Towards Uncertainty-aware Unsupervised Domain Adaptation for Videos and Time-Series with Causal Optimal Transport

Supplementary Material

5. Related Work

5.1. UDA in 1D Time Series

Labeling time-series data is typically labor-intensive due to the lack of explicit semantic cues, making unsupervised domain adaptation (UDA) a crucial strategy [16]. This is especially true when adapting models across domains such as different users or devices, where obtaining labeled data is costly. CoDATS [39], built upon DANN [11], incorporates adversarial training to align source and target domains and introduces weak supervision (WS) by minimizing the Kullback-Leibler (KL) divergence between target predictions and coarse label distributions. This is particularly relevant in domains like human activity recognition, where users may provide approximate distributions of their activity over time. However, this KL-based formulation is heuristic and lacks a formal probabilistic foundation. TransPL [15] integrates weak supervision into the pseudo-labeling process through Bayes' rule, offering a principled way to incorporate prior knowledge of target label distributions directly into the optimization framework.

Recent methods have tackled domain shifts in time series using various perspectives. SASA [5] aligns sparse associative structures using attention-based relationships between domains. RAINCOAT [13] exploits both time-domain and frequency-domain features for alignment, arguing that frequency representations show stronger domain invariance. CauDiTS [36] attempts to separate causal and spurious components in time-series data, under the assumption that causal features generalize across domains. However, these methods often overlook temporal state transitions, channel-level variability, or lack interpretability in the adaptation process.

A concurrent approach, SSSS-TSA [1], targets channel-level variations via a self-attention mechanism that identifies domain-relevant channels and employs distinct encoders and classifiers for each. Though this shares a similar motivation with our work, Causal-OT takes a fundamentally different approach by using Granger-causal graphs to model variable dependencies and guiding the adaptation process through causally regularized optimal transport.

Causal-OT outperforms existing methods by aligning not just marginal distributions but also the underlying causal structures, thereby avoiding spurious correlations. It further enhances pseudo-label reliability and robustness under domain shifts by preserving temporal dependencies and providing interpretable insights into cross-domain generalization.

5.2. UDA in Videos

Unlike image-based domain adaptation (DA), video-based DA remains relatively under-explored, with only a handful of studies addressing small-scale settings involving limited category overlap [10, 28, 40, 41]. Prior efforts include reducing background bias to improve generalization, projecting source and target videos into a shared feature space using shallow networks [9, 38], and adapting C3D features on a PCA-derived Grassmann manifold (AMLS). However, these datasets are too small to exhibit substantial domain shift, making it difficult to rigorously evaluate DA performance. To overcome this limitation, we used two large-scale cross-domain benchmarks, like UCF- HMDB (full) and Kinetics-Gameplay [7] and report results using multiple baseline methods. Our proposed framework jointly attends to temporal cues, aligns intermediate representations, and explicitly encodes temporal dynamics, leading to more robust video-domain adaptation.

5.3. Pseudo Labeling in UDA

In UDA, pseudo labeling refers to generating artificial labels (pseudo labels) for unlabeled target samples, which are then used as supervision to adapt the model to the target domain. Pseudo labeling methods generally fall into three broad categories. First, confidence-based methods rely on prediction confidence from a classifier trained on the source domain. For instance, Softmax-based selection [20] and agreement-based methods, like ATT [32] assign pseudo labels based on high-confidence predictions or consensus between multiple classifiers. Second, prototype-based methods, such as Nearest Class Prototype (NCP) [37] align target samples with learned source prototypes in the embedding space. Third, clustering-based strategies group unlabeled target samples based on their feature similarity [37]. Hybrid approaches, like SHOT [22] and T2PL [24] combine clustering with confidence-based filtering to discard noisy pseudo labels.

Despite their success in vision domains, most of these methods are not well-suited for time-series data [14], which pose unique challenges. Time-series exhibit strong temporal dependencies and complex multi-channel structures that static-data pseudo labeling approaches fail to exploit effectively. Ignoring these aspects can result in misleading pseudo labels and degraded adaptation performance.

To address these limitations, TransPL [15] introduces a novel pseudo-labeling framework that explicitly models temporal state transitions and channel-wise dynamics using vector-quantized (VQ) codebooks and transition matrices.

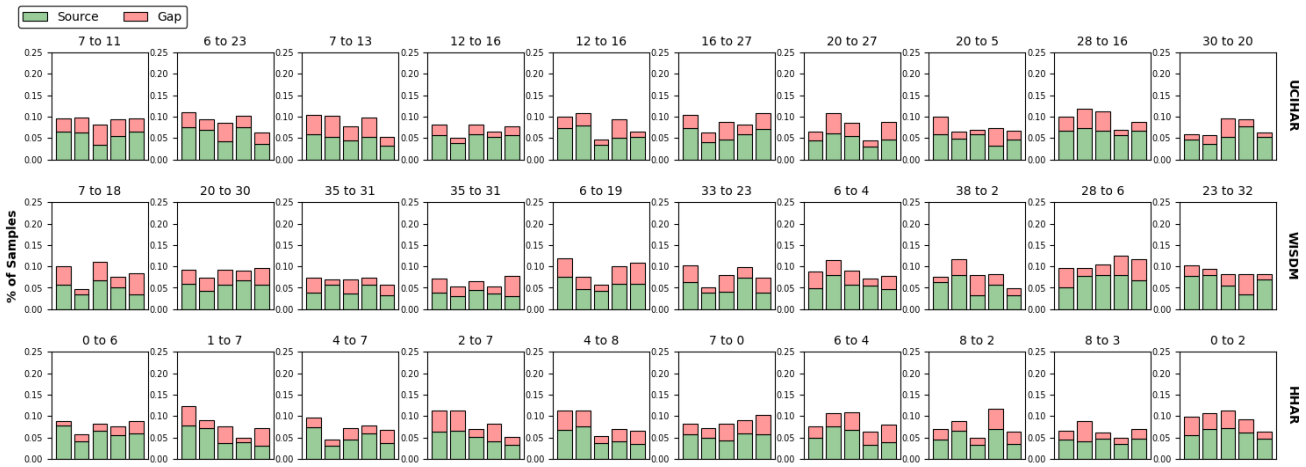
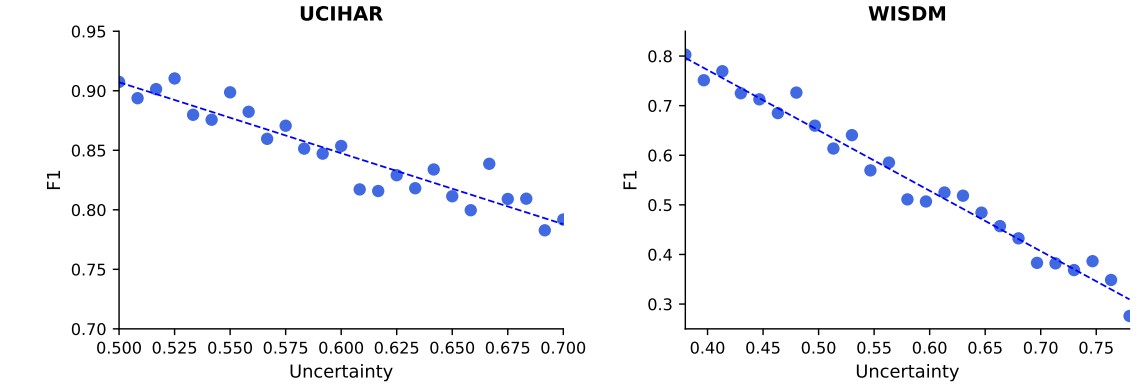
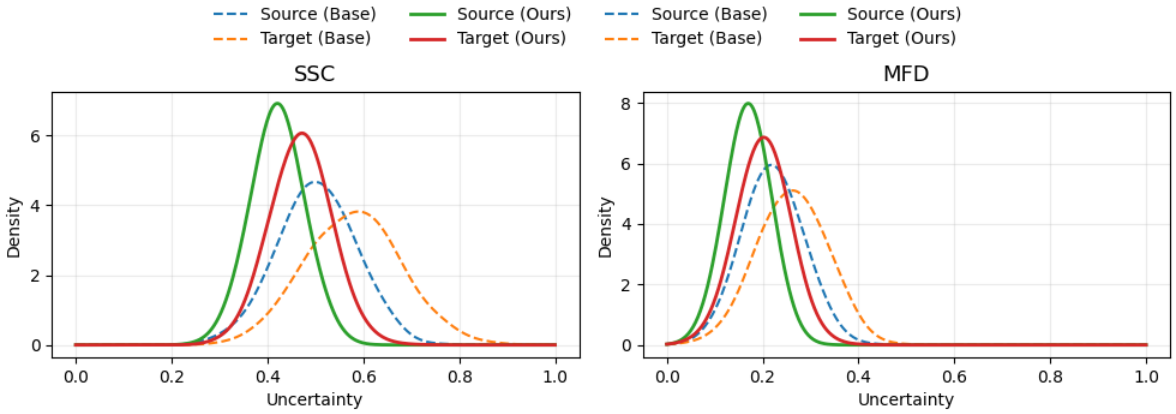


Figure 7. Class-wise label distributions of the source domain across multiple dataset splits, with red bars indicating the distribution gap relative to the target domain. Each subplot corresponds to a specific source–target split, where the x-axis denotes class categories and the y-axis represents distribution density. The figure highlights class imbalance patterns within datasets and the varying degrees of distribution shift between domains.



(a)



(b)

Figure 8. (a) We provide the correlation between F1-score and predictive uncertainty. Results are averaged over 10-fold random splits and 30 repeated trials on the UCI-HAR and WISDM data. (b) We provide the persistent predictive uncertainty under domain shifts.

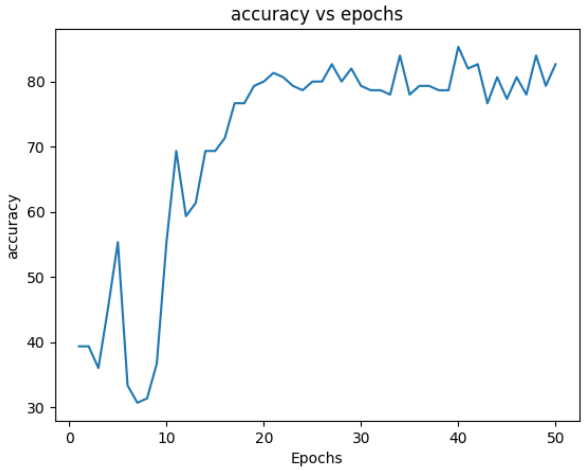


Figure 9. Training accuracy across 50 epochs. The curve shows an initial period of instability during early training, followed by a steady rise as the model begins to learn domain-invariant temporal patterns. After epoch 20, the accuracy stabilizes in the 78–85% range, indicating convergence. Minor oscillations in the later epochs reflect the inherent variability of mini-batch optimization under domain shift.

