PULSE: Privileged Knowledge Transfer from Rich to Deployable Sensors for Embodied Multi-Sensory Learning

Anonymous CVPR submission

Paper ID *****

Abstract

Multi-sensory systems for embodied intelligence, from wearable body-sensor networks to instrumented robotic platforms, routinely face a sensor-asymmetry *problem: the richest modality available during laboratory data collection is absent or impractical at deployment time due to cost, fragility, or interference with physical interaction. We introduce* ***PULSE****, a general framework for* privileged knowledge transfer *from an information-rich teacher sensor to a set of cheaper, deployment-ready student sensors. Each student encoder produces* shared *(modality-invariant) and* private *(modality-specific) embeddings; the shared subspace is aligned across modalities and then matched to representations of a frozen teacher via multi-layer hidden-state and pooled-embedding distillation. Private embeddings preserve modality-specific structure needed for self-supervised reconstruction, which we show is critical to prevent representational collapse. We instantiate PULSE on the wearable stress-monitoring task, using electrodermal activity (EDA) as the privileged teacher and ECG, BVP, accelerometry, and temperature as students. On the WESAD benchmark under leave-one-subject-out evaluation, PULSE achieves 0.994 AUROC and 0.988 AUPRC (0.965/0.955 on STRESS)* without *EDA at inference, exceeding all no-eda baselines and matches the performance of a full-sensor model that retains EDA at test time. We further demonstrate modality-agnostic transfer with ECG as teacher, extensive ablations on backbone design, pretraining strategies, and shared-private capacity, and discuss how the framework generalizes to broader embodied sensing scenarios involving tactile, inertial, and bioelectrical modalities.*

1. Introduction

Multi-sensory systems for embodied intelligence routinely face a sensor-asymmetry problem: the richest modality available during data collection is absent or impractical at deployment. This challenge is especially acute in wearable health monitoring, where *electrodermal activity* (EDA), the gold-standard physiological index of sympathetic arousal and acute stress [6, 7, 28–30], requires Ag/AgCl electrodes with a constant-current source and is highly vulnerable to motion artifacts [12]. Consequently, most commercial wearables and longitudinal studies provide only ECG/PPG (BVP), inertial signals, and skin temperature; models trained with EDA simply cannot rely on it at test time.

This creates a dilemma: EDA carries privileged information about stress-related autonomic activation that cheaper modalities lack, yet discarding EDA entirely during training wastes a valuable supervisory signal. The natural question is whether we can exploit EDA during training to improve the representations of deployment-grade sensors, without ever requiring EDA at inference.

Privileged knowledge transfer as a remedy. We address this through the Learning Using Privileged Information (LUPI) paradigm [20, 33]: a teacher model trained on EDA transfers its learned representations into student encoders that consume only cheap, deployable modalities (ECG, BVP, ACC, TEMP). Crucially, we *freeze* the pretrained EDA teacher, providing a stable optimization target, rather than updating all encoders symmetrically as in missing-modality robustness methods [13, 14, 24] that do not exploit the informational asymmetry between a privileged teacher and lighter-weight students.

The challenge: what to align, what to preserve. Naïvely aligning all student representations to the teacher risks over-constraining the students: each modality carries unique, modality-specific structure (e.g., QRS morphology in ECG, pulse wave shape in BVP) that does not and should not match EDA. Forcing full alignment destroys this structure and, as we show empirically, leads to representational collapse. The key design insight is to separate each student's representation into a *shared* subspace (modality-invariant content aligned to the teacher) and a *private* sub-

space (modality-specific features preserved for reconstruction), and to transfer knowledge only through the shared channel.

Identified gap. LUPI and knowledge distillation have shown benefits in vision [20], speech [23], and affect modeling from video [2–4, 22], while recent work distills PPG representations into accelerometer encoders at population scale [1]. EmotionKD [19] transfers across EEG and EDA for emotion recognition. However, to our knowledge, no prior study combines (i) a frozen privileged-sensor teacher with (ii) an explicit shared–private embedding decomposition and (iii) multi-depth hidden-state alignment for wearable biosignal transfer, a design that disentangles transferable content from modality-specific structure at every layer of the network.

Contributions. We propose **PULSE**: **P**rivileged knowledge transfer **U**sing **L**ow-cost **SE**nsors for wearable stress monitoring:

1. **EDA as a frozen teacher.** We pretrain an EDA encoder with self-supervised masked reconstruction, freeze it, and use its hidden-state embeddings at every layer plus a final pooled representation as privileged supervision. EDA is never required at inference.
2. **Shared–private embedding decomposition.** Each student encoder splits its output into a shared subspace (aligned across modalities and matched to the teacher) and a private subspace (preserving modality-specific information for reconstruction). We show both components are essential: removing private capacity collapses representations, while removing shared capacity eliminates the transfer pathway.
3. **Reconstruction as a collapse safeguard.** We demonstrate that knowledge distillation without an auxiliary reconstruction objective causes shared embeddings to collapse to a constant vector (cosine similarity ≈ 1.0). The reconstruction branch acts as an information-preserving regularizer, a finding with implications for any multi-sensory distillation setup.
4. **Comprehensive evaluation.** On WESAD [30] (15 subjects) and PhysioNet STRESS [11] (36 subjects), PULSE achieves 0.994 AUROC and 0.988 AUPRC(0.965/0.955 on STRESS) for binary stress detection *without* EDA at inference, surpassing both a no-teacher baseline and a full-sensor model that retains EDA. We provide ablations to illustrate the effectiveness of each of our design choices.

Broader relevance to multi-sensory embodied intelligence. While we instantiate PULSE on wearable physiological signals, the framework addresses a fundamental challenge shared across embodied sensing: how to compress information from a rich but impractical sensor into representations learned by cheaper, deployable modalities. The shared–private decomposition, multi-depth distillation, and reconstruction-based collapse prevention are architecture-agnostic design principles applicable wherever sensor asymmetry arises, whether in tactile-to-IMU transfer for robotic manipulation, high-fidelity-to-consumer-grade transfer in XR sensing, or bioelectrical-to-optical transfer in wearable health monitoring.

2. Related Work

Learning using privileged information (LUPI). Vapnik and Izmailov [33] introduced LUPI, where a teacher has access to additional features unavailable to the student. Lopez-Paz *et al*. [20] unified distillation and privileged information under a common framework, and Markov and Matsui [23] demonstrated gains in robust speech recognition. In affect modeling, Makantasis *et al*. [21, 22] transfer from laboratory physiological signals to video-only models for deployment in the wild.

Knowledge distillation in multi-sensory systems. EmotionKD [19] distills across heterogeneous biosignals (EEG and EDA) for single-modality emotion recognition at test time. Aslam *et al*. [2–4] develop optimal-transport and multi-teacher distillation for expression recognition. Abbaspourazad *et al*. [1] distill PPG representations into accelerometer encoders at population scale, producing generalist wearable foundation models.

Missing-modality robustness. PhysioOmni [14] and ADAPT [24] align modalities symmetrically to handle missing channels, while Ibtehaz and Mortazavi [13] pursue simultaneous alignment and reconstruction for multi-lead ECG. These approaches treat all modalities as peers rather than exploiting asymmetry between a privileged teacher and deployment students.

Self-supervised pretraining for time series. Several self-supervised objectives have been proposed for time-series representation learning. Contrastive Multi-Segment Coding (CMSC) [16] treats adjacent non-overlapping segments of the same recording as positive pairs and applies contrastive learning to capture temporal invariance. CLIP-style approaches [27] align representations across paired views (e.g., time and frequency domains, or different sensor channels) with an InfoNCE objective and have been adapted to biosignals [35]. TS-TCC [8] combines temporal and contextual contrasting with strong/weak augmentations. Masked autoencoders (MAE) [10], originally developed for vision, reconstruct randomly masked patches and have been adapted to physiological signals [18].

Positioning. Unlike symmetric alignment methods, PULSE uses a *frozen* privileged-modality teacher that provides a stable optimization target at every encoder layer. Unlike prior PKD for affect [2, 19], we introduce an explicit shared–private decomposition that prevents over-constraining modality-specific features, and we demonstrate that reconstruction loss is essential for preventing collapse during distillation, a finding with implications for any multi-sensory transfer setup.

