FLOW: Feature-Level Optimal Warping for Generalized Remote Physiological Measurement

Supplementary Material

A. Introduction to the Datasets

UBFC-rPPG [1] contains 42 RGB facial videos from 42 distinct subjects. Each video is captured at 640×480 pixel resolution and 30 frames per second (fps). Recordings take place under varied lighting conditions, including natural sunlight and indoor artificial illumination. Ground-truth physiological signals are recorded via a CMS50E pulse oximeter at 60 Hz, ensuring precise temporal alignment for evaluation.

PURE [34] comprises 60 high-quality RGB videos collected from 10 subjects performing six different head movement scenarios (static, talking, translation movements, etc.). Videos are recorded at 30 fps under consistent indoor lighting and controlled background settings, minimizing external interference. Synchronized physiological measurements are obtained using a CMS50E oximeter sampling at 60 Hz. PURE is particularly valuable for evaluating rPPG performance during facial movements.

BUAA-MIHR [45] is designed to assess algorithmic robustness across varying illumination intensities. The dataset features video sequences recorded under a range of controlled lighting conditions, from low-light (below 10 lux) to normal brightness. In our experiments, we only utilize videos captured under illumination levels $\geq$10 lux, as extremely dim lighting introduces significant image degradation requiring specialized enhancement techniques beyond this study's scope.

MMPD [?] comprises 660 videos, each lasting one minute, collected from 33 subjects with diverse skin tones and gender distributions. Each video is recorded at 30 fps with a resolution of 320×240 pixels, under four distinct lighting conditions (bright, warm, dim, and colored lighting). Subjects perform various daily activities, introducing intra-subject variability and further increasing dataset complexity.

B. Experimental Settings

Datasets and Evaluation Metrics We evaluate our method on four widely used remote photoplethysmography (rPPG) datasets: **UBFC-rPPG** [2], **PURE** [35], **BUAA-MIHR** [46] and **MMPD** [38]. Following prior works [37, 50], we adopt three standard evaluation metrics: mean absolute error (MAE), root mean square error (RMSE), and Pearson's correlation coefficient (R), to assess the accuracy of predicted heart rates (HRs). For both MAE and RMSE, lower values indicate smaller prediction errors, while higher values of R (closer to 1.0) indicate stronger linear correlation with the ground-truth HRs. MAE and RMSE are reported in beats per minute (bpm); for brevity, we omit these units in subsequent tables and discussions.

Implementation Details Our experiments are implemented in PyTorch, primarily based on the rPPG-Toolbox [20]. For preprocessing, we detect and crop the face region from the first frame of each video clip and apply a fixed bounding box across subsequent frames. Each video is resampled to a consistent frame rate of 30 fps, and a random chunk of 128 frames is selected, resized to 128×128 pixels. We adopt the PhysLLM [48] as the baseline. The hyperparameters $\alpha = 0.8$ and $l_{target} = 32$ are set by default. The LLM is trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 5×10^{-5}. The entire model is trained for 20 epochs on an NVIDIA H100 GPU with a batch size of 4.

C. Theoretical Proof of the Generalization Bound

In this section, we provide a detailed derivation of the proposed multi-source generalization bound under the conditional optimal transport (OT) geometry used in FLOW. Our goal is to connect the quality of cross-domain alignment—measured under a task-aware OT cost—to the prediction risk on an unseen target domain.

Preliminaries

Data and hypothesis space. Let $\{D_s\}_{s=1}^m$ be the source domains with mixing coefficients $\alpha_s \geq 0$ such that $\sum_{s=1}^m \alpha_s = 1$, and let T denote the (target) test domain. Each sample is denoted by $z = (x, hr)$, where $x \in \mathbb{R}^d$ is the visual-temporal feature and $hr \in \mathbb{R}$ is the continuous heart rate label. A hypothesis $h \in \mathcal{H}$ produces a prediction $\hat{y}$ from x. We assume a bounded loss $\ell(h(x), y) \in [0, 1]$. For a distribution P, the expected risk is

$$R_P(h) = \mathbb{E}_{(z,y)\sim P}[\ell(h(x), y)]. \tag{17}$$

Task-driven conditional cost. Following our formulation in the main paper, we define a conditional ground cost c that jointly accounts for feature distance and physiological consistency:

$$c\big((x, hr), (x', hr')\big) = \|x - x'\|_W^2 + \lambda_{hr}\Big(1 - \exp\big(-\tfrac{(hr-hr')^2}{2\sigma^2}\big)\Big), \tag{18}$$

where $W = \mathrm{diag}(w_1, \ldots, w_d)$ is the learnable frequency/feature weighting matrix, and $\lambda_{hr}, \sigma > 0$ control the influence and bandwidth of the heart-rate kernel term. This cost emphasizes both semantic similarity in feature space and coherence in underlying physiological signals.