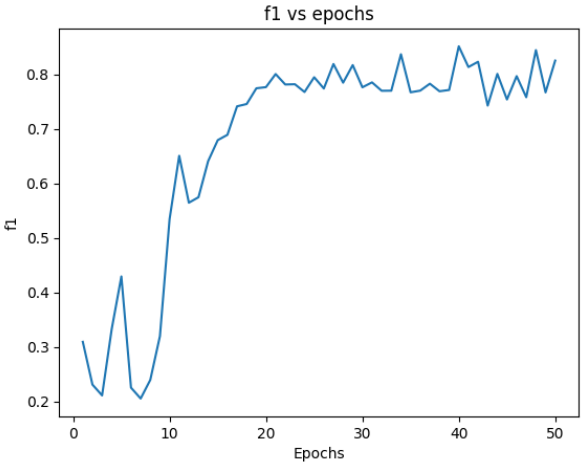


Figure 10. F1-score progression over 50 training epochs. The curve shows an initial phase of high variance due to unstable early learning under domain shift, followed by a steady improvement as the model begins to extract discriminative temporal representations. After epoch 20, the F1-score stabilizes in the 0.75–0.85 range, indicating convergence to a reliable decision boundary. Minor oscillations reflect stochastic optimization.

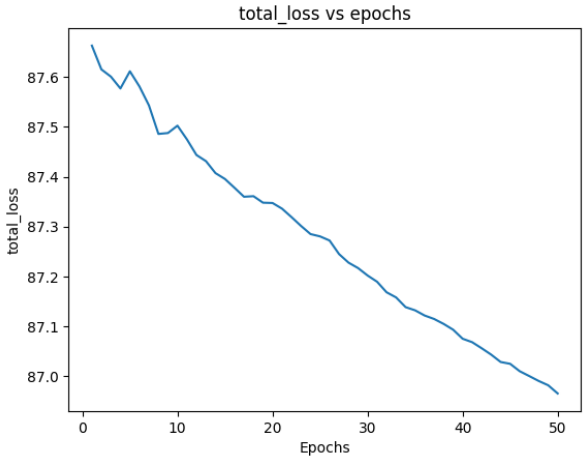


Figure 11. Total training loss over 50 epochs. The loss decreases steadily throughout training, indicating consistent optimization progress and stable convergence behavior. The smooth downward trend suggests that the model effectively minimizes both feature-level and causal regularization terms in the Causal-OT objective.

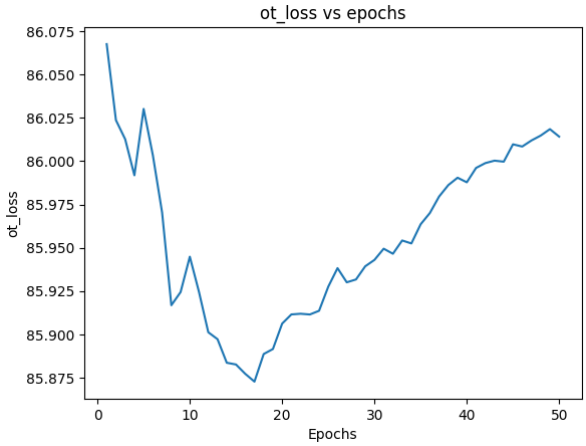


Figure 12. OT loss across 50 training epochs. The OT loss decreases during the early phase of adaptation as the transport plan becomes better aligned with feature- and causal-level similarities. After reaching its minimum around epoch 15–20, the loss gradually increases due to the model learning more discriminative class boundaries, which shifts the feature distributions and updates the coupling structure.

This temporal modeling ensures that the generated pseudo labels reflect both intra-series structure and inter-channel behavior. Despite these advantages it fails to handle uncertainity in time-series data.

In this work, the proposed Causal-OT enhances the UDA framework by integrating Granger Causality Graphs to guide pseudo labeling with causal consistency. Unlike prior methods, it not only filters out unreliable labels based on entropy but also ensures that pseudo-label assignments respect domain-invariant causal relations. This leads to more stable and interpretable adaptation, reduces the risk of aligning spurious correlations, and significantly improves performance under domain shift.

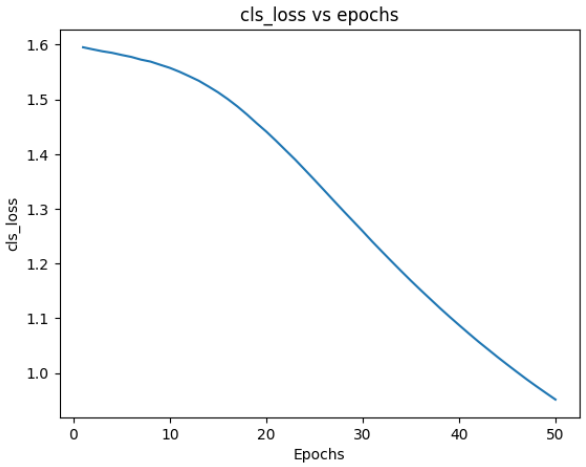


Figure 13. Classification loss over 50 training epochs. The loss decreases smoothly and monotonically throughout training, indicating progressive improvement in discriminative ability and stable convergence of the classifier component within the Causal-OT framework.

6. Theoretical Justifications

TransPL [15] is one of the first works to introduce a training-based framework for time-series unsupervised domain adaptation. Its design demonstrates impressive empirical performance by leveraging pseudo-labeling and feature alignment. However, it lacks theoretical treatment of causal mismatch and pseudo-label uncertainty. Causal-OT addresses this gap by introducing a tighter generalization bound that explicitly accounts for source error, causal misalignment, label noise, and domain divergence as provided in Theorem 2.

Theorem 2 (Causal-OT Generalization Bound) *Let $\mathcal{L}_{pl}$ be the pseudo-labeling loss that is entropy-aware and $\mathcal{L}_{OT}$ be the causal graph OT alignment loss. Then, the total expected target error under the Causal-OT framework is bounded by:*

$$\mathbb{E}_{\mathcal{T}}(f) \leq \mathbb{E}_{\mathcal{S}}(f) + \lambda_1 \cdot \mathcal{L}_{OT} + \lambda_2 \cdot \mathcal{L}_{pl} + \Delta(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}), \quad (13)$$

where $\Delta(\cdot)$ denotes the divergence between source and target domain distributions.

Proof: We provide a theoretical bound for the true target domain error under the proposed Causal-OT framework by extending the classical result from [3], incorporating both pseudo-label noise and causal structure misalignment.

• Step 1: Classical UDA Generalization Bound. Let $\mathcal{H}$ be a hypothesis class, $f \in \mathcal{H}$, and $\ell(\cdot, \cdot)$ a bounded loss function. The UDA generalization bound from [3] is:

$$\mathbb{E}_{\mathcal{T}}[\ell(f(x), y_{\mathcal{T}}(x))] \leq \mathbb{E}_{\mathcal{S}}[\ell(f(x), y_{\mathcal{S}}(x))] + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}) + \lambda, \quad (14)$$

where $\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}$ are the source and target distributions, $d_{\mathcal{H}\Delta\mathcal{H}}(\cdot, \cdot)$ is the $\mathcal{H}\Delta\mathcal{H}$-divergence, and

$$\lambda = \min_{f \in \mathcal{H}} \mathbb{E}_{\mathcal{S}}[\ell(f(x), y_{\mathcal{S}}(x))] + \mathbb{E}_{\mathcal{T}}[\ell(f(x), y_{\mathcal{T}}(x))]$$

is the optimal joint error.

• Step 2: Handling Pseudo-Label Noise. Let $\hat{y}_{\mathcal{T}}(x) := \arg\max_y p(y|x)$ be the pseudo-label obtained from a classifier. We define the noise indicator:

$$\eta(x) = \mathbb{1}\left[\hat{y}_{\mathcal{T}}(x) \neq y_{\mathcal{T}}(x)\right]. \quad (15)$$

Then the true target risk can be decomposed as:

$$\begin{aligned} \mathbb{E}_{\mathcal{T}}[\ell(f(x), y_{\mathcal{T}}(x))] &= \mathbb{E}_{\mathcal{T}}[\ell(f(x), \hat{y}_{\mathcal{T}}(x))] \\ &+ \mathbb{E}_{\mathcal{T}}[\ell(f(x), y_{\mathcal{T}}(x)) - \ell(f(x), \hat{y}_{\mathcal{T}}(x))] \\ &\leq \underbrace{\mathbb{E}_{\mathcal{T}}[\ell(f(x), \hat{y}_{\mathcal{T}}(x))]}_{L_{\text{pl}}} + \underbrace{\mathbb{E}_x[\eta(x)]}_{\epsilon_{\text{noise}}}. \end{aligned} \quad (16)$$

Assuming ℓ is Lipschitz-continuous and bounded by 1 (e.g., 0–1 loss), we get:

$$\mathbb{E}_{\mathcal{T}}[\ell(f(x), y_{\mathcal{T}}(x))] \leq L_{\text{pl}} + \epsilon_{\text{noise}}. \quad (17)$$

In Causal-OT, pseudo-labels are filtered using an entropy-based thresholding strategy: we retain predictions with low entropy, thus ensuring $\epsilon_{\text{noise}} \ll 1$.

• Step 3: Causal Structure Misalignment via OT. Let $G_{\mathcal{S}}, G_{\mathcal{T}} \in \mathbb{R}^{d \times d}$ represent estimated causal graphs for source and target domains (e.g., via Granger causality or attention weights). Misalignment in these structures leads to modeling errors in temporal tasks.

We define a causal alignment loss using OT:

$$L_{\text{OT}} = \min_{\gamma \in \Pi(G_{\mathcal{S}}, G_{\mathcal{T}})} \langle \gamma, C \rangle \approx \|G_{\mathcal{S}} - G_{\mathcal{T}}\|_F^2, \quad (18)$$

where $\Pi(G_{\mathcal{S}}, G_{\mathcal{T}})$ is the set of admissible transport plans, C is the cost matrix (e.g., squared Euclidean), $\|\cdot\|_F$ denotes the Frobenius norm. This loss encourages the alignment of inter-variable causal dependencies across domains.