3. Datasets and Preprocessing

Source. We use the publicly available **WESAD** benchmark [30], which provides synchronized physiological signals from 15 subjects wearing a RespiBAN chest band and an Empatica E4 wristband during a 35-min protocol (baseline → social-stress → comedy). Signals recorded on the wrist include *electrodermal activity* (EDA, 4 Hz), *photoplethysmography* (BVP, 64 Hz), skin *temperature* (TEMP, 4 Hz), and tri-axial *acceleration* (ACC, 32 Hz); the chest strap provides *ECG*, respiration, EMG, and ACC at 700 Hz. This multi-sensor setup exemplifies the sensor-asymmetry problem: EDA is the gold-standard stress indicator but requires specialized electrodes, while the remaining modalities are available on commodity wearables.

Resampling. All channels are resampled to a uniform 64 Hz target rate: first-order polyphase filtering for high-frequency signals (ECG, BVP, ACC), linear interpolation for low-frequency channels (EDA, TEMP).

Signal cleaning.

- **EDA**: tonic drift removed with a first-order 0.05 Hz high-pass; phasic band retained with a 0.05–1 Hz Butterworth filter.
- **BVP**: band-pass 0.5–2 Hz to isolate pulse wave.
- **ECG**: band-pass 0.5–40 Hz.
- **Motion**: net acceleration computed as $\|\mathrm{ACC}_{x,y,z}\|$ from the *wrist* accelerometer only; chest-band ACC is excluded to match wrist-only deployment hardware.

Segmentation. Cleaned, synchronized streams are cut into 60-second windows with a 0.25-second stride (96% overlap), following [30, 32]. Label 0 (*transient*) samples are discarded; only windows whose entire 60 s lie in a single class $\{1{=}\text{baseline}, 2{=}\text{stress}, 3{=}\text{amusement}\}$ are kept. Windows whose ECG or BVP standard deviation is lower than 0.02 (after z-scoring) are rejected to remove sensor drop-outs.

Normalization. Per-subject z-score parameters are estimated from all baseline (label=1) windows and applied to every channel, preserving inter-subject differences while removing long-term drift.

Leave-one-subject-out (LOSO) folds. The pipeline yields $\sim 8 \times 10^3$ valid windows per subject (3,840 samples per window). We build fifteen LOSO folds: in fold k, subject S_k is the test set and the remaining 14 subjects form the training set.

3.1. Second Dataset: PhysioNet STRESS

To evaluate cross-dataset generalization, we additionally use the **Wearable Device Dataset from Induced Stress and Structured Exercise Sessions** [11], a publicly available dataset on PhysioNet containing physiological recordings from 36 healthy volunteers. Data was collected using the Empatica E4 wristband during a structured acute stress induction protocol involving mathematical and emotional tasks. The device records *blood volume pulse* (BVP, 64 Hz), *electrodermal activity* (EDA, 4 Hz), *skin temperature* (TEMP, 4 Hz), and tri-axial *acceleration* (ACC, 32 Hz). Unlike WESAD, no chest-band ECG is available; this makes the dataset a natural testbed for the wrist-only deployment scenario.

We apply the same preprocessing pipeline as WESAD (resampling to 64 Hz, signal cleaning, 60 s windowing, per-subject baseline normalization) to the PhysioNet STRESS data, using the rest/baseline segments for normalization.

4. Methods

Our training pipeline, summarized in Fig. 1, proceeds in three stages: (1) self-supervised pretraining of per-modality masked autoencoders with cross-modal alignment, (2) privileged knowledge transfer from a frozen teacher, and (3) supervised finetuning with only deployment sensors. No privileged sensor data is required at inference.

4.1. Pretraining

Before knowledge transfer, the student `PhysioMAE` models (ECG, BVP, ACC, TEMP) are pretrained jointly in a self-supervised manner, while the teacher `PhysioMAE` (EDA) is pretrained separately. Inputs of length 3,840 samples at 64 Hz are split into non-overlapping patches (size 96), embedded, and processed by a transformer encoder (embedding dimension 1024, depth 8, heads 8). A lightweight decoder (dimension 512, depth 4, heads 8) reconstructs masked signal patches.

Shared–private decomposition. Each student encoder outputs two types of embeddings: *shared* embeddings (modality-invariant) and *private* embeddings (modality-specific), separated via a random binary mask. The teacher encoder outputs only private embeddings, as it does not undergo cross-modal alignment.

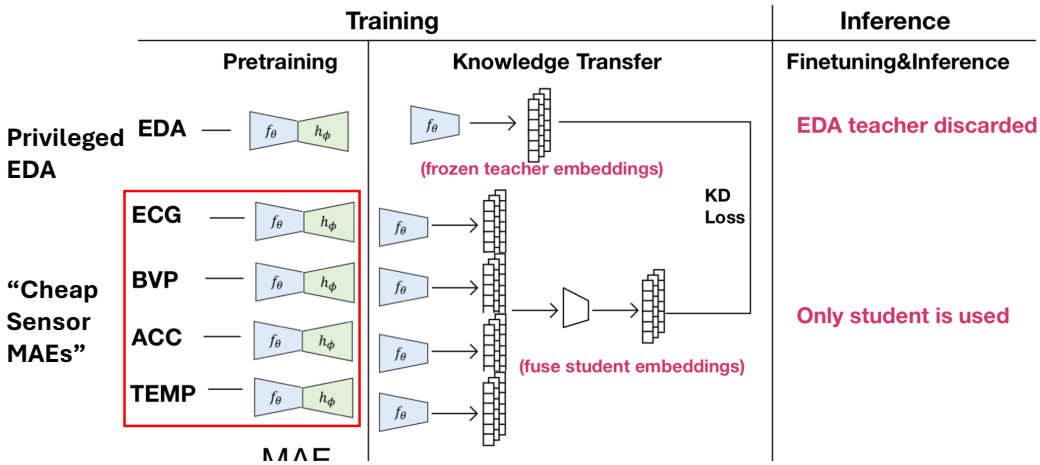

Figure 1. **The PULSE Framework.** Our framework uses privileged knowledge transfer from a frozen teacher encoder (here: EDA) to students built on low-cost, deployment-ready sensors. In the pretraining stage, student encoders learn modality-invariant *shared* embeddings alongside modality-specific *private* embeddings. Knowledge transfer aligns the students' shared embeddings with the privileged teacher. During finetuning, only the learned student embeddings are used for supervised classification; no privileged sensor is required at inference.

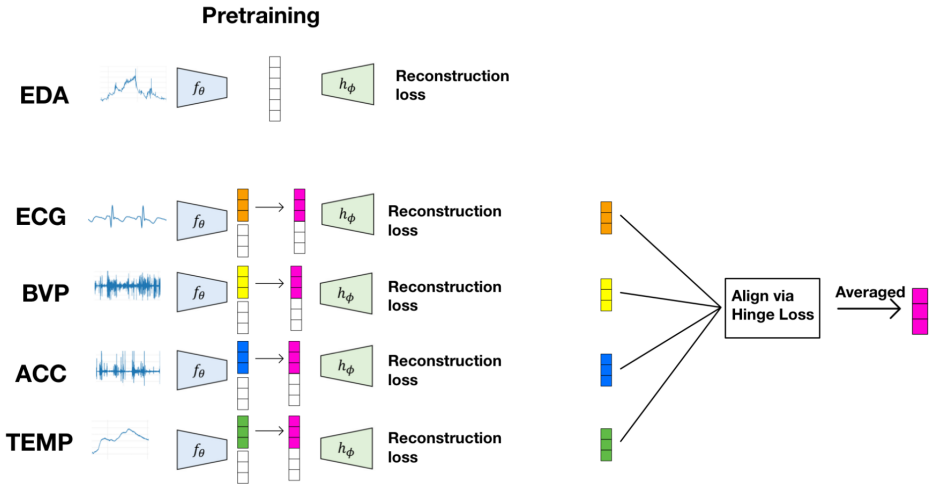

Figure 2. **Pretraining Setup.** Each student encoder outputs shared embeddings (colored) and private embeddings (white). Shared embeddings are aligned across modalities via a hinge loss, then averaged into a single modality-invariant embedding (magenta). This averaged embedding, together with private embeddings, drives reconstruction. The teacher MAE is pretrained separately via reconstruction loss only.