Conditional OT distance. Given two distributions P, Q on (x, hr), we define the conditional OT distance:

$$W_c(P, Q) = \inf_{\pi \in \Pi(P,Q)} \int c(z, z')\, d\pi(z, z'), \tag{19}$$

where $\Pi(P, Q)$ is the set of couplings with marginals P and Q. Throughout, we assume c induces a valid (or pseudo-)metric structure compatible with the OT geometry.

OT barycenter over sources. We consider the (conditional) OT barycenter B of $\{D_s\}_{s=1}^m$:

$$B = \arg\min_{\nu} \sum_{s=1}^{m} \alpha_s W_c(D_s, \nu). \tag{20}$$

Intuitively, B captures a geometry-aware "anchor" distribution that balances all source domains under W_c.

Regularized OT and residual terms. In practice, we employ a debiased Sinkhorn divergence to approximate W_c, together with structure-preserving regularization terms (e.g., identity-preserving constraints) used in FLOW. We collect these deviations into:

- Δ_{sink}: the residual bias between the ideal W_c and its debiased Sinkhorn approximation;
- Δ_{id}: the residual bias introduced by identity-preserving regularization (and related structure-preserving penalties) that slightly perturb the ideal OT geometry.

Both Δ_{sink} and Δ_{id} are treated as small non-negative constants controlled by regularization strength and optimization accuracy.

Risk Discrepancy under Conditional OT

We first relate the risk difference between two domains to their conditional OT distance.

Lemma 1 (Task-related discrepancy bound). Let $g(z) = \mathbb{E}[\ell(h(x), y) \mid z]$. Assume g is L_c-Lipschitz with respect to the metric induced by c, i.e.,

$$|g(z) - g(z')| \le L_c\, d_c(z, z') \le L_c\, c(z, z'). \tag{21}$$

Then for any distributions P, Q,

$$|R_P(h) - R_Q(h)| \le L_c\, W_c(P, Q) + \Delta_{\text{id}}. \tag{22}$$

Proof. Let $\pi^\star \in \Pi(P, Q)$ be an optimal coupling for W_c. Then

$$R_P(h) - R_Q(h) = \mathbb{E}_P[g(z)] - \mathbb{E}_Q[g(z')] = \iint (g(z) - g(z'))\, d\pi^\star(z, z'). \tag{23}$$

By the Lipschitz property, $|g(z) - g(z')| \le L_c\, c(z, z')$, thus

$$|R_P(h) - R_Q(h)| \le L_c \iint c(z, z')\, d\pi^\star(z, z') = L_c W_c(P, Q) + \Delta_{\text{id}}, \tag{24}$$

where Δ_{id} accounts for the slight distortion introduced by identity-preserving regularization in the practical alignment. □

From Sources to Barycenter and Target

Using Lemma 1, we control the risk at the barycenter B and then transfer it to the target domain T.

Lemma 2 (Multi-source to barycenter). For any $h \in \mathcal{H}$,

$$R_B(h) \le \sum_{s=1}^{m} \alpha_s R_{D_s}(h) + L_c \sum_{s=1}^{m} \alpha_s W_c(D_s, B) + \Delta_{\text{id}}. \tag{25}$$

Proof. Apply Lemma 1 with $(P, Q) = (D_s, B)$ and then average with weights α_s:

$$R_B(h) \le \sum_{s} \alpha_s R_{D_s}(h) + L_c \sum_{s} \alpha_s W_c(D_s, B) + \Delta_{\text{id}}. \tag{26}$$

□

Lemma 3 (Barycenter to target). For any $h \in \mathcal{H}$,

$$R_T(h) \le R_B(h) + L_c W_c(B, T) + \Delta_{\text{id}}. \tag{27}$$

Proof. Apply Lemma 1 with $(P, Q) = (B, T)$ directly. □

Lemma 4 (Two-hop barycenter inequality). Assume W_c satisfies the triangle inequality and B is the OT barycenter defined above. Then

$$\sum_{s=1}^{m} \alpha_s W_c(D_s, B) + W_c(B, T) \le 2 \sum_{s=1}^{m} \alpha_s W_c(D_s, T). \tag{28}$$