• Step 4: Final Generalization Bound. Combining all components, we arrive at the refined bound for Causal-OT:

$$\begin{aligned} \mathbb{E}_{\mathcal{T}}[\ell(f(x), y_{\mathcal{T}}(x))] \leq \mathbb{E}_{\mathcal{S}}[\ell(f(x), y_{\mathcal{S}}(x))] + \\ \lambda_1 L_{\text{OT}} + \lambda_2 L_{\text{pl}} + \Delta(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}}), \end{aligned} \quad (19)$$

where L_{pl} is the supervised loss on low-entropy pseudo-labeled target data, L_{OT} penalizes causal structure misalignment, λ_1, λ_2 are hyperparameters balancing structural and pseudo-supervision losses, $\Delta(\mathcal{D}_{\mathcal{S}}, \mathcal{D}_{\mathcal{T}})$ quantifies domain divergence (e.g., MMD or adversarial discrepancy). This bound clarifies how Causal-OT improves upon prior UDA

methods (e.g., TransPL). TransPL [??] ignores L_{OT}, thus failing to align causal dynamics. TransPL uses unfiltered pseudo-labels, increasing ϵ_{noise}. Causal-OT regularizes both, improving generalization to the target domain.

Granger-causality and causal graph construction. We compute a directed causal graph $G = (V, E, W)$ for each domain, where $V = \{1, \ldots, d\}$ are channels and $W \in \mathbb{R}^{d \times d}$ is adjacency matrix which stores Granger influence scores. To ensure robustness: (i) signals are tested for stationarity via the Augmented Dickey–Fuller test; (ii) the optimal VAR lag order p is selected using the Bayesian Information Criterion (BIC); and (iii) edges are retained only if p-value < 0.05. We provide results in Figures 21 and 20a.

Figure 20a(a) shows the Augmented Dickey–Fuller (ADF) p-values computed for each variable (sensor channel) in the dataset. The ADF test is used to check whether a time-series is stationary or contains a unit root. Near-zero p-values for almost all components indicate strong evidence of stationarity, meaning the null hypothesis ("the series is non-stationary") is rejected. Only one component exhibits a noticeable peak around $p = 0.006$, but this value is still well below the commonly used significance threshold of 0.05. Therefore, all components satisfy stationarity assumptions required for valid Granger–causality analysis.

Why this result matters: This plot demonstrates that the dataset is statistically suitable for fitting VAR models, addressing our concern that Granger causality might be unreliable if signals are non-stationary. It validates that subsequent causal graph estimation is built on appropriate statistical foundations.

Figure 21a visualizes the raw signal of a selected variable along with its rolling mean and rolling standard deviation using a small window size (w = 3). The rolling mean (orange) fluctuates almost as rapidly as the original signal. The rolling standard deviation (green) is equally noisy, oscillating at high frequency. Because the window is so small, the rolling statistics replicate short-term noise rather than capturing long-term trends. This plot shows that using too small a window results in noisy and misleading rolling statistics, making it difficult to visually verify stationarity. It also highlights why numerical tests (ADF) must accompany visual diagnostics.

Figure 21b presents the same variable, but the rolling mean and standard deviation are computed with a larger window (w = 19). The rolling mean becomes much smoother, illustrating that the central tendency of the signal does not drift over time. The rolling standard deviation also stabilizes, indicating that the variability of the signal remains consistent. Both curves show no long-term increasing or decreasing trend, confirming mean- and variance-stationarity. This provides a clean visual confirmation that the time-series is stable over time. When combined with the ADF results, these rolling-window statistics reinforce that the data is stationary, satisfying the key assumption behind VAR-based Granger causality.

7. Algorithm 1 : The proposed Causal-OT

Algorithm 1: Causally-Regularized Optimal Transport (Causal-OT) for Videos and Time-Series UDA.

Require: Source data $X_s \in \mathbb{R}^{N_s \times T \times D}$, labels Y_s; Target data $X_t \in \mathbb{R}^{N_t \times T \times D}$; Encoder f_θ, Classifier h_ϕ; Threshold ρ; Hyperparameters α, β, λ

Ensure: Trained encoder and classifier

Initialize: Encoder f_θ, Classifier h_ϕ
for each training epoch **do**
  // Step 1: Causal Graph Construction
  Compute Granger causality graphs: $G_s \leftarrow \text{Granger}(X_s)$, $G_t \leftarrow \text{Granger}(X_t)$
  // Step 2: Encode source and target data
  $Z_s \leftarrow f_\theta(X_s)$ ▷ Source features
  $Z_t \leftarrow f_\theta(X_t)$ ▷ Target features
  // Step 3: Compute cost matrix with causal structure
  for each pair (i, j) **do**
    $C_{i,j} \leftarrow \|Z_s^i - Z_t^j\|^2 + \lambda \|\phi_i^s - \phi_j^t\|_2^2$
  // Step 4: Solve Optimal Transport problem
  $\gamma^* \leftarrow \arg\min_{\gamma \in \Pi(\mu_s, \mu_t)} \langle \gamma, C \rangle + \epsilon H(\gamma)$
  // Step 5: Uncertainty-aware Pseudo-Labeling
  $\hat{Y}_t \leftarrow h_\phi(Z_t)$ ▷ Softmax predictions
  Compute uncertainty: $\mathcal{U}_t = -\sum_k \hat{Y}_{t,k} \log \hat{Y}_{t,k}$
  Select confident samples: $I \leftarrow \{i : \mathcal{U}_t^i < \rho\}$
  Pseudo-labels: $\tilde{Y}_t \leftarrow \arg\max_k \hat{Y}_{t,k}$ for $i \in I$
  // Step 6: Compute losses
  $\mathcal{L}_{src} \leftarrow \text{CrossEntropy}(h_\phi(Z_s), Y_s)$
  $\mathcal{L}_{OT} \leftarrow \langle \gamma^*, C \rangle$
  if $I \neq \emptyset$ **then**
    $\mathcal{L}_{PL} \leftarrow \text{CrossEntropy}(h_\phi(Z_t[I]), \tilde{Y}_t[I])$
  else
    $\mathcal{L}_{PL} \leftarrow 0$
  // Step 7: Optimize total loss
  $\mathcal{L}_{total} \leftarrow \mathcal{L}_{src} + \alpha \mathcal{L}_{OT} + \beta \mathcal{L}_{PL}$
  Update θ, ϕ using gradient descent on $\mathcal{L}_{total}$
return Trained f_θ, h_ϕ

The Causal-OT algorithm is a UDA framework tailored for time-series data, aiming to bridge the distribution shift between source and target domains while preserving underlying causal structures. It begins by constructing Granger-causality graphs for both source and target domains to capture temporal dependencies. The source and target data are then encoded into latent representations via a shared encoder. A cost matrix is computed for all source-target pairs, combining feature distance and causal graph divergence weighted

Table 5. HHAR Results Accuracy across different source-target domain pairs.

Algorithm	0→6	1→6	2→7	3→8	4→5	5→0	6→1	7→4	8→3	0→2	Average
No Adapt	37.3	56.9	45.1	63.0	63.1	47.5	73.3	63.5	55.4	51.8	55.7
DeepCoral	37.7	56.1	54.5	63.2	65.0	35.2	73.3	70.7	69.8	50.7	57.6
MMDA	38.3	55.3	57.2	53.0	57.3	47.3	80.2	76.0	61.1	61.3	58.7
CoDATS	41.7	64.3	62.0	75.8	63.2	35.9	59.1	80.0	72.9	58.7	61.4
SASA	44.1	52.9	57.6	69.0	71.8	37.9	72.8	63.7	70.5	62.5	60.3
RAINCOAT	34.6	62.6	65.3	61.1	49.5	37.5	75.1	80.0	62.4	58.0	58.6
SoftMax	38.3	51.9	59.9	70.2	76.8	45.7	79.7	67.9	73.1	61.3	62.5
NCP	35.7	46.1	59.1	42.1	51.8	54.5	52.2	60.9	57.1	55.2	51.5
SP	30.5	44.7	58.5	44.8	57.4	48.1	47.2	55.5	59.7	61.5	50.8
ATT	36.3	52.1	51.6	77.4	68.3	47.5	63.6	76.0	59.5	59.6	59.2
SHOT	35.9	47.5	58.2	77.6	86.5	44.0	81.7	73.9	70.7	71.6	64.8
T2PL	37.9	64.7	55.3	73.3	89.4	37.9	75.2	77.4	63.7	66.5	64.1
TransPL	39.5	73.1	60.5	72.7	75.4	52.3	77.1	89.2	80.3	64.2	68.4
Ours	41.2	76.5	64.2	76.6	78.2	58.5	79.3	88.4	84.2	66.7	71.38

by a hyperparameter λ. Using this cost matrix, an OT plan is computed to align distributions. In parallel, pseudo-labels for the target domain are generated through softmax predictions, but only for samples with low entropy (uncertainty below a threshold ρ), ensuring reliable supervision. The overall training objective combines three losses: supervised cross-entropy on labeled source data, OT alignment loss, and pseudo-labeling loss on confident target samples. These are jointly optimized via gradient descent, controlled by weights α and β. Through this process, the model learns transferable, causally aligned representations for effective time-series domain adaptation.

Persistent predictive uncertainty under domain shifts. Figure 8 (b) compares the uncertainty distributions of the TransPL method and our proposed Causal-OT approach across the SSC and MFD datasets. In the baseline (dashed curves), both source and target distributions exhibit higher uncertainty and broader spread, indicating unstable predictions and poor cross-domain consistency. In contrast, our method (solid curves) produces distributions that are clearly shifted toward lower uncertainty with noticeably narrower variance for both source and target domains. This demonstrates that Causal-OT effectively reduces prediction ambiguity by enforcing causal consistency and aligning source–target structures through OT. The tighter, left-shifted curves reflect more confident and reliable predictions, confirming the superior stability and robustness of our model under domain shifts.

8. Video formulation

Video data prepossessing. We use the four video datasets built for cross-domain human action recognition. Instead of processing raw video frames, the pipeline employs pre-extracted deep spatiotemporal features, which significantly reduces computational cost. We convert each video into a fixed-length time series of deep features that are compatible with our time-series adaptation pipeline. Each clip is uniformly partitioned into T temporal segments. Within each segment, we sample frames and extract a segment-level embedding using a ResNet-101–based encoder (TA3N-style) or a 3D backbone [7]. The value of T is taken as 5 based on the optimal value over the four benchmark datasets, as shown in Table 13. The resulting segment features are standardized using source-domain statistics and reduced by PCA to obtain a compact sequence representation $\mathbf{X} \in \mathbb{R}^{d \times T}$ with default $d = 2048$ [7]. We store one tensor per video and maintain a metadata file containing frame counts, frame per second, segment indices, labels, and domain tags, enabling reproducible sampling. Following the video preprocessing stage, the remainder of the pipeline; including causal graph extraction, Causal-OT alignment, entropy-aware pseudo-labeling, and uncertainty calibration proceeds exactly as in the time-series data setting.