Alignment loss. Shared embeddings are aligned using a hinge loss with in-batch negatives. Let $\mathcal{P}$ denote positive (matched) cross-modal pairs. For a positive pair $(i, j) \in \mathcal{P}$ with shared embeddings s_i, s_j and in-batch negatives $\mathcal{N}(j), \mathcal{N}(i)$:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \Bigg[\frac{1}{|\mathcal{N}(j)|} \sum_{j'\in\mathcal{N}(j)} \max\big(0, \cos\langle s_i, s_{j'}\rangle - \cos\langle s_i, s_j\rangle + \alpha\big) + \frac{1}{|\mathcal{N}(i)|} \sum_{i'\in\mathcal{N}(i)} \max\big(0, \cos\langle s_{i'}, s_j\rangle - \cos\langle s_i, s_j\rangle + \alpha\big) \Bigg], \tag{1}$$

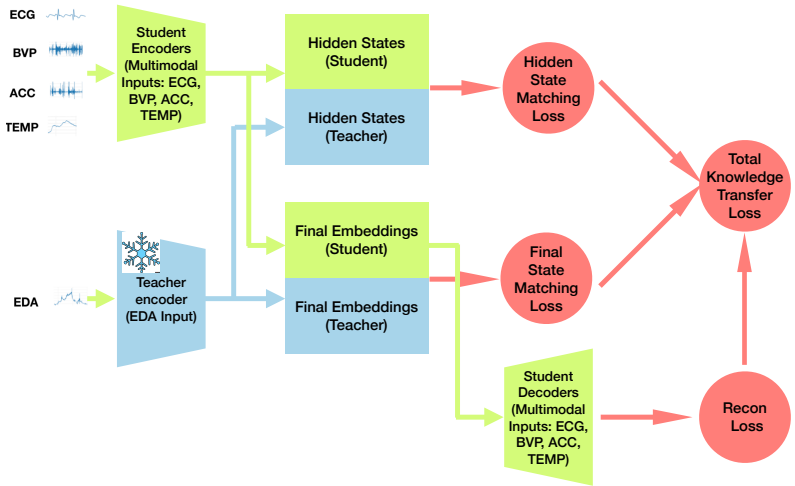

Figure 3. **Knowledge Transfer.** The frozen EDA teacher encoder and the student encoders (ECG, BVP, ACC, TEMP) process their respective inputs in parallel. Knowledge is transferred through two complementary objectives: (1) a hidden-state matching loss that aligns fused student hidden representations with the teacher's intermediate representations at every encoder layer, and (2) a final-state matching loss that matches the pooled last-layer embeddings. An auxiliary reconstruction loss from the student decoders preserves modality-specific information and prevents representational collapse. The three terms are combined into the total knowledge transfer loss. The teacher encoder remains frozen throughout; only student parameters are updated.

where α=0.2 is the margin.

Reconstruction loss. The averaged shared embedding and private embeddings are decoded to reconstruct masked patches Ω:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \|x_\omega - \hat{x}_\omega\|^2. \tag{2}$$

Total pretraining loss. $\mathcal{L}_{\text{pre}} = \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}$, with λ_{align}=λ_{rec}=1. We pretrain for 300 epochs with Adam (lr=10^{-4}, batch size 128).

4.2. Knowledge Transfer

We freeze the teacher encoder and distill it into the four student encoders. All models share the same ViT-style configuration.

Transfer heads and fusion. A lightweight transfer head (i) layer-normalizes student embeddings, (ii) projects into shared/private subspaces via per-modality linear layers, and (iii) fuses shared embeddings across modalities with a linear fusion layer initialized to exact averaging.

What is transferred. We align (a) *hidden tokens* at all student layers against corresponding teacher layers and (b) a *final pooled embedding* from the last layer. No logits or labels are used.

For each matched layer ℓ:

$$\mathcal{L}_{\text{hid}} = \frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \Big(1 - \cos\big\langle \texttt{Fuse}(\{S_m^\ell\}), T^\ell \big\rangle\Big), \tag{3}$$

and for the final embedding:

$$\mathcal{L}_{\text{emb}} = 1 - \cos\big\langle \text{mean}_t\, \texttt{Fuse}(\{S_m^{\text{final}}\}), \text{mean}_t\, T^{\text{final}} \big\rangle. \tag{4}$$

Optional regularizers. Two optional terms can be toggled: (i) a decorrelation penalty $\mathcal{L}_{\text{perp}}$ between shared and private embeddings, and (ii) a reconstruction loss from each student's MAE decoder (masking ratio 0.5). We find reconstruction essential for avoiding collapse (**??**).

Total loss. $\mathcal{L} = \lambda_{\text{hid}}\mathcal{L}_{\text{hid}} + \lambda_{\text{emb}}\mathcal{L}_{\text{emb}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{perp}}\mathcal{L}_{\text{perp}}$, with defaults λ_{hid}=λ_{emb}=1, λ_{rec}=0.1, λ_{perp}=0. We train for 100 epochs with Adam (lr=10^{-4}, batch size 128).

4.3. Finetuning and Inference

At test time, the privileged modality is absent. Student encoders consume only deployment-grade sensors, form a fused shared representation, and classify via a lightweight 2-layer MLP (hidden dimension 4). Encoders remain frozen; only the MLP head is trained.

Overfitting control. We uniformly subsample training windows by $1/40$ to reduce correlation between overlapping windows and use a deliberately small classification head to limit capacity, improving cross-subject generalization.

Optimization. We train each fold for 300 epochs with Adam (lr=10^{-3}), batch size 128, cosine LR schedule. Model selection uses highest validation AUPRC.

5. Experiments and Results

We evaluate five training configurations summarized in Tab. 1 and report aggregate metrics in Tab. 2.

Metric choice. We report threshold-independent AUROC and AUPRC as primary metrics, avoiding the instability of threshold-dependent accuracy under cross-subject evaluation. Accuracy at a fixed threshold is included as a secondary summary.

Main observations. Symmetric alignment (B) yields only a minor improvement over the no-teacher baseline (A) when the privileged modality is removed at test time: enforcing cross-modal consistency alone does not transfer enough privileged information. PULSE (C) delivers consistent gains across all metrics. The frozen teacher provides a stable optimization target that shapes the students' shared representations toward the teacher's signal structure, and this benefit persists when only cheap sensors are used at inference.

Notably, PULSE (C) *matches* the full-sensor model (D) on AUROC/AUPRC, despite not using EDA at test time. We attribute this to: (i) the frozen teacher acting as a data-dependent regularizer; (ii) joint optimization in D overfitting to idiosyncratic sensor artifacts; and (iii) knowledge transfer effectively denoising and compressing the privileged signal into the student representation.

5.1. Three-Class Classification

The gap between PULSE and the no-teacher baseline *widens* in the three-class setting (Tab. 3), suggesting that fine-grained distinctions between affective states benefit disproportionately from privileged supervision.

5.2. Results on PhysioNet STRESS

We evaluate PULSE on the PhysioNet STRESS dataset [11] to test generalization beyond WESAD. Since this dataset provides only wrist-worn Empatica E4 signals (BVP, EDA, ACC, TEMP) without chest-band ECG, the student set consists of BVP, ACC, and TEMP, with EDA as the privileged teacher. Results under LOSO across 36 subjects are in Tab. 4.

6. Ablation Studies

We present three ablations that validate the core design principles of PULSE. Additional hyperparameter sweeps and secondary ablations are in the supplementary material (Sec. B).

6.1. Hidden-State Matching Depth

Supervising all hidden layers yields the best performance (Tab. 5), and removing hidden-state matching entirely degrades metrics *below* the no-KD baseline, indicating multi-depth guidance is essential for stable cross-modal transfer. This suggests intermediate representations carry complementary structural information about the privileged modality that final-layer-only transfer cannot capture.

6.2. Shared–Private Capacity

Both extremes hurt (Tab. 6): no private capacity (ratio=0) eliminates modality-specific features needed for reconstruction; no shared capacity (ratio=1) removes the pathway for privileged transfer. A balanced split (0.5) maximizes ranking metrics, validating the shared–private decomposition as essential rather than incidental.

