. In contrast, BioKD improves distillation stability by explicitly modeling sample-wise supervision reliability and selectively regulating teacher guidance. This suggests that the key issue is not simply whether the teacher is strong on average, but whether its supervision should be trusted for each sample. This finding also clarifies the difference between BioKD and previous methods. Conventional KD and feature-based distillation methods mainly focus on knowledge transfer [26], [30], [34], [35], while recent cross-modal methods often focus on alignment or sample selection [37], [40], [55], [56]. However, our results suggest that reducing the modality gap does not necessarily guarantee reliable supervision when the teacher itself is built on noisy physiological signals. In this setting, the main challenge is not only how to transfer knowledge, but also when teacher supervision should be trusted.

More broadly, these results provide a different perspective on the relationship between external and internal signals in affective computing. They suggest that physiological signals may not be ideal as mandatory inference-time inputs, but they can still serve as training-only privileged information that complements video with internal affective cues. Although BioKD is still limited by teacher quality and proxy-based reliability modeling, the results indicate that explicitly modeling supervision reliability is necessary for stable physiology-to-video knowledge distillation.

## VI. CONCLUSION

This paper presented BioKD, a reliability-aware physiology-to-video knowledge distillation framework for video-based emotion recognition. By treating physiological signals as training-only privileged information, BioKD selectively transfers complementary affective knowledge while retaining facial video as the only inference-time input. Its sample-wise reliability gate and progressive distillation strategy jointly regulate unreliable physiological supervision and reduce negative transfer during cross-modal learning. Experiments on DEAP and AMIGOS demonstrated consistent improvements over representative distillation baselines under both trial-wise and subject-wise protocols. Further analyses confirmed the importance of distinguishing teacher confidence from supervision reliability. Future work will investigate more adaptive reliabil-

ity estimation and broader generalization across datasets and video backbones.

## ACKNOWLEDGMENTS

This work was supported by Guangdong Basic and Applied Basic Research Foundation under Grant No. 2025A1515110098.

## REFERENCES

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information fusion*, vol. 37, pp. 98–125, 2017.

[2] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020.

[3] Z. Ahmad and N. Khan, "A survey on physiological signal-based emotion recognition," *Bioengineering*, vol. 9, no. 11, p. 688, 2022.

[4] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[5] F. Ahmed, A. H. Bari, and M. L. Gavrilova, "Emotion recognition from body movement," *IEEE Access*, vol. 8, pp. 11 761–11 781, 2019.

[6] F. Li and D. Zhang, "Multimodal physiological signals from wearable sensors for affective computing: A systematic review," *Intelligent Sports and Health*, vol. 1, no. 4, pp. 210–222, 2025.

[7] T. Pan, Y. Ye, H. Cai, S. Huang, Y. Yang, and G. Wang, "Multimodal physiological signals fusion for online emotion recognition," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 5879–5888.

[8] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal emotion recognition using visual, vocal and physiological signals: a review," *Applied Sciences*, vol. 14, no. 17, p. 8071, 2024.

[9] J. J. Gross and R. W. Levenson, "Emotional suppression: physiology, self-report, and expressive behavior." *Journal of personality and social psychology*, vol. 64, no. 6, p. 970, 1993.

[10] M. Pantic and L. J. Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[11] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[12] H. D. Critchley, J. Eccles, and S. N. Garfinkel, "Interaction between cognition, emotion, and the autonomic nervous system," in *Handbook of clinical neurology*. Elsevier, 2013, vol. 117, pp. 59–77.

[13] Y. Zhu, L. Han, G. Jiang, P. Zhou, and Y. Wang, "Hierarchical moe: Continuous multimodal emotion recognition with incomplete and asynchronous inputs," *arXiv preprint arXiv:2508.02133*, 2025.

[14] E. Lopez, A. Uncini, and D. Comminiello, "Phemonet: A multimodal network for physiological signals," in *2024 IEEE 8th Forum on Research and Technologies for Society and Industry Innovation (RTSI)*. IEEE, 2024, pp. 260–264.

[15] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger, "A review of classification algorithms for eeg-based brain–computer interfaces: a 10 year update," *Journal of neural engineering*, vol. 15, no. 3, p. 031005, 2018.

[16] Y. Liu, Y. Yao, Z. Wang, J. Plested, and T. Gedeon, "Generalized alignment for multimodal physiological signal learning," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–10.