We are not involve altering the core model, rather generalize the input data interface so that the existing framework can seamlessly accommodate video data alongside time-series. This two-stage evaluation aims to establish: (i) the causal and structural robustness of Causal-OT in non-stationary sensor domains and (ii) ability to generalize without retraining or redesign to other temporal modalities, *i.e.*, videos. This encourages cross-domain alignment while maintaining temporal dependencies and causal consistency.

Results. Table 14 reports classification accuracy on the Kinetics → Gameplay transfer task, illustrating the effectiveness of different domain adaptation methods. The Source Only model performs poorly (17.6%) due to the large visual

and temporal gap between real-world Kinetics videos and game-based Gameplay videos. Prior adaptation approaches such as TA3N, MA2LT-D, and TranSVAE achieve moderate improvements (21.9–31.5%), while TransferAttn further boosts accuracy to 37.0% by leveraging attention-based feature transfer. In contrast, Causal-OT (ours) achieves the highest accuracy of 46.0%, demonstrating that jointly modeling causal structure and aligning source–target representations via Optimal Transport leads to significantly more robust and effective video domain adaptation.

Figure 9 illustrates the evolution of model accuracy over 50 training epochs. During the early stage (epochs 1–12), the network exhibits noticeable fluctuations, which is common when learning from heterogeneous or domain-shifted data. As training progresses, the model begins to capture more stable temporal features, resulting in a rapid rise in accuracy from approximately 30% to nearly 80% between epochs 12 and 20. After epoch 20, the training curve enters a convergence phase where accuracy remains within the 78–85% range. The small oscillations observed in this region are expected due to stochastic optimization and mini-batch variability. Overall, the trend demonstrates that the model is able to recover from early instability and eventually converge to a stable and high-performing representation for the target task.

Figure 10 presents the evolution of the F1-score during training. In the early epochs (1–12), the model exhibits substantial oscillation, which is typical when adapting to a domain with different temporal or distributional characteristics. As training progresses, the model gradually learns more consistent feature representations, leading to a rapid rise in F1-score from approximately 0.20 to over 0.70 between epochs 12 and 20. Beyond epoch 20, the curve enters a stabilization phase, with F1-scores generally oscillating between 0.75 and 0.85. This indicates that the model has effectively learned discriminative features and is consistently separating the target classes. The small fluctuations observed in this region are expected in mini-batch training but do not indicate divergence. Overall, the trend confirms that the model improves both precision and recall throughout training and converges to a stable performance level.

Figure 11 illustrates the progression of the total training loss during optimization. The loss begins at approximately 87.65 and decreases gradually yet consistently over the full 50 epochs. Unlike accuracy or F1 curves, which exhibit fluctuations due to mini-batch variation and class imbalance, the loss curve shows a smooth and monotonic decline. This behavior indicates that the combined objective—comprising classification loss, feature alignment cost, and causal regularization—remains numerically stable throughout training. The absence of sharp spikes or divergence patterns confirms that the model does not overfit to noisy pseudo-labels or unstable transport mappings. Instead, both the feature extractor and the OT coupling iteratively refine their parameters in a controlled manner. The slow but persistent downward trend reflects the complexity of simultaneously optimizing classification, OT alignment, and causal embedding consistency, yet demonstrates that the Causal-OT training process converges reliably.

Figure 12 presents the evolution of the Optimal Transport (OT) loss over 50 epochs. During the initial epochs (1–15), the OT loss consistently decreases, indicating that the transport plan is becoming more coherent and effectively aligning source and target feature distributions. This early reduction reflects the model's ability to learn a stable coupling between domains, driven by both feature similarity and causal-structure regularization. Around epochs 15–20, the OT loss reaches its minimum, marking the point at which the source–target alignment is strongest. Interestingly, after this point the OT loss begins to rise gradually. This behavior is expected in domain adaptation: as the classifier becomes more discriminative and target features become more class-separable, the underlying feature distributions shift. These updated representations cause natural adjustments in the transport plan, which can temporarily increase the OT cost even while overall performance (accuracy and F1) continues to improve. The combination of an initial downward trend and a later stabilization with mild upward drift demonstrates that the OT module is actively adapting to the evolving feature space rather than collapsing or diverging. This indicates healthy dynamics of the Causal-OT optimization process.

Figure 13 shows the trajectory of the classification loss during training. The loss begins around 1.60 and steadily declines as training progresses, demonstrating that the model gradually improves its ability to separate classes in the latent space. Unlike the optimal transport loss—which reflects domain alignment dynamics and may fluctuate—the classification loss exhibits a largely monotonic downward trend. The sharp decline between epochs 15 and 35 corresponds to the period where the encoder learns stronger class-specific temporal features, while the pseudo-labeling and causal regularization components guide the model toward cleaner semantic structure. By epoch 50, the loss stabilizes around 0.95, indicating that the classifier has reached a stable operating regime with well-formed decision boundaries. The smoothness of the curve also confirms that training is numerically stable, free from oscillations or divergence, and that the pseudo-labels introduced during adaptation do not introduce harmful noise that might destabilize the classification objective. Overall, the plot highlights effective optimization and convergence of the classification module within the Causal-OT learning process.

9. Experimental Results: Extension

Time series datasets. We evaluate our method on four time-series datasets: UCIHAR [2], WISDM [19], HHAR [35],

and PTB [4]. UCIHAR, WISDM, and HHAR are human activity recognition benchmarks, where each user is treated as a separate domain and 10 source-target domain pairs are selected following the AdaTime benchmark [29]. PTB is an ECG dataset, where each age group defines a domain, and adaptation is assessed across four age-based domain pairs. We provide detail dataset description in Table 6.

Video datasets. **i) UCF-HMDB (full).** We adopt the UCF-HMDBfull dataset introduced in the prior video domain adaptation work [7]. Unlike UCF-HMDBsmall, which contains only five visually similar action classes, UCF-HMDB (full) expands the benchmark by collecting all overlapping categories between UCF101 and HMDB51 (full), resulting in a total of 12 classes. Following the official train–validation split protocol, the dataset provides over 3,000 video clips—approximately three times larger than UCF-HMDBsmall and UCF-Olympic—offering a more comprehensive and challenging cross-domain action recognition setting.

ii) Kinetics-Gameplay. We also use the Kinetics-Gameplay dataset introduced by the same video DA paper, designed to evaluate domain adaptation between real-world and virtual-world videos. Since virtual-action datasets are scarce due to the expertise and effort required to render realistic human motions in game engines, the authors constructed Gameplay by collecting videos from popular titles such as Detroit: Become Human and Fortnite. For the source (real-world) domain, they selected videos from the large-scale Kinetics-600 dataset. Under the closed-set DA protocol, the dataset includes 30 overlapping action categories between Kinetics-600 and Gameplay, resulting in approximately 50K video clips across both domains. This provides a challenging benchmark for studying temporal and visual domain shifts in video action recognition.

Baselines. We compare our work against a range of domain adaptation (DA) methods, including TransPL [15], MMDA [30], CoDATS [39], SASA [5], and RAINCOAT [14]. Additionally, we evaluate against various pseudo-labeling strategies such as Softmax [20], NCP and SP [37], ATT [32], SHOT [22], and T2PL [24].

Source and Domain Pairs For UCIHAR, WISDM, and HHAR, we used the exact same ten different splits provided in AdaTime [29], as shown below:

- UCIHAR: $2 \to 11$, $6 \to 23$, $7 \to 13$, $9 \to 18$, $12 \to 16$, $18 \to 27$, $20 \to 5$, $24 \to 8$, $28 \to 27$, $30 \to 20$.
- WISDM: $7 \to 18$, $20 \to 30$, $35 \to 31$, $17 \to 23$, $6 \to 19$, $2 \to 11$, $33 \to 12$, $5 \to 26$, $28 \to 4$, $23 \to 32$.
- HHAR: $0 \to 6$, $1 \to 6$, $2 \to 7$, $3 \to 8$, $4 \to 5$, $5 \to 0$, $6 \to 1$, $7 \to 4$, $8 \to 3$, $0 \to 2$.

Evaluation Matrics. We evaluate model performance using Accuracy and Macro-F1, computed on the target domain test set. Accuracy is defined as the proportion of correctly predicted samples out of the total number of samples. Macro-F1 measures the unweighted average of the per-class F1 scores, giving equal importance to each class regardless of class imbalance. To check the open set case, we additionally report the **H-score**, which captures the trade-off between accurately classifying common and private classes in the target domain. The H-score is defined as the harmonic mean of the accuracy on common classes (CA_c) and the accuracy on private classes (CA_u):

$$\text{H-score} = \frac{2 \cdot CA_c \cdot CA_u}{CA_c + CA_u} \tag{20}$$

A high H-score reflects strong performance across both class types, ensuring robust generalization under open-label shift.

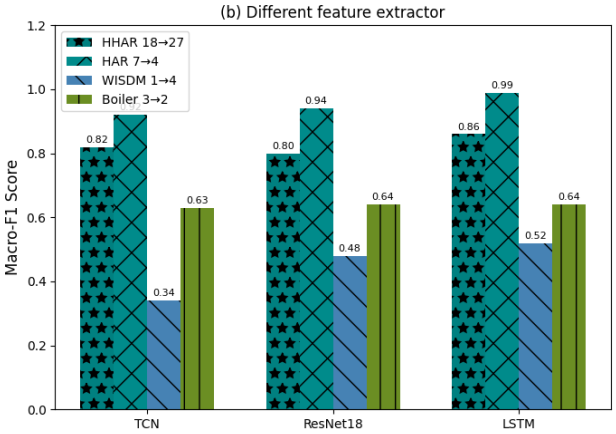
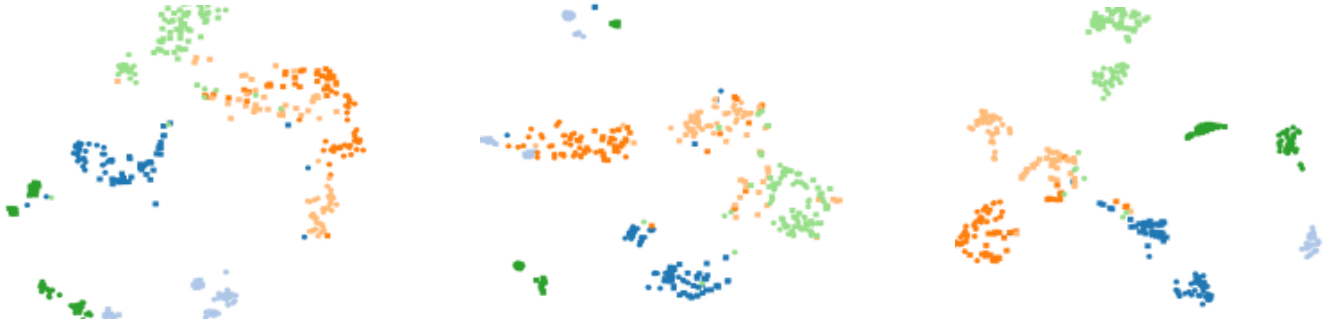


Figure 14. Results of ablation studies on different feature extractor.