7. Discussion: Generalization to Broader Embodied Sensing

While our experiments use wearable physiological signals, the core insights of PULSE address challenges shared across embodied multi-sensory systems:

Sensor asymmetry is ubiquitous. Dense tactile arrays on robotic grippers provide rich contact information during lab data collection but are too fragile or expensive for field deployment, where only sparse force sensors and IMUs are available. High-fidelity bioelectrical sensors (EDA, EMG, EEG) capture neural and autonomic processes during controlled human studies but are impractical during dexterous manipulation, where electrodes interfere with grasping. In XR, research headsets carry a full complement of physiological sensors, but consumer hardware retains only optical heart-rate and inertial measurement. In all cases, PULSE's formulation applies: freeze a teacher trained on the rich modality and distill into deployment-grade students.

Shared–private decomposition as a general principle. The decomposition into modality-invariant and modality-specific subspaces is motivated by a general observation: not all information in one sensor should be forced to match another. In tactile-to-IMU transfer, the spatially-resolved pressure distribution (private to tactile) differs fundamentally from the whole-body dynamics captured by the IMU; only the shared contact-event structure should be aligned. Our ablations confirm that collapsing private capacity (ratio=0) or shared capacity (ratio=1) degrades performance.

Reconstruction as a collapse safeguard. We demonstrate that knowledge distillation without an auxiliary reconstruction objective leads to representational collapse.

Table 1. Training and evaluation configurations. "Cheap sensors" refers to deployment-grade modalities (ECG, BVP, ACC, TEMP). The privileged sensor (EDA) is used only during training.

ID	Name	Train-time inputs	Teacher?	Test-time inputs	Purpose
A	No-teacher baseline	Cheap sensors only	–	Cheap sensors only	Achievable without privileged data
B	Symmetric alignment	Cheap sensors + EDA (all updated jointly)	–	Cheap sensors only	Standard cross-modal alignment
C	**PULSE**	Cheap sensors + EDA (EDA frozen → students match)	Yes	Cheap sensors only	Privileged knowledge transfer
D	Full-sensor baseline	Same as B	–	Cheap sensors + EDA	Upper bound if privileged sensor is kept
E	Teacher only (EDA)	Pretrain EDA MAE; finetune classifier	–	EDA only	Teacher quality verification

Table 2. Binary stress classification (baseline vs. stress) under LOSO. Mean ± std across 15 subjects. PULSE uses no privileged sensor at inference.

ID / Name	Test-time inputs	AUROC	AUPRC	Accuracy (%)
A / No-teacher baseline	Cheap sensors	0.963 ± 0.050	0.937 ± 0.101	91.64 ± 6.61
B / Symmetric alignment	Cheap sensors	0.972 ± 0.031	0.944 ± 0.061	88.83 ± 6.24
C / **PULSE**	Cheap sensors	$\mathbf{0.994 \pm 0.011}$	$\mathbf{0.988 \pm 0.022}$	$\mathbf{96.08 \pm 4.52}$
D / Full-sensor baseline	Cheap sensors + EDA	0.983 ± 0.028	0.963 ± 0.048	90.74 ± 5.58
E / Teacher only (EDA)	EDA	0.962 ± 0.067	0.924 ± 0.122	87.20 ± 17.69

Table 3. Three-class classification (baseline / stress / amusement) under LOSO. PULSE uses no EDA at inference.

Model	AUROC	AUPRC	Acc. (%)
no-EDA baseline	0.891 ± 0.090	0.810 ± 0.130	71.62 ± 10.27
symmetric alignment	0.811 ± 0.103	0.710 ± 0.126	69.12 ± 8.82
PULSE	$\mathbf{0.956 \pm 0.058}$	$\mathbf{0.894 \pm 0.115}$	$\mathbf{85.40 \pm 7.18}$
full-sensor baseline	0.812 ± 0.128	0.703 ± 0.124	73.38 ± 7.89

Table 4. Binary stress classification on PhysioNet STRESS [11] under LOSO (36 subjects). Student modalities: BVP, ACC, TEMP. EDA is privileged (training only).

Model	AUROC	AUPRC	Acc. (%)
No-teacher baseline	0.827 ± 0.117	0.716 ± 0.168	82.61 ± 11.2
Symmetric alignment	0.734 ± 0.196	0.678 ± 0.191	78.57 ± 13.74
PULSE	$\mathbf{0.965 \pm 0.045}$	$\mathbf{0.955 \pm 0.045}$	$\mathbf{88.88 \pm 6.93}$
Full-sensor baseline	0.835 ± 0.135	0.722 ± 0.141	76.98 ± 11.44

When we train without reconstruction (KD-only), the shared embeddings collapse toward a constant vector across modalities: the mean pairwise cosine $\approx$ 1.0 for ECG/BVP/ACC (TEMP 0.998), and the feature variance is near-zero, indicating almost no dispersion. After adding the reconstruction objective, feature variance returns by several orders of magnitude, and mean pairwise cosine drops to 0.027–0.137, showing that samples are no longer trivially aligned (Fig. 4). This pattern holds consistently across modalities, supporting our claim that reconstruction acts as an information-preserving regularizer: it counteracts the KD objective's tendency to minimize alignment by shrinking representation variance, thereby preventing representational collapse and yielding meaningful geometry for downstream LOSO generalization.

Multi-depth distillation. Our ablations in Table 5 show all-layer hidden-state matching substantially outperforms

Table 5. Effect of hidden-state matching depth during knowledge transfer.

Configuration	AUROC	AUPRC	Acc. (%)
Layers 3,5,7	.989	.977	93.97
$\mathcal{L}_{\text{hid}}$=0 (final only)	.953	.922	90.95
All layers	**.994**	**.988**	**96.08**
No KD (baseline)	.963	.937	91.64

Table 6. Private-mask ratio controlling shared vs. private capacity.

Ratio	AUROC	AUPRC	Acc. (%)
0 (all shared)	.959	.927	70.41
0.2	.980	.966	**94.46**
0.5 (default)	**.989**	**.977**	93.97
0.8	.975	.958	93.76
1 (no transfer)	.945	.906	70.41

Modality	mean pairwise cosine	mean feature variance
ECG	1.000	2.57e–05
BVP	1.000	6.26e–05
ACC	1.000	7.53e–05
TEMP	0.998	7.96e–04

Without reconstruction

Modality	mean pairwise cosine	mean feature variance
ECG	0.070	2.52e–02
BVP	0.0645	1.51e–02
ACC	0.137	9.20e–02
TEMP	0.0273	2.06e–03

With reconstruction

Figure 4. Effect of adding reconstruction during KD. Without reconstruction, shared embeddings collapse (cosine $\approx$ 1.0, near-zero variance). Adding reconstruction restores variance by several orders of magnitude, preventing collapse and enabling meaningful shared geometry.

final-only transfer. This suggests that intermediate representations carry complementary structural information about the privileged modality, a finding likely transferable to other encoder architectures processing heterogeneous sensor streams.

Modality-Agnostic Transfer: ECG as Teacher To evaluate whether privileged knowledge transfer in PULSE depends specifically on EDA or can generalize to other physiological modalities, we conducted an ablation using **ECG as the privileged teacher**. In this setup, the pretrained ECG encoder is frozen and used to distill knowledge into the remaining deployable modalities (BVP, ACC, and TEMP) through the same shared/private alignment and hidden-state matching objectives. Tab. 11 reports the results under the LOSO protocol.

Using ECG as the teacher yields moderate improvements in AUROC and AUPRC compared to the no-KD counterpart, while accuracy remains comparable within variance. This demonstrates that the PULSE framework is *modality-agnostic*: privileged distillation benefits are not limited to EDA, though the magnitude of gain depends on the informativeness of the teacher modality. EDA remains the strongest privileged signal for stress-related supervision due to its direct coupling with sympathetic arousal, yet ECG-based distillation still provides measurable transfer, confirming that PULSE can flexibly leverage alternative privileged channels when EDA is unavailable.