Sketch. By barycenter optimality, for any ν (in particular $\nu = T$),

$$\sum_{s} \alpha_s W_c(D_s, B) \le \sum_{s} \alpha_s W_c(D_s, T). \tag{29}$$

By triangle inequality, $W_c(D_s, T) \le W_c(D_s, B) + W_c(B, T)$, hence

$$W_c(B, T) \le W_c(D_s, T) - W_c(D_s, B) + \text{(non-negative term)}. \tag{30}$$

Aggregating over s and combining the two relations yields an upper bound of the two-hop quantity by a constant factor of the direct discrepancies $\{W_c(D_s, T)\}$; the above inequality is a convenient sufficient form. A fully rigorous derivation can be obtained by exploiting convexity of OT distances and the characterization of Wasserstein barycenters; we omit the routine details for brevity. □

Conditional OT Generalization Bound

We now combine the above lemmas to obtain the main bound.

Theorem 1 (Conditional OT barycenter generalization bound). For any $h \in \mathcal{H}$,

$$R_T(h) \le \sum_{s=1}^{m} \alpha_s R_{D_s}(h) + L_c\Big(\sum_{s=1}^{m} \alpha_s W_c(D_s, B) + W_c(B, T)\Big) + \Delta_{\text{sink}} + \Delta_{\text{id}} + \Delta_{\text{est}}. \tag{31}$$

where Δ_{est} denotes standard statistical estimation errors from finite samples.

Proof. Starting from Lemma 3,

$$R_T(h) \le R_B(h) + L_c W_c(B, T) + \Delta_{\text{id}}, \tag{32}$$

then substitute Lemma 2 for $R_B(h)$ to obtain

$$R_T(h) \le \sum_{s} \alpha_s R_{D_s}(h) + L_c \sum_{s} \alpha_s W_c(D_s, B) + L_c W_c(B, T) + \Delta_{\text{id}}. \tag{33}$$

Approximating W_c with the debiased Sinkhorn divergence contributes an additional Δ_{sink}, and empirical estimation of risks contributes Δ_{est}, leading to the stated inequality. □

Comparison with Traditional Geometric Bounds

We compare our conditional OT-based discrepancy with the standard Euclidean Wasserstein-1 bound for multi-source domain generalization.

Classical bound. A typical geometric bound using the Wasserstein-1 distance W_1 has the form:

$$R_T(h) \le \sum_{s=1}^{m} \alpha_s R_{D_s}(h) + L_1 \sum_{s=1}^{m} \alpha_s W_1(D_s, T) + \lambda^\star + \tilde{\Delta}, \tag{34}$$

where L_1 is a Lipschitz constant under the Euclidean metric, $\lambda^\star$ denotes an irreducible joint error term, and $\tilde{\Delta}$ collects statistical/approximation errors.

Theorem 2 (Relative tightness under conditional geometry). Assume: (i) bounded feature domain $\|x\| \le M_x$; (ii) bounded weight matrix $\|W\|_{\text{op}} \le \Lambda_W$; (iii) heart-rate kernel term is L_{hr}-Lipschitz; and (iv) the conditional metric removes irrelevant (non-physiological) directions so that there exists $0 < r < 1$ with $L_c \le r L_1$. Then there exists a constant $\kappa > 0$ such that

$$W_c(P, Q) \le \kappa W_1(P, Q), \tag{35}$$

and, combining Lemma 4 with Theorem 1, our discrepancy term satisfies

$$\begin{aligned} \text{Diff}_{\text{ours}} &\lesssim 2L_c \sum_{s=1}^{m} \alpha_s W_c(D_s, T) \\ &\le 2r\kappa L_1 \sum_{s=1}^{m} \alpha_s W_1(D_s, T). \end{aligned} \tag{36}$$

up to $(\Delta_{\text{sink}} + \Delta_{\text{id}} + \Delta_{\text{est}})$. Under standard boundedness and normalization assumptions, both r and κ remain moderate constants. In particular, when $2r\kappa < 1$, which is a mild condition that can be encouraged in practice via feature normalization and regularization, the conditional OT geometry yields a strictly tighter or comparable discrepancy term than the Euclidean Wasserstein-1 counterpart. □

D. Training and Inference Procedure of FLOW

In this section, we provide additional details regarding the training and inference pipeline of **FLOW**. During training, TRM first stabilizes intermediate representations, after which PCOT computes soft cross-temporal correspondences to align features across domains, as summarized in Algorithm 1. The regularization terms jointly ensure consistent prototype usage and preserve feature identity throughout the alignment process. At inference time, FLOW operates without requiring domain labels or additional preprocessing, producing temporally coherent and domain-invariant representations that support robust physiological signal prediction, as illustrated in Algorithm 2. For clarity, we summarize the full procedure in the following subsections.