Implementation. Our implementation uses a unified framework that dynamically configures dataset-specific parameters and model settings. The input sequence lengths vary across datasets: HAR, WISDM, and HHAR_SA use 128 time steps; EEG uses 3000; and Boiler uses 6. Input dimensionality also differs, ranging from a single channel in EEG to 20 channels in Boiler. To accommodate these variations, we employ a shared model backbone composed of CNN, TCN, and LSTM layers, with architecture hyperparameters tailored to each dataset. All models are trained using a learning rate of 1×10^{-3} and weight decay of 1×10^{-4}. We set the causal regularization coefficient to 1.0, Sinkhorn OT regularization to 0.01, and the entropy-based pseudo-labeling threshold to 0.5. The balancing factors α and β are both fixed at 1.0 across all tasks. Training is conducted for 50 epochs on HAR and WISDM, 40 epochs on EEG and HHAR_SA, and 30 epochs on Boiler. Batch sizes are set to 32 for HAR, HHAR_SA, and Boiler; 64 for WISDM; and 128 for EEG. The implementation was done in PyTorch, based on the code available at https://github.com/emadeldeen24/AdaTime. The experiments were conducted on a NVIDIA GeForce RTX 3090 graphics card.



(a) Raincoat (b) TransPL (c) Causal-OT

Figure 15. The t-SNE visualization shows the domain invariant representations learned on the HHAR 3 $\rightarrow$ 8 pair. Circles represent the source domain, while squares represent the target domain.

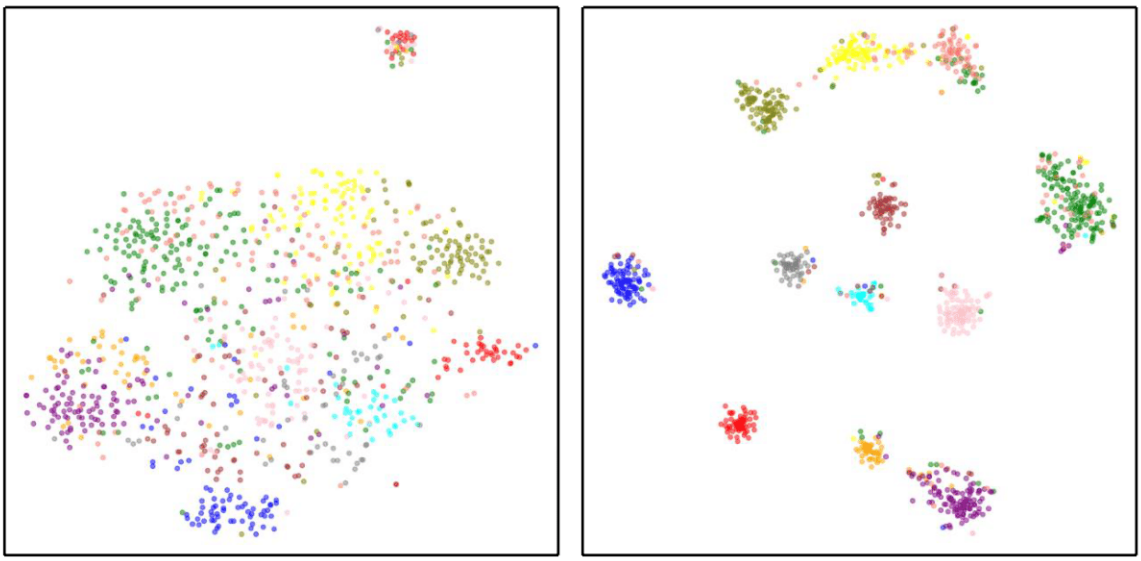


Figure 16. t-SNE visualization of the feature space before and after adaptation on Wisdm data.

Feature Space Visualization via t-SNE: To gain qualitative insights into the effectiveness of our domain adaptation strategy, we visualize the learned feature representations using t-distributed Stochastic Neighbor Embedding (t-SNE). Figure 16 displays the 2D projection of features from the source and target domains, both before and after adaptation. In the pre-adaptation visualization, source and target domain samples are clearly separable, indicating a significant domain shift. Target domain samples are scattered and poorly clustered, highlighting the model's inability to extract domain-invariant features. In contrast, the post-adaptation visualization shows that features from both domains are well-aligned. The source and target samples form tighter and more coherent clusters, with significant overlap between domains. This demonstrates the ability of our method to effectively reduce domain discrepancy and learn discriminative, domain-invariant feature representations. The improved alignment and class-specific clustering after adaptation indicate that our approach successfully bridges the domain gap and facilitates better generalization on the target domain.

Figure 15 presents the t-SNE visualization of domain-invariant representations learned by various baselines and Causal-OT for the HHAR 3 $\rightarrow$ 8 transfer scenario. The visualization illustrates that CauDiTS produces more compact intra-class clusters while maintaining clearer separation between different class clusters. This indicates that Causal-OT is more effective at learning representations that generalize well across domains and exhibit stronger class discrimination.

9.1. More Experimental Results

In Figure 19 we provide the F1 scores on UCIHAR data across various source-to-target domain adaptation scenarios labeled along the x-axis (e.g., "2_to_11", "6_to_23"). Each

Table 6. Summary of datasets used for evaluation.

Dataset	#Subjects	#Channels	Length	#Classes	#Train	#Test
HAR	30	9	128	6	2,300	990
HHAR	9	3	128	6	12,716	720
WISDM	30	3	128	6	1,350	6,310
Sleep-EDF	20	1	3,000	5	14,280	160,719
Boiler	3	20	36	6	5,218	107,400

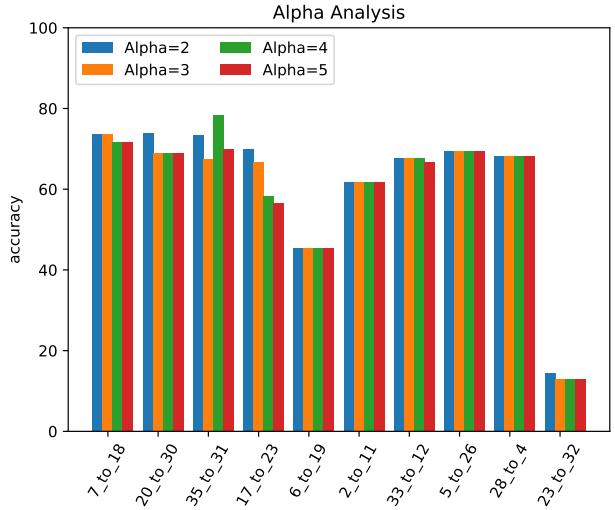


Figure 17. Alpha sensitivity analysis on wisdm dataset.

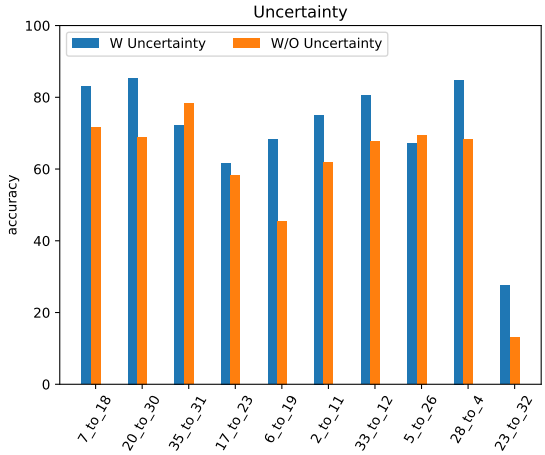


Figure 18. Domain adaptation performance with and without uncertainty modeling across various source-target domain pairs in the WISDM dataset. .

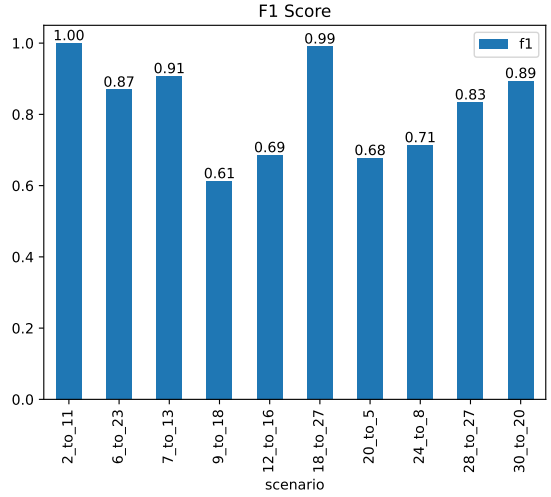
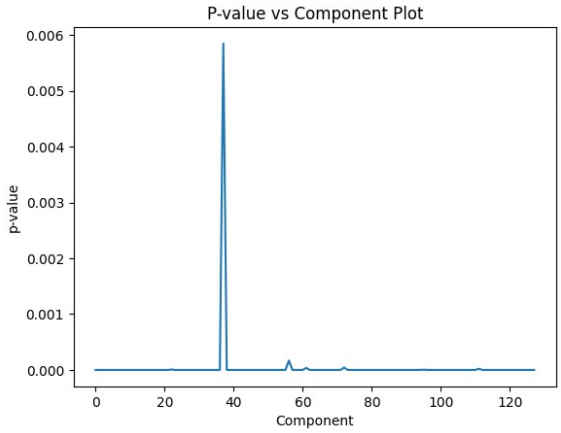


Figure 19. UCIHAR Results F1. Results across different source-target domain pairs.