8. Conclusion

We present PULSE, a privileged knowledge transfer framework that addresses the sensor-asymmetry problem in multi-sensory embodied systems. By decomposing student representations into shared and private embeddings, aligning only the shared subspace with a frozen privileged-sensor teacher, and using reconstruction to prevent collapse, PULSE transfers information from rich but impractical sensors into deployment-ready modalities. On WESAD, PULSE achieves state-of-the-art stress detection without EDA at inference, matching if not surpassing the full-sensor model. We further validate the framework on the PhysioNet STRESS dataset, demonstrating generalization across different populations and stressor protocols.

The framework's design principles, namely shared–private decomposition, multi-depth distillation, hinge-based alignment, and reconstruction-based regularization, are sensor- and task-agnostic. We envision applications in tactile-to-inertial transfer for robotic manipulation, bioelectrical-to-optical transfer for consumer wearables, and multi-sensory XR systems where privileged data is available only during development. Future work will validate PULSE on additional datasets and sensing domains, explore multi-teacher ensembles, and study cross-dataset transfer to quantify domain generalization.

References

[1] Salar Abbaspourazad, Anshuman Mishra, Joseph Futoma, Andrew C. Miller, and Ian Shapiro. Wearable accelerometer foundation models for health via knowledge distillation. *arXiv preprint arXiv:2412.11276*, 2025. 2

[2] Muhammad Haseeb Aslam, Soufiane Belharbi, Marco Pedersoli, Alessandro L. Koerich, Simon Bacon, and Eric Granger. Distilling privileged multimodal information for expression recognition using optimal transport. *arXiv preprint arXiv:2401.15489*, 2024. 2, 3

[3] Muhammad Haseeb Aslam, Marco Pedersoli, Alessandro Lameiras Koerich, and Eric Granger. Multi teacher privileged knowledge distillation for multimodal expression recognition. *arXiv preprint arXiv:2408.09035*, 2024.

[4] Muhammad Haseeb Aslam, Marco Pedersoli, Alessandro L. Koerich, and Eric Granger. Leveraging diversity for privileged multi-teacher knowledge distillation for facial expression recognition. *SSRN preprint 5270694*, 2025. 2

[5] Behnam Behinaein, Anubhav Bhatti, Dirk Rodenburg, Paul Hungler, and Ali Etemad. A transformer architecture for stress detection from ECG. In *UbiComp*, pages 132–134, 2021. 10, 11

[6] Wolfram Boucsein. *Electrodermal Activity*. Springer, 2nd edition, 2012. 1

[7] Hugo D. Critchley. Electrodermal responses: What happens in the brain. *The Neuroscientist*, 8(2):132–142, 2002. 1

[8] Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. In *IJCAI*, pages 2352–2359, 2021. 2

[9] Sayandeep Ghosh, SeongKi Kim, Muhammad Fazal Ijaz, Pawan Kumar Singh, and Mufti Mahmud. Classification of mental stress from wearable physiological sensors using image-encoding-based deep neural network. *Biosensors*, 12 (12), 2022. 11

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2

[11] Alejandro Hongn, Federico Bosch, Luis E. Prado, and María Paula Bonomini. Wearable device dataset from induced stress and structured exercise sessions. *PhysioNet*, 2025. Version 1.0.1. 2, 3, 6, 7

[12] Md-Billal Hossain, Hugo F. Posada-Quintero, Youngsun Kong, Riley McNaboe, and Ki H. Chon. Automatic motion artifact detection in electrodermal activity data using machine learning. *Biomed. Signal Process. Control*, 71:103176, 2022. 1

[13] Nabil Ibtehaz and Masood S. Mortazavi. Modally reduced representation learning of multi-lead ECG signals through simultaneous alignment and reconstruction. In *ICLR 2024 Workshop on Learning from Time Series For Health*, 2024. 1, 2

[14] Wei-Bang Jiang, Xi Fu, Yi Ding, and Cuntai Guan. PhysioOmni: Towards robust multimodal physiological foundation models with arbitrary missing modalities. *arXiv preprint arXiv:2504.19596*, 2025. 1, 2

[15] Alperen Karan and Atabey Kaygun. Time series classification via topological data analysis. *Expert Systems with Applications*, 183:115326, 2021. 11

[16] Dani Kiyasseh, Tingting Zhu, and David A. Clifton. CLOCS: Contrastive learning of cardiac signals across space, time, and patients. In *ICML*, 2021. 2

[17] Leonidas Liakopoulos, Nikolaos Stagakis, Evangelia I. Zacharaki, and Konstantinos Moustakas. Cnn-based stress and emotion recognition in ambulatory settings. In *2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pages 1–8. IEEE, 2021. 11

[18] Ran Liu, Ellen Zippi, Hadi Pour Ansari, Chris Sandino, Jingping Nie, Hanlin Goh, Erdrin Azemi, and Ali Moin. Frequency-aware masked autoencoders for multimodal pretraining on biosignals. In *ICLR Workshop*, 2024. 2

[19] Yucheng Liu, Ziyu Jia, and Haichao Wang. EmotionKD: A cross-modal knowledge distillation framework for emotion recognition based on physiological signals. In *ACM MM*, 2023. 2, 3

[20] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016. 1, 2

[21] Konstantinos Makantasis, David Melhart, Antonios Liapis, and Georgios N. Yannakakis. Privileged information for modeling affect in the wild. *arXiv preprint arXiv:2107.10552*, 2021. 2

[22] Konstantinos Makantasis, Kosmas Pinitas, Antonios Liapis, and Georgios N. Yannakakis. From the lab to the wild: Affect modeling via privileged information. *IEEE Transactions on Affective Computing*, 2024. 2

[23] Konstantin Markov and Tomoko Matsui. Robust speech recognition using generalized distillation framework. In *Proc. Interspeech*, pages 2364–2368, 2016. 2

[24] Julie Mordacq, Léo Milecki, Maria Vakalopoulou, et al. ADAPT: Multimodal learning for detecting physiological changes under missing modalities. *arXiv preprint arXiv:2407.03836*, 2024. 1, 2

[25] Eric Oliver and Sagnik Dakshit. Cross-modality investigation on WESAD stress classification. *arXiv preprint arXiv:2502.18733*, 2025. 11

[26] Pooja Prajod and Elisabeth André. On the generalizability of ECG-based stress detection models. In *IEEE ICMLA*, pages 549–554, 2022. 11

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[28] L. G. Roos and G. M. Slavich. Wearable technologies for health research: Opportunities, limitations, and practical considerations. *Brain, Behavior, and Immunity*, 113:444–452, 2023. 1

[29] Roberto Sánchez-Reolid, María T. López, and Antonio Fernández-Caballero. Machine learning for stress detection from electrodermal activity: A scoping review. *Sensors*, 2025.

[30] P. Schmidt, A. Reiss, R. Dürichen, C. Marberger, and K. Van Laerhoven. Introducing WESAD: A multimodal dataset for wearable stress and affect detection. In *Proc. 20th ACM Int. Conf. on Multimodal Interaction (ICMI)*, pages 400–408, 2018. 1, 2, 3

[31] Ghanapriya Singh, Orchid Chetia Phukan, Rinki Gupta, and Anand Nayyar. Hybrid deep learning model for wearable sensor-based stress recognition for IoMT system. *Int. J. Commun. Syst.*, 37(3), 2024. 11

[32] Žiga Stržinar, Araceli Sanchis, Agapito Ledezma, Óscar Sipele, Boštjan Pregelj, and Igor Škrjanc. Stress detection using frequency spectrum analysis of wrist-measured electrodermal activity. *Sensors*, 23(2):963, 2023. 3

[33] Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16:2023–2049, 2015. 1, 2

[34] Shiqi Yang, Yuan Gao, Yao Zhu, Li Zhang, Qinlan Xie, Xuesong Lu, Fang Wang, and Zhushanying Zhang. A deep learning approach to stress recognition through multimodal physiological signal image transformation. *Scientific Reports*, 15, 2025. 11

[35] Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *NeurIPS*, 35: 3988–4003, 2022. 2

A. Comparison with Existing Work

We show a comparison of our results with some existing work in Tab. 7. Most WESAD reports either (i) rely on a single sensor (often ECG) or (ii) include EDA at inference, and many use task definitions and splits that differ from ours (e.g., multi-class vs. binary; subject-dependent vs. LOSO). Consequently, these numbers provide useful context but are not like-for-like baselines for our setting where EDA is absent at test time. Behinaein et al [5] reports 91.1% accuracy only after fine-tuning/calibration using 10% of the held-out test subject; the zero-calibration LOSO accuracy is 80.4%.