[17] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[18] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 6, p. e1563, 2024.

[19] V. Strizhkova, H. Kachmar, H. Chaptoukaev, R. Kalandadze, N. Kukhilava, T. Tsmindashvili, N. Abo-Alzahab, M. A. Zuluaga, M. Balazia, A. Dantcheva *et al.*, "Mvp: Multimodal emotion recognition based on video and physiological signals," in *European Conference on Computer Vision*. Springer, 2024, pp. 101–116.

[20] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Computer methods and programs in biomedicine*, vol. 140, pp. 93–110, 2017.

[21] V. Vapnik, R. Izmailov *et al.*, "Learning using privileged information: Similarity control and knowledge transfer." *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 2023–2049, 2015.

[22] Y. Zhu, T. Li, and Y. Wang, "Real-time cross-modal cybersickness prediction in virtual reality," *arXiv preprint arXiv:2501.01212*, 2025.

[23] M. H. Aslam, M. O. Zeeshan, M. Pedersoli, A. L. Koerich, S. Bacon, and E. Granger, "Privileged knowledge distillation for dimensional emotion recognition in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 3338–3347.

[24] J. Hoffman, S. Gupta, and T. Darrell, "Learning with side information through modality hallucination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 826–834.

[25] P. Jemioło, D. Storman, M. Mamica, M. Szymkowski, W. Żabicka, M. Wojtaszek-Główka, and A. Ligęza, "Datasets for automated affect and emotion recognition from cardiovascular signals using artificial intelligence—a systematic review," *Sensors*, vol. 22, no. 7, p. 2538, 2022.

[26] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[27] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 5191–5198.

[28] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" *Advances in neural information processing systems*, vol. 32, 2019.

[29] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, "Your classifier is secretly an energy based model and you should treat it like one," *arXiv preprint arXiv:1912.03263*, 2019.

[30] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International journal of computer vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[31] A. M. Mansourian, R. Ahmadi, M. Ghafouri, A. M. Babaei, E. B. Golezani, Z. Y. Ghamchi, V. Ramezanian, A. Taherian, K. Dinashi, A. Miri *et al.*, "A comprehensive survey on knowledge distillation," *arXiv preprint arXiv:2503.12067*, 2025.

[32] Y. Wu, Q. Mi, and T. Gao, "A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions," *Biomimetics*, vol. 10, no. 7, p. 418, 2025.

[33] B. Pan, K. Hirota, Z. Jia, and Y. Dai, "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods," *Neurocomputing*, vol. 561, p. 126866, 2023.

[34] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets. arxiv 2014," *arXiv preprint arXiv:1412.6550*, 2014.

[35] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[36] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," *arXiv preprint arXiv:1612.03928*, 2016.

[37] S. Gupta, J. Hoffman, and J. Malik, "Cross modal distillation for supervision transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2827–2836.

[38] Y. Chen, T. He, J. Fu, L. Wang, J. Guo, T. Hu, and H. Cheng, "Vision-language meets the skeleton: Progressively distillation with cross-modal knowledge for 3d action representation learning," *IEEE Transactions on Multimedia*, 2024.

[39] Z. Jia, Y. Liu, H. Wang, and T. Jiang, "Cross-modal knowledge distillation for enhanced unimodal emotion recognition," *IEEE Transactions on Affective Computing*, 2025.

[40] Y. Liu, Z. Jia, and H. Wang, "Emotionkd: a cross-modal knowledge distillation framework for emotion recognition based on physiological signals," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 6122–6131.

[41] H. Ma, J. Wang, H. Lin, B. Zhang, Y. Zhang, and B. Xu, "A transformer-based model with self-distillation for multimodal emotion recognition in conversations," *IEEE Transactions on Multimedia*, vol. 26, pp. 776–788, 2023.

[42] M. Hussain, C. Chen, M. Hussain, M. Anwar, M. Abaker, A. Abdelmaboud, and I. Yamin, "Optimised knowledge distillation for efficient social media emotion recognition using distilbert and albert," *Scientific Reports*, vol. 15, no. 1, p. 30104, 2025.