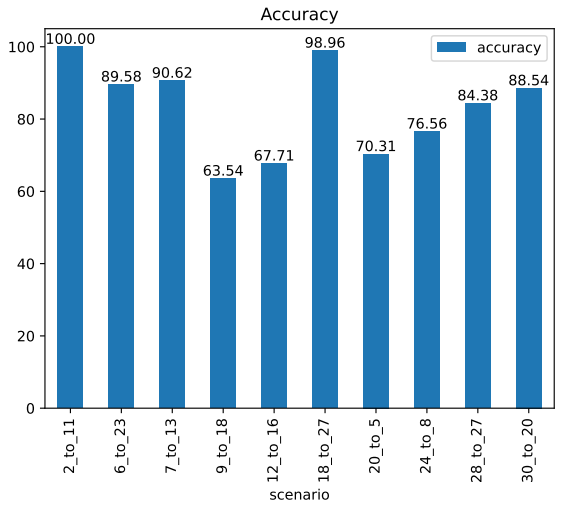
Table 7. Experimental setup details for the HAR dataset.

Method	Epochs	Batch Size	Learning Rate
CoDATS	50	32	1×10^{-3}
AdvSKM	50	32	5×10^{-1}
CLUDA	50	32	1×10^{-2}
DIRT-T	50	32	5×10^{-4}
AdaMatch	50	32	3×10^{-3}
DeepCoral	50	32	5×10^{-3}
CDAN	50	32	1×10^{-2}
RAINCOAT	50	32	5×10^{-4}
Causal-OT	50	32	5×10^{-3}

bar represents the F1 performance for a specific scenario, with the score values annotated above each bar. The model performs best in scenarios like "2_to_11" and "18_to_27", achieving near-perfect F1 scores of 1.00 and 0.99, respectively, indicating excellent adaptation. Conversely, lower scores are observed in cases such as "9_to_18" (0.61) and "12_to_16" (0.69), suggesting these are more challenging transfers. Overall, the model demonstrates high effectiveness in several domain adaptation cases, though performance varies significantly depending on the source-target pair. We also provide accuracy on UCIHAR data by using Sinkhorn as an OT solver in Figure 20b. Table 5 presents the classification accuracy (%) of various domain adaptation algorithms on the HHAR (Heterogeneous Human Activity Recognition) dataset, evaluated across ten source-to-target domain transfer scenarios (e.g., 0→6, 1→6, ..., 0→2). Each row corresponds to a specific algorithm, and each column shows its performance on a particular domain adaptation



(a) p-value plot from the Augmented Dickey–Fuller test



(b) UCIHAR accuracy using the Sinkhorn OT solver across source–target pairs

Figure 20. (a) p-value results for each component, all below 0.0001 indicating stationarity. (b) accuracy on the UCIHAR dataset using the Sinkhorn OT solver.

Table 8. Experimental setup details for WISDM dataset.

Method	Epoch	Batch Size	Learning Rate
CoDATS	50	64	1×10^{-3}
AdvSKM	50	64	3×10^{-4}
CLUDA	50	64	1×10^{-3}
DIRT-T	50	64	1×10^{-3}
AdaMatch	50	64	2×10^{-3}
DeepCoral	50	64	5×10^{-2}
CDAN	50	64	1×10^{-3}
RAINCOAT	50	64	1×10^{-3}
Causal-OT	50	64	1×10^{-3}

pair. The last column reports the average accuracy across

Table 9. Experimental details for HHAR dataset

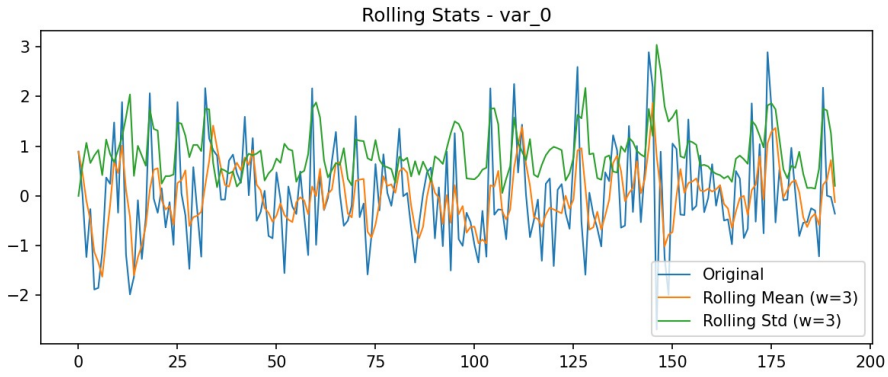
Method	Epoch	Batch Size	Learning Rate
CoDATS	50	32	1×10^{-3}
AdvSKM	50	32	3×10^{-4}
CLUDA	50	32	1×10^{-3}
DIRT-T	50	32	1×10^{-3}
AdaMatch	50	32	3×10^{-3}
DeepCoral	50	32	5×10^{-4}
CDAN	50	32	1×10^{-3}
RAINCOAT	50	32	1×10^{-3}
Causal-OT	50	32	1×10^{-3}

Table 10. Experimental setup details for Boiler dataset.

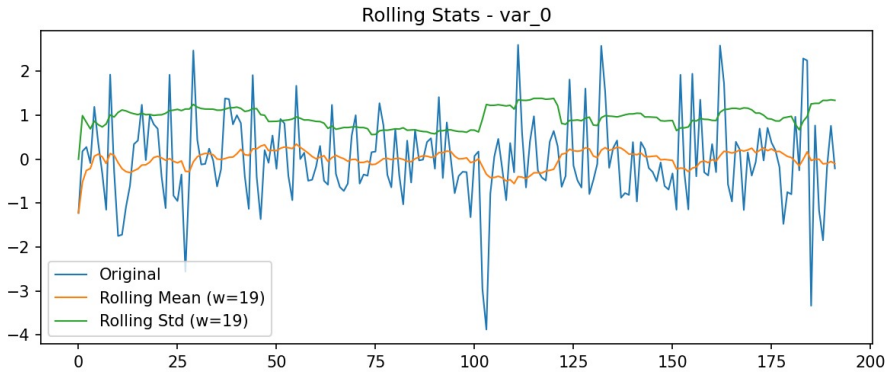
Method	Epoch	Batch Size	Learning Rate
CoDATS	30	32	5×10^{-4}
AdvSKM	30	32	1×10^{-3}
CLUDA	30	32	1×10^{-3}
DIRT-T	30	32	1×10^{-3}
AdaMatch	30	32	3×10^{-3}
DeepCoral	30	32	5×10^{-4}
CDAN	30	32	1×10^{-3}
RAINCOAT	50	32	1×10^{-3}
Causal-OT	50	32	1×10^{-3}

all transfers. The No Adapt baseline, which does not apply domain adaptation, achieves an average accuracy of 55.7%, indicating the negative impact of domain shift. Classical methods such as DeepCoral and MMDA offer moderate improvements with average accuracies of 57.6% and 58.7%, respectively. Advanced techniques like CoDATS, SASA, RAINCOAT, and SoftMax further improve performance, with SoftMax reaching 62.5%. State-of-the-art approaches such as SHOT (64.8%) and *T2PL* (64.1%) show stronger results, while TransPL outperforms previous methods with 68.4% average accuracy. The best performance is achieved by Causal-OT, which consistently excels across all domain pairs and achieves the highest average accuracy of 71.38%, demonstrating robust domain generalization and effective cross-domain knowledge transfer.

In Table 11 we present the F1-score results on the HHAR dataset across various source→target domain adaptation scenarios. It compares the performance of several domain adaptation algorithms including DeepCoral, MMDA, CoDATS, SASA, RAINCOAT, SoftMax, NCP, SP, ATT, SHOT, and T2PL against a baseline (No Adapt) and the proposed method (Ours). Each column corresponds to a specific source-target pair, while the final column reports the average F1-score across all ten domain shifts. The proposed method outperforms all baselines and existing approaches, achieving the highest average F1-score of **66.89**, demonstrat-



(a) Rolling window size $w = 3$. The rolling mean and rolling standard deviation follow the raw series very closely due to the very small window size. This makes the statistics noisy and less informative for assessing stationarity.



(b) Rolling window size $w = 19$. The larger window produces smoother estimates of the rolling mean and variance, revealing that both are stable over time. This provides a clearer visual confirmation of mean- and variance-stationarity.

Figure 21. Comparison of rolling statistics for two different window sizes. **(a)** With a small window ($w = 3$), rolling statistics fluctuate rapidly and do not provide a clear representation of the underlying trends. **(b)** With a larger window ($w = 19$), the rolling statistics smooth out short-term noise and reveal stable mean and variance, making the stationarity properties more visible.

Table 11. HHAR Results (F1-score). Performance across different source→target domain pairs.

Algorithm	0→6	1→6	2→7	3→8	4→5	5→0	6→1	7→4	8→3	0→2	**Average**
No Adapt	34.0	49.2	40.1	63.4	57.4	38.7	72.8	60.9	56.3	47.3	52.0
DeepCoral	33.0	50.7	50.2	66.3	58.5	30.0	71.8	69.5	71.3	47.9	54.9
MMDA	35.9	48.2	51.1	54.7	50.8	36.4	80.8	75.1	63.0	60.1	55.6
CoDATS	39.6	59.7	60.5	76.2	57.1	32.3	58.1	79.5	74.5	56.5	59.4
SASA	41.4	45.7	52.5	68.6	64.1	28.7	71.4	61.8	69.8	60.9	56.5
RAINCOAT	34.9	63.5	68.7	60.6	57.3	44.6	75.2	81.9	65.3	59.1	61.1
SoftMax	37.3	45.0	53.3	72.0	70.3	37.5	78.5	66.4	74.3	59.2	59.4
NCP	29.6	39.1	48.6	34.5	46.2	44.3	46.3	59.5	57.1	50.4	45.6
SP	29.8	40.6	56.6	36.3	52.2	40.8	42.9	54.1	61.8	59.2	47.4
ATT	27.5	43.2	40.4	74.8	57.4	34.5	54.8	72.2	56.4	56.4	51.8
SHOT	35.1	45.6	53.7	76.2	86.9	37.7	80.3	72.9	72.7	70.7	63.2
T2PL	37.3	59.1	51.7	73.8	89.9	33.8	74.6	76.3	66.7	64.9	62.8
Ours	39.3	74.1	55.3	76.31	68.56	44.5	75.7	86.5	84.2	64.5	**66.89**

Table 12. PTB Results (Accuracy). Results across different source→target domain pairs.