B. Additional Ablation Studies

We report additional hyperparameter sweeps and secondary ablations. Unless noted, metrics are mean $\pm$ std across 15 LOSO folds.

B.1. Modality Dropout at Test Time

We assess the sensitivity of PULSE to missing test-time modalities. Results are obtained by removing signal modalities during finetuning and inference. Relative to the full setting (ECG+BVP+ACC+TEMP), removing TEMP yields small changes. Removing ACC in addition (ECG+BVP) produces the largest additional drop. Going to ECG only leads to a further but modest decline.

Variability increases as modalities are removed (e.g., AUPRC std: $0.033 \rightarrow 0.036 \rightarrow 0.088 \rightarrow 0.117$), indicating reduced stability under stronger input constraints. Overall, PULSE degrades gracefully and maintains strong performance even with ECG only, consistent with effective transfer of privileged-sensor structure into each student.

B.2. Modality Dropout During Pretraining

Pretraining with all four cheap sensors yields the best results; dropping modalities degrades AUPRC and accuracy most sharply, indicating reduced transfer of privileged EDA structure. Compared with Tab. 9, removing modalities during pretraining degrades performance more than removing them at test time, underscoring the benefits of multimodal representation learning over single-modality training.

B.3. Modality-Agnostic Transfer: ECG as Teacher

To evaluate whether privileged knowledge transfer in PULSE depends specifically on EDA or can generalize to other physiological modalities, we conducted an ablation using **ECG as the privileged teacher**. In this setup, the pretrained ECG encoder is frozen and used to distill knowledge into the remaining deployable modalities (BVP, ACC, and TEMP) through the same shared/private alignment and hidden-state matching objectives. Tab. 11 reports the results under the LOSO protocol.

Table 7. Selected WESAD results from prior work, augmented with our results. Metrics are reported as in the original papers and are *not* directly comparable across differing task setups (binary vs. 3/4-class) and evaluation protocols.

Paper (Year)	**Test-time sensors**	**Eval protocol**	**Accuracy (%)**	**AUROC**
Ours (PULSE, 2025)	ECG + BVP + ACC + TEMP	LOSO, binary	93.97 ± 5.77	0.989 ± 0.017
Ours (Symmetric alignment, 2025)	ECG + BVP + ACC + TEMP	LOSO, binary	88.83 ± 6.24	0.972 ± 0.031
Ours (No-EDA baseline, 2025)	ECG + BVP + ACC + TEMP	LOSO, binary	91.64 ± 6.61	0.963 ± 0.050
Ours (Full-sensor baseline, 2025)	ECG + BVP + ACC + TEMP + EDA	LOSO, binary	90.74 ± 5.58	0.983 ± 0.028
[26]	ECG only	LOSO, binary	90.8	N/A
[5]	ECG only	LOSO, binary†	91.1	N/A
[15]	ECG only	LOSO, binary	88.7	N/A
[17]	ECG only	LOSO, binary	82.4	N/A
[25]	ECG / BVP / EDA / RESP / TEMP (single-modality)	Subject-dependent (random 85:15)	up to 99.95 (validation; multiclass)	N/A
[34]	EDA + ACC + BVP + RESP + TEMP	LOSO, 3-class	90.96	N/A
[31]	ECG + EDA + RESP (+TEMP/ACC)	LOSO, 3-class	90.45	N/A
[9]	EDA + ECG + ACC + RESP + TEMP	LOSO, 4-class	94.77	N/A

Table 8. Representational collapse without reconstruction. Mean pairwise cosine and feature variance across modalities.

	KD-only (no recon.)	
Modality	Cosine	Variance
ECG	1.000	2.57e-05
BVP	1.000	6.26e-05
ACC	1.000	7.53e-05
TEMP	0.998	7.96e-04
	With reconstruction	
ECG	0.070	2.52e-02
BVP	0.065	1.51e-02
ACC	0.137	9.20e-02
TEMP	0.027	2.06e-03

Using ECG as the teacher yields moderate improvements in AUROC and AUPRC compared to the no-KD counterpart, while accuracy remains comparable within variance. This demonstrates that the PULSE framework is *modality-agnostic*: privileged distillation benefits are not limited to EDA, though the magnitude of gain depends on the informativeness of the teacher modality. EDA remains the strongest privileged signal for stress-related supervision due to its direct coupling with sympathetic arousal, yet ECG-based distillation still provides measurable transfer, confirming that PULSE can flexibly leverage alternative privileged channels when EDA is unavailable.

B.4. Hinge-Loss Margin

A modest margin (0.2) provides the strongest signal for aligning student shared embeddings with the EDA teacher; margins that are too small (0) or too large (≥ 0.6) underperform.

B.5. Shared–Private Capacity (Full Sweep)

Both insufficient private capacity (ratio 0: no private embeddings to encode modality-specific features) and excessive private capacity (ratio 1: no shared embeddings for alignment and knowledge transfer) hurt performance. A *balanced* split between shared and private capacity (0.5) most effectively supports privileged EDA transfer, while a smaller private fraction (0.2) trades a slight drop in AUROC/AUPRC for the highest accuracy.

B.6. Fusion Strategy

We examined whether learning per-modality fusion weights could improve performance compared to static averaging of modality embeddings. In the **adaptive fusion** variant, a small trainable gating module predicts fusion weights for each modality during both knowledge distillation and fine-tuning, initialized to uniform values and updated end-to-end. The default PULSE configuration uses a **static average** of shared embeddings across modalities.

Adaptive fusion offered no consistent improvement and slightly degraded performance. We attribute this to the high inter-subject variability under LOSO evaluation: fusion weights learned on training subjects do not necessarily generalize to unseen participants. Static averaging enforces

Table 9. Ablations on test-time input availability for PULSE. Reported means and standard deviations are computed across folds.

Name	Test-time inputs	AUROC	AUPRC	Accuracy (%)
PULSE (full)	ECG + BVP + ACC + TEMP	$\mathbf{0.989 \pm 0.017}$	$\mathbf{0.977 \pm 0.033}$	$\mathbf{93.97 \pm 5.77}$
PULSE (−TEMP)	ECG + BVP + ACC	0.985 ± 0.021	0.972 ± 0.036	90.85 ± 6.53
PULSE (−TEMP, −ACC)	ECG + BVP	0.966 ± 0.072	0.948 ± 0.088	87.74 ± 8.53
PULSE (ECG only)	ECG	0.960 ± 0.082	0.936 ± 0.117	86.22 ± 10.37

Table 10. Effect of removing modalities during pretraining. Full-sensor pretraining yields the strongest downstream performance.

Setup	AUROC	AUPRC	Accuracy
PULSE (ECG + BVP + ACC + TEMP)	$\mathbf{0.989 \pm 0.017}$	$\mathbf{0.977 \pm 0.033}$	$\mathbf{93.97 \pm 5.77}$ %
ECG + BVP + ACC	0.949 ± 0.084	0.923 ± 0.117	86.12 ± 8.22 %
ECG + BVP	0.973 ± 0.048	0.942 ± 0.112	90.44 ± 9.92 %
Only ECG	0.960 ± 0.077	0.927 ± 0.137	75.74 ± 10.81 %

a more stable modality combination and avoids overfitting to subject-specific patterns.

B.7. Alignment Loss

To quantify the contribution of the cross-modal alignment objective, we conducted an ablation removing $\mathcal{L}_{\text{align}}$ during pretraining. In this setting, each modality's encoder is trained solely via its reconstruction objective (MAE) without enforcing consistency across shared embeddings.

Removing the alignment loss leads to a sharp drop in both ranking metrics and accuracy, confirming that explicit cross-modal regularization is essential for learning a coherent shared latent space.

B.8. Hidden-State Matching (Full Results)

To assess the effect of hidden-state supervision during knowledge distillation, we performed ablations over which layers of the student were matched to the frozen EDA teacher. We compare three configurations: (i) removing hidden-state matching entirely ($\mathcal{L}_{\text{hid}} = 0$), equivalent to final-representation-only transfer; (ii) matching only intermediate layers (3, 5, 7); and (iii) matching all hidden layers.