Algorithm 1: Training of FLOW

Input: Multi-source videos $\{X_i\}$, domains $\{d_i\}$, rPPG signals $\{s_i\}$; hyper-parameters $\lambda_{OT}, \lambda_{sc}, \lambda_{\text{id}}, \alpha, \beta, \eta, E, B$.

Output: Parameters $\theta_b, \theta_t, \theta_p, \theta_r$; prototype bank $P = \{P_d\}$.

Initialize $f_{\text{backbone}}, f_{\text{TRM}}$, PCOT, h_{reg}, optimizer, P;

for $epoch = 1$ **to** E **do**

  sample minibatch $\{(X_i, d_i, s_i)\}_{i=1}^{B}$;

  `// 1. HR labels from FFT`

  $y_i \leftarrow \text{FFT\_HR}(s_i)$ for all i;

  `// 2. Backbone + TRM`

  $F_i \leftarrow f_{\text{backbone}}(X_i; \theta_b)$;

  $C_i \leftarrow f_{\text{TRM}}(F_i; \theta_t)$;

  flatten $\{C_i\}$ to $F_{\text{flat}} = \{F_n\}_{n=1}^{N}$, repeat d_i, y_i to $d_{\text{flat}}, y_{\text{flat}}$;

  `// 3. PCOT with label-aware cost`

  for $n = 1$ **to** N **do**

    select domain prototypes P_n from $P_{d_{\text{flat}}[n]}$;

    for $j = 1$ **to** K **do**

      $C[n, j] = \alpha \, \text{dist}(F_n, P_n[j]) + \beta \, |y_{\text{flat}}[n] - \bar{y}_{P_n[j]}|$;

  compute transport plan $\Pi^\star = \text{OT\_Solver}(C)$;

  $\mathcal{L}_{OT} = \sum_{n,j} \Pi^\star_{n,j} C[n, j]$;

  $\mathcal{L}_{sc} = \frac{1}{N} \sum_n \min_j \text{dist}(F_n, P_n[j])^2$;

  $A_n = \sum_j \Pi^\star_{n,j} P_n[j]$ for all n;

  `// 4. Regression loss`

  aggregate $\{A_n\}$ by sample to get A_i;

  $\hat{y}_i = h_{\text{reg}}(A_i; \theta_r)$;

  $\mathcal{L}_{\text{task}} = \frac{1}{B} \sum_i (\hat{y}_i - y_i)^2$;

  `// 5. Identity-preserving loss`

  $\mathcal{L}_{\text{id}} = \frac{1}{N} \sum_{n=1}^{N} \|A_n - F_n\|_2^2$;

  `// 6. Total loss and update`

  $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{OT} \mathcal{L}_{OT} + \lambda_{sc} \mathcal{L}_{sc} + \lambda_{\text{id}} \mathcal{L}_{\text{id}}$;

  update $\theta_b, \theta_t, \theta_p, \theta_r$ by SGD on $\mathcal{L}$;

  $P \leftarrow \text{UpdatePrototypes}(P, \{A_n\}, d_{\text{flat}})$;

return $\theta_b, \theta_t, \theta_p, \theta_r, P$;

Algorithm 2: Inference of FLOW

Input: Trained $\theta_b, \theta_t, \theta_p, \theta_r$, prototype bank P, test video X.

Output: Predicted heart rate $\hat{y}$.

`// 1. Backbone + TRM`

$F \leftarrow f_{\text{backbone}}(X; \theta_b)$;

$C \leftarrow f_{\text{TRM}}(F; \theta_t)$;

flatten C to $F_{\text{flat}} = \{F_n\}_{n=1}^{N}$;

`// 2. PCOT`

for $n = 1$ **to** N **do**

  choose prototypes P_n (e.g., corresponding domain or all domains);

  for $j = 1$ **to** K **do**

    $C[n, j] = \text{dist}(F_n, P_n[j])$;

$\Pi = \text{OT\_Solver}(C)$ (or nearest-prototype);

$A_n = \sum_j \Pi_{n,j} P_n[j]$ for all n;

`// 3. Regression`

aggregate $\{A_n\}$ to video-level A;

$\hat{y} = h_{\text{reg}}(A; \theta_r)$;

return $\hat{y}$;