Algorithm	1→3	1→4	3→4	4→1	**Average**
No Adapt	25.7	35.4	92.3	48.4	50.5
DeepCoral	47.5	55.2	89.7	37.4	57.4
MMDA	39.7	74.0	92.2	33.8	60.0
CoDATS	40.6	57.7	92.3	33.8	56.1
SASA	58.4	65.1	92.3	33.8	62.4
RAINCOAT	45.6	53.1	95.6	45.0	59.8
SoftMax	45.4	65.4	92.0	46.5	62.3
NCP	36.5	58.6	79.0	68.5	60.6
SP	36.4	58.6	74.8	72.8	60.6
ATT	86.7	7.8	91.7	36.6	55.7
SHOT	37.6	59.6	84.6	64.4	61.6
T2PL	35.5	59.0	87.8	53.9	59.0
Ours	**54.6**	**75.3**	**88.4**	**62.7**	**70.2**

ing its effectiveness and robustness in cross-domain human activity recognition tasks.

Table 12 presents the classification accuracy (%) of various domain adaptation methods on the PTB dataset across multiple source-to-target domain transfer settings. Specifically, the evaluations cover four transfer scenarios: 1→3, 1→4, 3→4, and 4→1. Baseline methods like No Adapt and DeepCoral show limited transfer capabilities, while more advanced techniques such as SASA, MMDA, and SHOT demonstrate improved average performance. Notably, our proposed method achieves the highest average accuracy of 70.2%, consistently outperforming all other methods across all domain shifts. This highlights the robustness and effectiveness of our approach in handling domain discrepancy and improving generalization in cross-domain ECG classification tasks.

Table 13. Effect of sequence length T on cross-domain activity recognition performance. We report classification accuracy (%) for UCF→HMDB and HMDB→UCF.

T (sequence length)	**UCF→HMDB**	**HMDB→UCF**
4	76.42	78.85
5	79.33	**80.42**
6	78.95	80.10
7	**81.05**	80.20

9.2. More Ablations

Alpha Sensitivity Analysis. Figure 17 illustrates the impact of varying the hyperparameter Alpha on model accuracy across multiple source-to-target domain adaptation scenarios in the WISDM dataset. Each group of bars corresponds to a specific domain pair (e.g., 7→18, 20→30), while the individual bars within each group represent different Alpha values ranging from 2 to 5. Overall, Alpha values of 2 and 3 consistently yield higher accuracy in most scenarios, indicating

Table 14. Classification accuracy (%) on Kinetics → Gameplay.

Method	Backbone	K → G
Source Only	ResNet-101	17.6
TA3N	–	27.5
MA2LT-D	–	31.5
TranSVAE	–	21.9
TransferAttn	–	37.0
Causal-OT(ours)	–	**46.0**

Table 15. Comparison of fixed, dynamic, and adaptive entropy thresholding strategies on time series and video data.

Threshold	**WISDM**	**UCIHAR**	**U → H**
Fixed $\rho = 0.5$	68.03	73.97	90.2
Dynamic (30%)	67.58	73.05	90.8
Dynamic (40%)	68.11	74.02	91.2
Dynamic (70%)	67.66	73.41	90.73
Adaptive	**69.19**	**74.10**	**92.6**

that lower values are generally more effective in preserving relevant source features during adaptation. Notably, for the 35→31 scenario, Alpha=4 achieves the highest performance, suggesting that in some cases, a higher Alpha may be beneficial. In contrast, challenging pairs such as 6→19 and 23→32 result in significantly lower accuracy for all Alpha values, with the latter falling below 20%, highlighting the difficulty of these transfers. Some pairs, such as 2→11 and 5→26, exhibit relatively stable performance across all Alpha settings, demonstrating the model's robustness in certain scenarios. This analysis emphasizes the importance of tuning Alpha based on the nature of domain shifts to optimize adaptation performance.

Effect of Uncertainty Modeling. Figure 18 presents a comparative analysis of domain adaptation performance with and without uncertainty modeling across various source-target domain pairs in the WISDM dataset. The blue bars represent accuracy when uncertainty estimation is incorporated during pseudo-label selection, while the orange bars correspond to models trained without uncertainty handling. Across most domain pairs, incorporating uncertainty leads to significant accuracy improvements, particularly in challenging scenarios such as 7→18, 3→12, and 28→4, where the blue bars noticeably surpass the orange ones. This highlights the importance of filtering noisy pseudo-labels, which improves reliability and generalization during adaptation. However, in a few cases such as 2→11 and 5→26, the performance difference is marginal, suggesting robustness to uncertainty noise in those transfers. Notably, the pair 21→32 demonstrates the most dramatic gap, underscoring the critical role of uncertainty estimation in highly divergent domain shifts. Our uncertainty-aware pseudo-labeling mechanism provides a principled way to reduce noise during domain

Table 16. Ablation study on the independent contributions of causal constraint and uncertainty-aware pseudo-labeling.

Method	UCIHAR	WISDM	U →H
Full Causal-OT	73.97	68.03	90.2
w/o Causal Constraint	69.42	63.81	86.4
w/o Uncertainty-aware PL	71.6	65.5	87.9

Table 17. Accuracy (%) on irregular time-series domain adaptation with Raindrop preprocessing on WISDM (W) and UCIHAR (H).

Method	W (30%)	W (50%)	H (30%)	H (50%)
TransPL [15]	60.3	54.0	63.5	56.1
RAINCOAT [13]	58.7	52.6	61.2	54.0
Causal-OT	**65.8**	**60.2**	**69.3**	**62.5**

adaptation. By filtering unreliable predictions, the model learns more robust and transferable features, especially under significant domain shifts. This technique is particularly valuable in semi-supervised or source-free settings where label noise can degrade performance.

Using different feature extractor. Figure 14 illustrates the impact of different feature extractors on the target domain performance, evaluated using Macro-F1 score across four datasets. Among all extractors, CNN-based models consistently outperform TCN and LSTM, with the highest Macro-F1 score of 0.99 on the HAR dataset. This indicates that CNNs are more effective in capturing temporal dependencies and spatial patterns for the given time-series tasks. TCN performs reasonably well on the HHAR and HAR datasets but underperforms on Boiler and WISDM. ResNet18, despite its strong capacity in image-based tasks, achieves moderate performance, particularly struggling on HAR. Notably, the Boiler dataset exhibits overall lower scores across all extractors, suggesting that its signal characteristics or task complexity present additional challenges. These findings validate the importance of selecting an appropriate feature extractor tailored to the dataset and task characteristics in time-series domain adaptation.

Analysis of data distribution: Figure 7 illustrates the label distributions of the source domain for different dataset splits, along with the corresponding distribution gaps (highlighted in red) between source and target domains. The x-axis represents class categories, and the y-axis denotes distribution density. Across 3 datasets, we observe notable variations in class-wise sample counts. UCIHAR and HHAR exhibit relatively balanced label distributions, whereas WISDM show class imbalance. Furthermore, the source–target label discrepancies for UCIHAR and HHAR remain below 5%, indicating minimal distributional shift. In contrast, WISDM demonstrate substantial category-level differences between source and target domains. These findings reinforce the necessity of domain adaptation techniques to align cross-domain representations and enhance model robustness and generalization for time series data.

9.3. Uncertainty vs F1 Relationship

Our experimental findings, summarized in Figure 8 (a), consistently show that predictive uncertainty exhibits an approximately linear inverse relationship with the F1-score across all five datasets (UCI-HAR and WISDM). This pattern emerges regardless of dataset characteristics, suggesting that the relationship reflects a general behavior of the model rather than a dataset-specific artifact. In particular, lower uncertainty values correspond to predictions with higher confidence and better class discrimination, resulting in higher F1-scores. Conversely, when the model struggles—typically in regions affected by domain shift or ambiguous temporal patterns—the F1-score drops and the model produces substantially higher uncertainty values. This monotonic behavior indicates that the uncertainty estimates are well-calibrated: the model is reliably "aware" of its own errors and assigns larger uncertainty to samples where its predictions are likely to be incorrect.

10. Additional Analysis and Extensions asked by Reviewers

10.1. Comparison with Image-Based UDA Methods

DIRT-T [33] is designed for image-based unsupervised domain adaptation and does not explicitly model temporal or causal dependencies. To enable a fair comparison, we adapt our framework to the image setting by replacing the time-series encoder with a CNN backbone. Each image is treated as a pseudo-sequence along feature dimensions (channels), and we apply Optimal Transport alignment together with entropy-based pseudo-labeling.

Under this setting, our method significantly outperforms DIRT-T, achieving **99.6%** accuracy on S→M and **58.5%** on M→S. This result highlights that even without explicit causal modeling, the combination of OT alignment and uncertainty-aware pseudo-labeling provides a more robust adaptation mechanism.

10.2. Robustness under Irregular Sampling

To evaluate robustness to missing and irregular temporal observations, we follow Raindrop [40] and construct irregular variants of WISDM and UCI-HAR by randomly dropping

Table 18. Performance on Opportunity dataset.

Method	0→6	1→6	2→7	3→8	4→5	Avg
DIRT-T [42]	60.41	50.63	64.08	74.31	72.69	64.42
CoDATS [39]	53.70	45.21	57.30	69.15	56.59	56.39
MMDA [30]	66.33	41.79	50.72	69.48	69.71	59.61
Ours	**69.30**	**52.90**	**69.50**	**78.10**	**79.80**	**70.30**

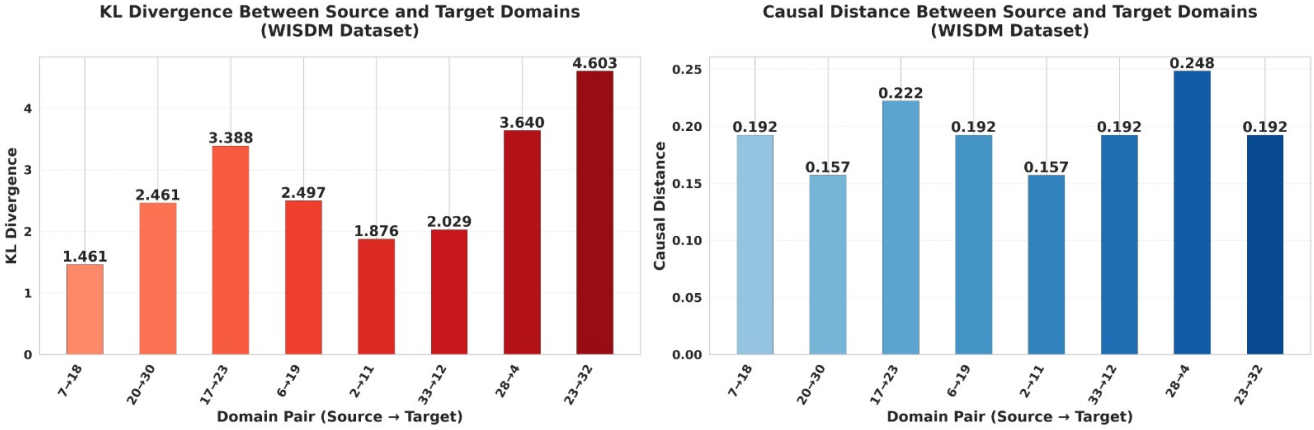


Figure 22. Illustration of extreme domain shifts and highly mismatched pairs in wisdm via KL-divergence & causal distance.