Removing hidden-state matching degrades all metrics relative to both PULSE and the no-EDA baseline, indicating that intermediate feature supervision is critical for transferring sympathetic-arousal structure. Supervising *all* layers yields the highest AUROC, AUPRC, and accuracy, suggesting that multi-depth guidance stabilizes optimization and enables richer cross-modal transfer.

C. Why Hinge Loss (vs. Anchored Contrastive)

During pretraining we first tried an **anchored contrastive** objective: each modality's shared embeddings were aligned to ECG (the "anchor") via cosine similarity. In practice, the alignment loss collapsed to $\sim$0 within the first few iterations and then remained nearly constant, providing little training signal thereafter, as shown in Fig. 5. This early collapse motivated switching to a **hinge-based alignment** with a small positive margin, which maintained non-trivial gradients throughout training.

D. Training Dynamics

In Figs. 6 to 8, we show the convergence behavior in different stages of training (pretraining, knowledge transfer, and finetuning).

Table 11. ECG as frozen teacher distilled to BVP/ACC/TEMP students.

Setup	AUROC	AUPRC	Accuracy (%)
ECG teacher → (BVP, ACC, TEMP)	$\mathbf{0.956 \pm 0.067}$	$\mathbf{0.930 \pm 0.085}$	86.63 ± 5.86
(BVP, ACC, TEMP) w/o KD	0.949 ± 0.058	0.902 ± 0.102	$\mathbf{87.06 \pm 8.77}$

Table 12. Effect of hinge-loss margin for shared-embedding alignment. A small positive margin (0.2) is optimal; too large a margin degrades performance, likely by over-separating positive pairs and harming transfer.

Margin	**AUROC**	**AUPRC**	**Accuracy**
0.0	0.972 ± 0.039	0.954 ± 0.060	$70.41 \pm 8.60\,\%$
0.2 (default)	$\mathbf{0.989 \pm 0.017}$	$\mathbf{0.977 \pm 0.033}$	$\mathbf{93.97 \pm 5.77}\,\%$
0.4	0.977 ± 0.051	0.966 ± 0.071	$89.08 \pm 13.39\,\%$
0.6	0.983 ± 0.052	0.975 ± 0.070	$77.57 \pm 8.85\,\%$
0.8	0.957 ± 0.077	0.938 ± 0.112	$70.41 \pm 8.60\,\%$
1.0	0.969 ± 0.070	0.933 ± 0.160	$76.43 \pm 8.69\,\%$

Table 13. Varying the private-mask ratio that reserves capacity for modality-specific features. A balanced split (0.5) maximizes AUROC/AUPRC, while a slightly smaller private region (0.2) peaks accuracy. Removing knowledge transfer (ratio 1) hurts all metrics.

Private mask ratio	**AUROC**	**AUPRC**	**Accuracy**
0	0.959 ± 0.062	0.927 ± 0.100	$70.41 \pm 8.60\,\%$
0.2	0.980 ± 0.044	0.966 ± 0.076	$\mathbf{94.46 \pm 7.78}\,\%$
0.4	0.919 ± 0.116	0.894 ± 0.148	$79.51 \pm 13.92\,\%$
0.5 (default)	$\mathbf{0.989 \pm 0.017}$	$\mathbf{0.977 \pm 0.033}$	$93.97 \pm 5.77\,\%$
0.6	0.967 ± 0.072	0.956 ± 0.084	$92.62 \pm 9.89\,\%$
0.8	0.975 ± 0.047	0.958 ± 0.082	$93.76 \pm 5.96\,\%$
1 (no knowledge transfer)	0.945 ± 0.078	0.906 ± 0.127	$70.41 \pm 8.60\,\%$

Table 14. Fusion strategy during finetuning.

Model	AUROC	AUPRC	Accuracy (%)
PULSE (static fusion; default)	$\mathbf{0.989 \pm 0.017}$	$\mathbf{0.977 \pm 0.033}$	$\mathbf{93.97 \pm 5.77}$
PULSE (adaptive fusion)	0.967 ± 0.046	0.951 ± 0.064	90.91 ± 7.35

Table 15. Ablation on alignment loss $\mathcal{L}_{\text{align}}$ during pretraining. Removing alignment substantially degrades performance, indicating that shared-space regularization is critical for multimodal representation learning.

Model	AUROC	AUPRC	Accuracy (%)
no-EDA baseline (with alignment loss)	$\mathbf{0.963 \pm 0.050}$	$\mathbf{0.937 \pm 0.101}$	$\mathbf{91.64 \pm 6.61}$
no-EDA baseline (without alignment loss)	0.945 ± 0.078	0.906 ± 0.127	70.41 ± 8.60

Table 16. Effect of hidden-state matching during KD. Full results with standard deviations.

Model	AUROC	AUPRC	Accuracy (%)
PULSE (match layers 3,5,7)	0.989 ± 0.017	0.977 ± 0.033	93.97 ± 5.77
PULSE ($\mathcal{L}_{\text{hid}}$=0; final-only)	0.953 ± 0.096	0.922 ± 0.155	90.95 ± 9.73
PULSE (match all layers)	$\mathbf{0.994 \pm 0.011}$	$\mathbf{0.988 \pm 0.022}$	$\mathbf{96.08 \pm 4.52}$
no-EDA baseline (no KD)	0.963 ± 0.050	0.937 ± 0.101	91.64 ± 6.61

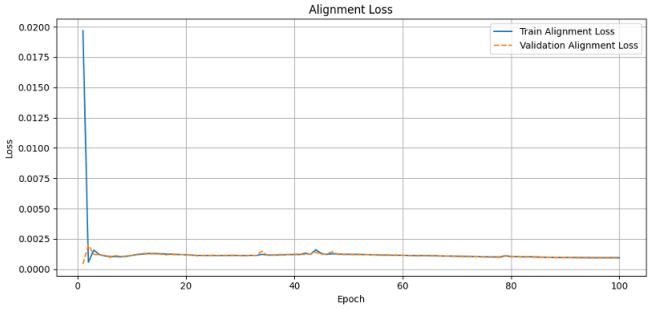

Figure 5. Learning curve showing anchored contrastive loss in early experiments.