[43] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.
[44] K. Makantasis, K. Pinitas, A. Liapis, and G. N. Yannakakis, "From the lab to the wild: Affect modeling via privileged information," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 380–392, 2023.
[45] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.
[46] P. J. Bota, C. Wang, A. L. Fred, and H. P. Da Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals." *IEEE access*, vol. 7, no. 99, pp. 140 990–141 020, 2019.
[47] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3713–3722.
[48] Z. Lai, C. Qing, J. Tan, W. Luo, and X. Xu, "Online multi-level contrastive representation distillation for cross-subject fnirs emotion recognition," in *Proceedings of the 1st International Workshop on Brain-Computer Interfaces (BCI) for Multimedia Understanding*, 2024, pp. 29–37.
[49] Y. Yang, C. Wang, L. Gong, M. Wu, Z. Chen, Y. Gao, T. Wang, and X. Zhou, "Uncertainty-aware self-knowledge distillation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
[50] S. Saganowski, B. Perz, A. G. Polak, and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876–1897, 2022.
[51] S. Zhang, C. Tang, and C. Guan, "Visual-to-eeg cross-modal knowledge distillation for continuous emotion recognition," *Pattern Recognition*, vol. 130, p. 108833, 2022.
[52] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "Amigos: A dataset for affect, personality and mood research on individuals and groups," *IEEE transactions on affective computing*, vol. 12, no. 2, pp. 479–493, 2018.
[53] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
[54] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9163–9171.
[55] F. Huo, W. Xu, J. Guo, H. Wang, and S. Guo, "C2kd: Bridging the modality gap for cross-modal knowledge distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 006–16 015.
[56] S. Wu, H. Liang, Y. Zhang, Y. Chen, and Z. Jia, "A cross-modal densely guided knowledge distillation based on modality rebalancing strategy for enhanced unimodal emotion recognition," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 2025, pp. 4236–4244.
[57] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1365–1374.
[58] Z. Huang and N. Wang, "Like what you like: Knowledge distill via neuron selectivity transfer," *arXiv preprint arXiv:1707.01219*, 2017.
[59] A. Scarciglia, V. Catrambone, C. Bonanno, and G. Valenza, "Physiological noise: Definition, estimation, and characterization in complex biomedical signals," *IEEE Transactions on Biomedical Engineering*, vol. 71, no. 1, pp. 45–55, 2023.

## VII. Biography Section

**Bojing Hou** is currently a Ph.D. student in the Computational Media and Arts Thrust at The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. Her research interests lie in multimodal affective computing, with a particular focus on reliable cross-modal knowledge transfer, uncertainty-aware multimodal learning, and weakly supervised emotion recognition.

**Ruohao LI** is currently a Ph.D. Candidate in the Thrust of Computational Media and Arts at the Hong Kong University of Science and Technology (Guangzhou). His research interests include Human-Robot Interaction, Context-aware Interactive System, and Immersive Technologies. He serves as a reviewer for CHI, IMWUT, ACM MM, etc.

**Yitong Zhu** is currently a Ph.D. student in the Thrust of Computational Media and Arts at the Hong Kong University of Science and Technology (Guangzhou). Her recent research interests include Sentiment Analysis, Human Factors Engineering and Cognitive Neuroscience. She has served as a reviewer of ACL, CVPR and ACM MM etc.

**Hongjun Liu** is currently a Ph.D. candidate in the Center of Data Science at New York University, advancing agentic AI and reasoning-intensive NLP through executable skill interventions, process-level rewards, and multimodal systems. She serves as a reviewer for ACL, EMNLP, ICLR, Neurips, etc.

**Dr. Luwen Yu** is currently an Assistant Professor in Computational Media and Arts at The Hong Kong University of Science and Technology (Guangzhou). She received her bachelor's and master's degrees in Design from Huazhong University of Science and Technology, and her PhD from the University of Leeds. Her research interests include colour science and design, colour perception and cognition, colour and emotion, visual attention, and aesthetic neuroscience, with a particular interest in multimodal psychophysiological methods and their applications in design and digital media.

**Dr. Yuyang Wang** is an Assistant Professor in the Computational Media and Arts (CMA) Thrust at the Hong Kong University of Science and Technology (Guangzhou) Information Hub. He was with the CMA thrust from 2022 to 2024, where he held the position of Postdoctoral Research Fellow. He received a PhD in computer science from the Arts et Métiers Institute of Technology, ParisTech in France. Under the framework of the French-German Institute for Industry of the Future, he was a visiting researcher at the Karlsruhe Institute of Technology in Germany. His research interests include affective computing,multimodal BCI and immersive human-machine interaction.