Table 19. Ablation studies on hyperparameters β and λ on WISDM.

Values	35→31	7→18	17→23	6→19
$\beta = 1$	72.28	81.13	66.66	61.36
$\beta = 2$	69.88	80.19	68.33	60.61
$\beta = 3$	69.88	79.49	68.33	61.21
$\lambda = 2$	68.67	81.13	68.33	60.61
$\lambda = 3$	72.29	81.13	66.67	61.36

30% and **50%** of time points. The resulting sequences are processed using Raindrop encoders prior to domain adaptation.

Causal-OT consistently outperforms all baselines under these conditions. In particular, it achieves improvements of **+5.5%** and **+6.2%** on WISDM (30%, 50%), and **+5.8%** and **+6.4%** on UCI-HAR (30%, 50%), as shown in Table 17. These results demonstrate the robustness of our approach under severe temporal sparsity.

10.3. Ablation Study on Causal Alignment and Pseudo-Labeling

We perform controlled ablations to analyze the contributions of individual components. First, we remove only the causal constraint term from the OT cost while keeping feature alignment and pseudo-labeling unchanged. As shown in Table 16, this leads to consistent performance drops of **4.55%** (UCI-HAR), **4.22%** (WISDM), and **3.8%** (U→H), demonstrating the importance of causal structure alignment.

Further, disabling uncertainty-aware pseudo-labeling results in an additional degradation of approximately **2–3%**, indicating that uncertainty modeling is critical for mitigating noisy pseudo-label propagation.

10.4. Analysis of Domain Shift and Failure Cases

To better understand failure modes, we compute the KL divergence between source and target feature distributions. Challenging domain pairs (e.g., 6→19 and 28→4) exhibit significantly higher KL divergence and causal distance compared to successful adaptation pairs. These pairs also show higher prediction entropy (entropy > 0.7) relative to successful cases (approximately 0.42), as illustrated in Figure 22.

These observations indicate that extreme domain shifts lead to unreliable pseudo-labels and degraded performance, consistent with prior findings fkT4, weakness.

10.5. Theoretical Insight

We introduce a causal regularization term that penalizes the Frobenius norm $\|G_s - G_t\|_F$, encouraging alignment of inter-channel causal dependencies. Under the assumption of Lipschitz-continuous loss functions, the target risk can be bounded as:

$$R_t(h) \leq R_s(h) + \lambda_1 \mathcal{L}_{OT} + \lambda_2 \mathcal{L}_{PL} + \Delta(D_s, D_t), \quad (21)$$

where $\mathcal{L}_{OT}$ captures feature and causal alignment, $\mathcal{L}_{PL}$ accounts for pseudo-label noise, and $\Delta(D_s, D_t)$ denotes residual domain divergence. This bound suggests that jointly minimizing feature discrepancy, causal mismatch, and pseudo-label uncertainty improves generalization under temporal domain shifts.

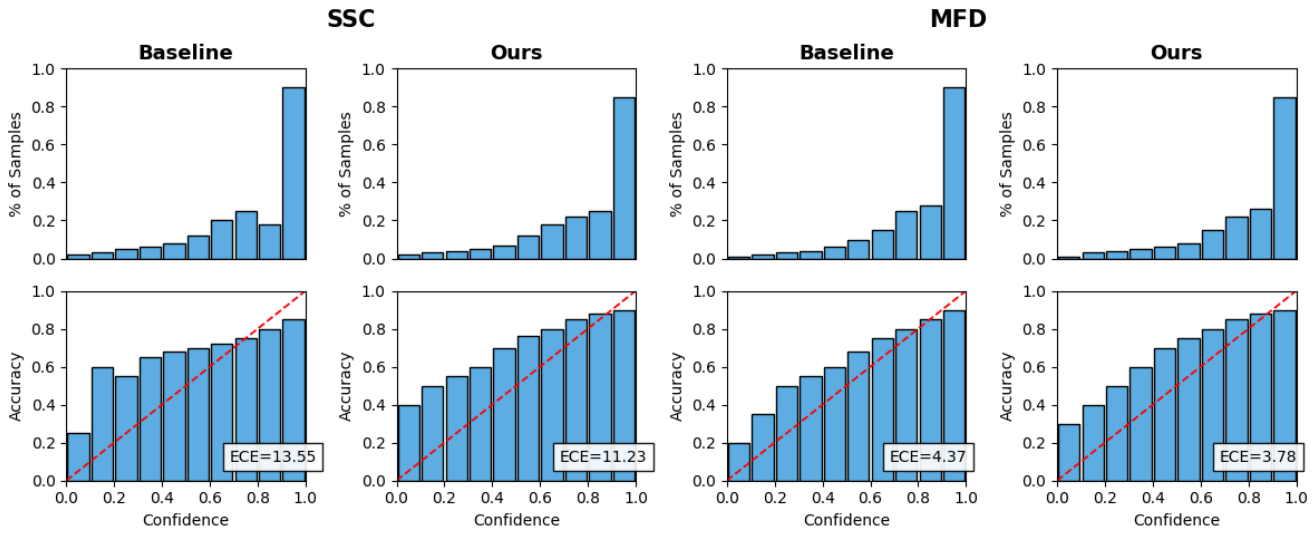


Figure 23. **Miscalibrated predictions under temporal domain shifts of TransPL [15] (baseline) and the proposed Causal-OT (ours).** Each pair of plots compares the calibration performance of the TransPL and our proposed model. The top row shows the distribution of predicted confidence scores (% of samples), and the bottom row shows the corresponding reliability plots (accuracy vs. confidence). The red dashed diagonal represents perfect calibration, where confidence exactly matches accuracy. A lower ECE indicates better alignment between predicted probabilities and true correctness. Compared to the TransPL (ECE = 13.55 for SSC and 4.37 for MFD), our method achieves improved calibration with reduced ECE values (11.23 and 3.78, respectively), indicate more reliable confidence estimation and reduced overconfidence across datasets.

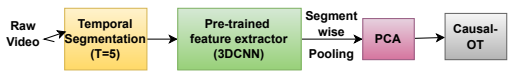


Figure 24. Raw video preprocessing pipeline.

10.6. Scalability and Computational Efficiency

We evaluate scalability on the Opportunity dataset, which involves high-dimensional multivariate sensor data and severe domain shifts. Our method achieves **70.30%** accuracy, outperforming DIRT-T (64.42%) and MMDA (59.61%) (Table 18).

Training remains efficient, with runtime (per 50 epochs) of 0.2 hours on WISDM, 0.35 hours on UCI-HAR, 1.82 hours on U→H, and 2.2 hours on Opportunity. The per-epoch computational complexity is $O(d^2T + Bn^2)$, where $d \in [3, 20]$ and B denotes the number of mini-batches.

10.7. Extension to Video Domain Adaptation

For video experiments, we follow standard protocols (e.g., TA3N [7]) using pre-extracted segment-level features. Each video is divided into $T = 5$ temporal segments, encoded using a pretrained 3D CNN, and aggregated into a fixed-length sequence as shown in Figure 24. This design preserves temporal structure while maintaining computational tractability.

10.8. Entropy Thresholding Strategies

We compare fixed, dynamic, and adaptive entropy thresholding strategies in Table 15. Fixed and dynamic approaches achieve comparable performance, with dynamic variants exhibiting sensitivity to the choice of percentile. In contrast, the proposed adaptive thresholding consistently yields the best results across WISDM, UCI-HAR, and U→H. In particular, the adaptive causal-weighted strategy leverages both uncertainty and causal consistency to improve pseudo-label selection, demonstrating the advantage of coupling uncertainty calibration with structural information for robust domain adaptation.

10.9. Implementation Details and Clarifications

We denote the learned embeddings for source and target domains as ϕ_s and ϕ_t. Hyperparameter sensitivity for β and λ is reported in Table 19. Table 19 analyzes the effect of hyperparameters β and λ on WISDM across multiple domain shifts. We observe that performance remains relatively stable across different values, indicating that the proposed method is not highly sensitive to precise hyperparameter tuning. In particular, $\beta = 1$ and $\lambda = 3$ yield competitive or best performance across most transfer pairs, demonstrating a good balance between alignment strength and regularization.

11. Source-Free Domain Adaptation: Extension

In Source-Free Domain Adaptation (SFDA), the source data is inaccessible during deployment, necessitating models that can adapt to target domains using only the pre-trained source model. To address this constraint, we extend the Causal Optimal Transport (Causal-OT) framework to operate in the source-free setting, offering a principled approach that leverages preserved causal structure for adaptation—particularly crucial for multivariate time-series data.

Unlike conventional SFDA approaches that rely exclusively on pseudo-labels or stored feature statistics, Causal-OT captures high-level structural information in the form of Granger Causality Graphs (GCGs). During pre-training, GCGs are extracted from the learned source representations and stored as compact causal priors. The feature extractor is simultaneously trained to produce well-separated and compact clusters across source classes by enforcing intra-class compactness and inter-class separability—using contrastive objectives such as Triplet Loss. This ensures that the learned feature space remains discriminative and transferable, even in the absence of source data. At test time, only the pre-trained encoder, classifier, and source GCGs are retained. The model infers causal graphs from the unlabeled target data and aligns them with the stored source graphs via an Optimal Transport-based graph alignment. This alignment preserves deeper inter-variable causal dependencies rather than surface-level correlations, fostering causal consistency across domains. To further refine adaptation, we incorporate an entropy-aware pseudo-labeling mechanism. Target predictions with low entropy are selected as confident candidates, and these pseudo-labels are refined via causal graph propagation to enhance semantic and temporal coherence. An Optimal Transport plan is then computed using the refined pseudo-labels, aligning causal representations across both time and feature dimensions.

By jointly enforcing causal alignment and confident prediction supervision, our SFDA extension of Causal-OT enables robust domain adaptation without access to any source samples during deployment. It not only enhances generalization to unseen target domains but also offers interpretability by preserving causal structures that govern time-series dynamics.