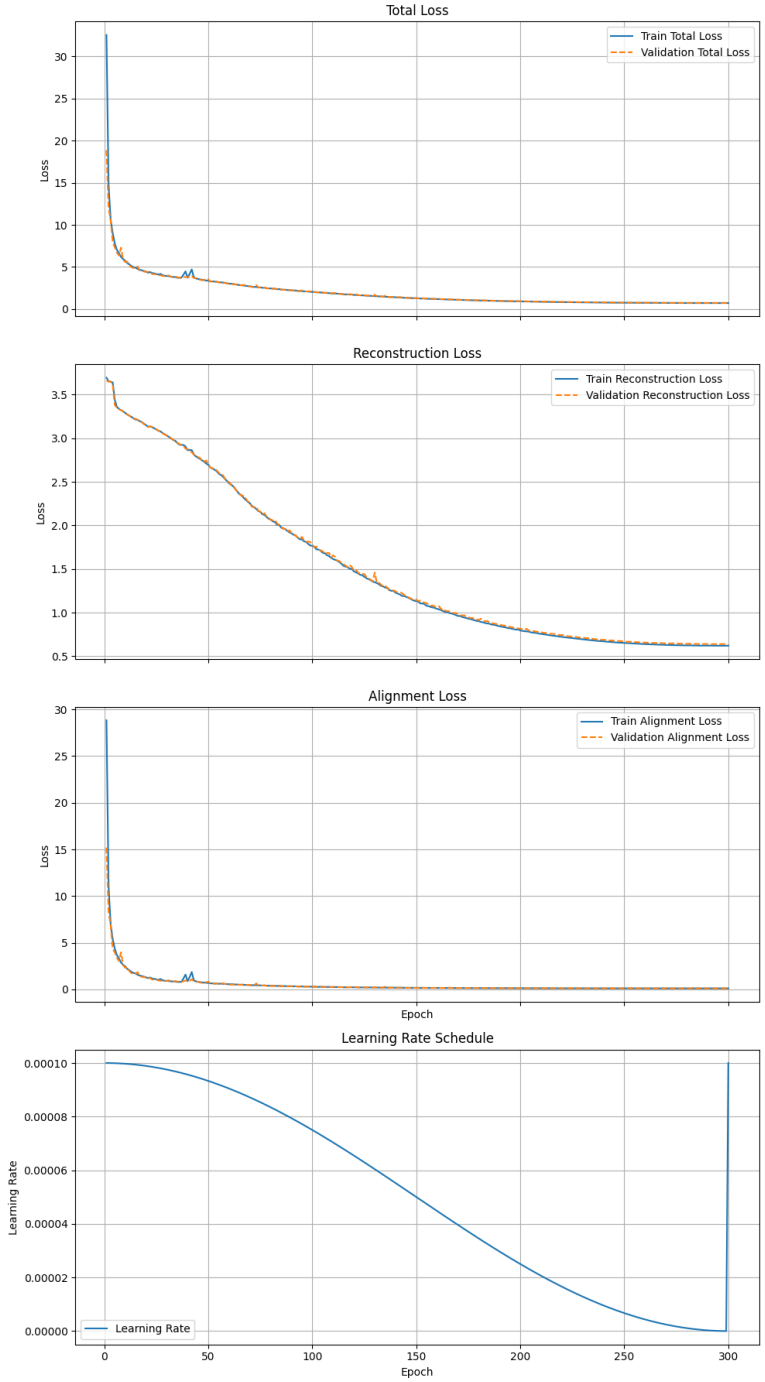

Figure 6. Convergence behavior in pretraining.

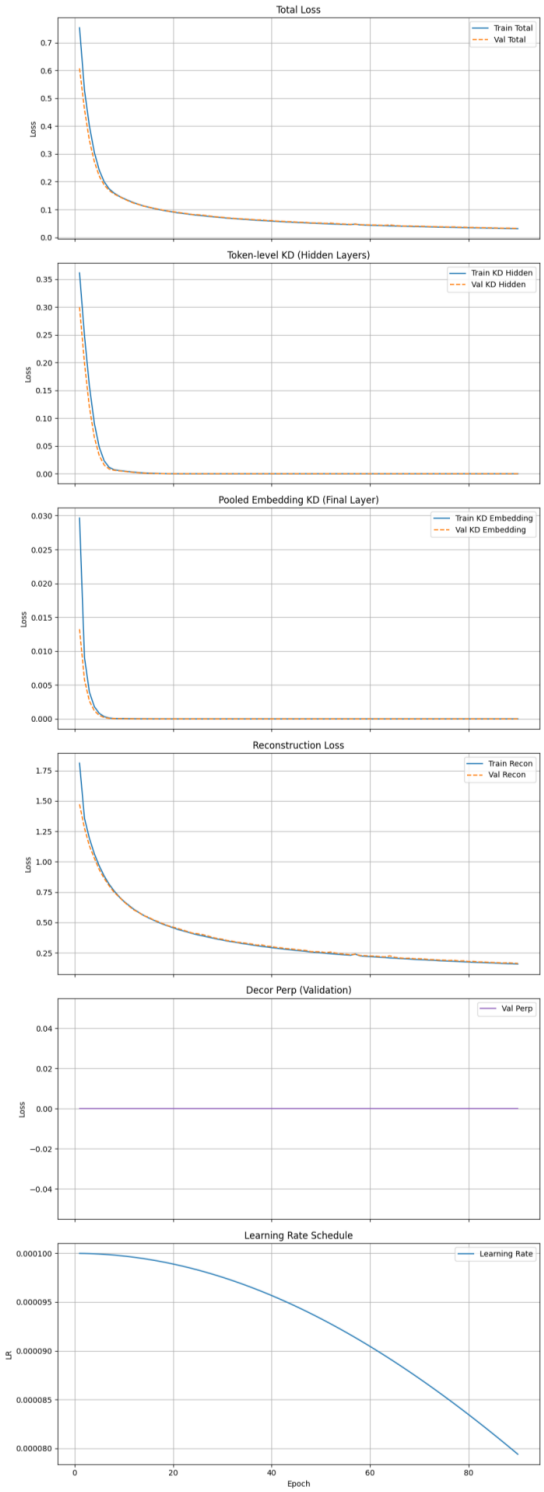

Figure 7. Convergence behavior in knowledge transfer.

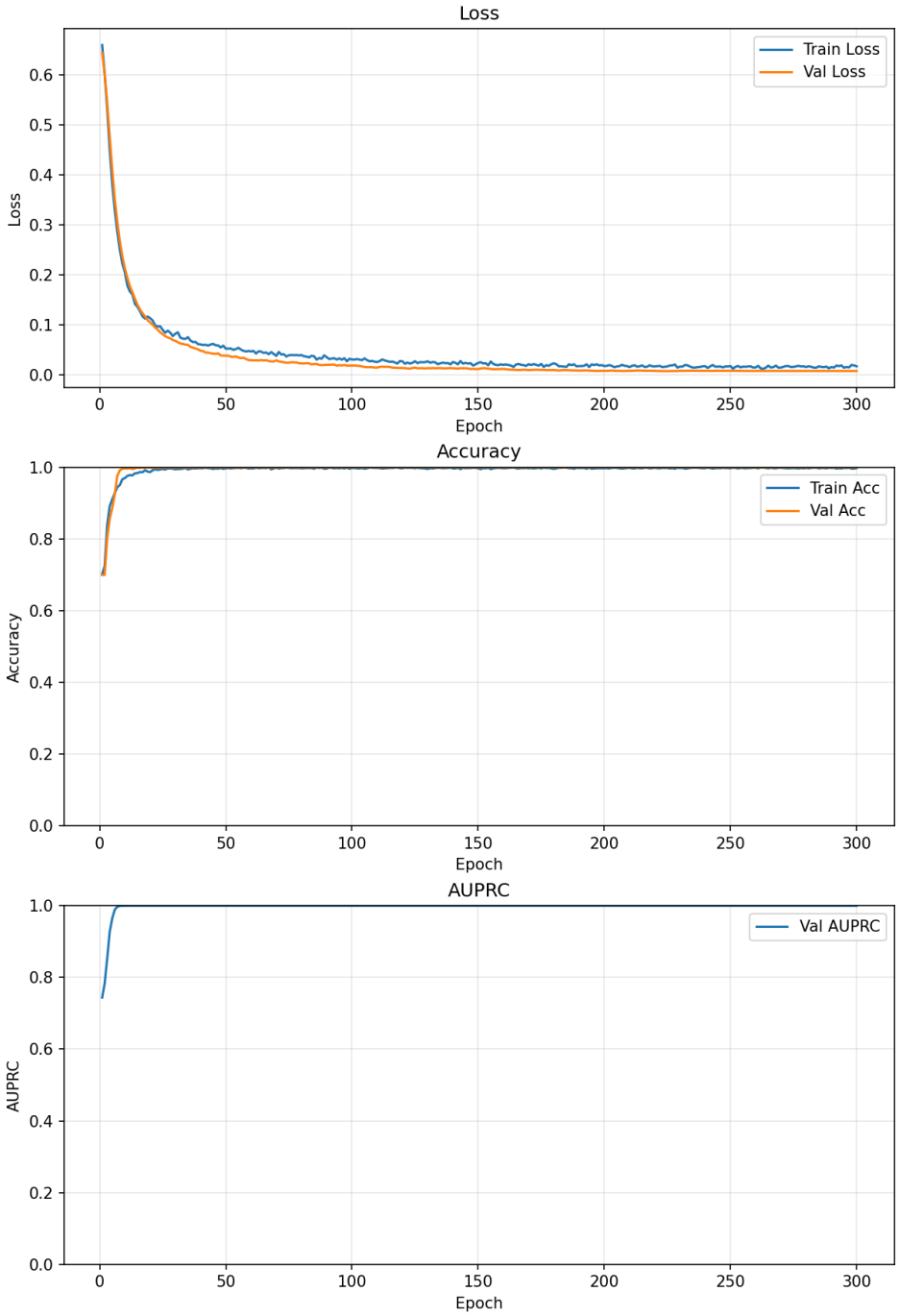

Figure 8. Convergence behavior in finetuning.