# DisentangledTMR: Privacy-Preserving Skeleton Motion Retargeting via Factorized Transformers

Thomas Carr[1,2], Depeng Xu[1], Shuhan Yuan[3], and Aidong Lu[1]

[1] University of North Carolina at Charlotte
[2] Incerta Intelligence
[3] Utah State University
thomas.carr@incertaintel.ai, {dxu7, aidong.lu}@charlotte.edu, shuhan.yuan@usu.edu

**Abstract.** Skeleton-based motion data leak personally identifiable information through both static skeletal structure and dynamic motion patterns, enabling re-identification even without facial features. We present DisentangledTMR, a Transformer Motion Retargeting (TMR) architecture that achieves privacy through explicit architectural disentanglement. Two encoders with complementary inductive biases, temporal convolutions for action and spatial graph convolutions for identity, feed a factorized decoder that fuses their representations through separate cross-attention streams and adaptive gating. A three-stage training curriculum progressively establishes disentanglement, reconstruction, and end-to-end refinement, and a tunable partial-retargeting ratio trades privacy for compatibility with pre-trained downstream models. On three benchmarks, DisentangledTMR substantially reduces re-identification while preserving action recognition, outperforming single-encoder baselines.

**Keywords:** Motion privacy · Skeleton anonymization · Motion retargeting · Disentangled representations · Factorized attention

## 1 Introduction

Skeleton-based motion data are widely used for action recognition, virtual reality, and human-computer interaction, yet they leak personally identifiable information (PII) despite omitting facial and appearance cues [5,6,25]. Static structural cues such as limb ratios and bone lengths serve as biometric signatures, while gait and individual execution styles enable identity inference from movement alone [18,27]. Even realistic full-body anonymization preserves gait-based identity [32], and in VR, unprotected body motion can undermine other privacy protections [4]. As motion-sensing technologies become ubiquitous in healthcare, smart homes, and immersive environments [6], robust anonymization is needed to balance analytical utility with individual privacy. Existing skeleton anonymization approaches like adversarial training [22,33,34], explanation-based perturbation [8], and differential privacy [28] all modify sequences directly, while motion retargeting transfers source motion onto target skeletons, explicitly replacing

structural identity while preserving action dynamics [7]. However, existing retargeting approaches use single-encoder architectures that create an inherent conflict between preserving action and removing identity.

We introduce *DisentangledTMR* (Transformer Motion Retargeting, TMR),[4] which advances privacy-centric retargeting through *architectural disentanglement*: separate encoders with complementary inductive biases (temporal convolutions for action, spatial GCN for identity) paired with a factorized decoder that fuses disentangled representations under privacy-aware objectives. Although the action encoder receives position alongside velocity and acceleration, its temporal convolution and attention primitives are poorly suited to extracting frame-invariant structure such as bone ratios, while spatial GCNs on a time-averaged pose cannot represent dynamic trajectories. This mismatch makes entanglement *structurally difficult* rather than relying solely on loss functions to discourage it. We show that this architectural separation provides greater privacy benefit than any individual disentanglement loss (Sec. 4.5), suggesting that inductive bias design is an underexplored axis for skeleton privacy. We use a three-stage training protocol for stable learning: Stage 1 pretrains the encoders with disentanglement losses to establish factor separation, Stage 2 trains the decoder for reconstruction and physical plausibility, and Stage 3 fine-tunes the full system end-to-end. We evaluate *DisentangledTMR* on three benchmark datasets, where the anonymized skeletons achieve lower re-identification (RI) accuracy while maintaining higher action recognition (AR) accuracy.

Our contributions include:

- a dual-encoder retargeting architecture with a factorized decoder and adaptive per-joint fusion gating;
- a disentanglement objective combining encoder-level losses with output-level adversarial supervision, suppressing identity in both the embeddings and the decoded output;
- a three-stage training procedure preventing reconstruction gradients from overriding disentanglement;
- a partial-retargeting ratio $\beta$ that exposes a tunable privacy–compatibility trade-off, enabling pre-trained downstream models to retain utility; and
- evaluation on three benchmarks with ablations and cross-dataset generalization.

## 2 Related Work

### 2.1 Motion Privacy and Skeleton Anonymization

Skeleton sequences reveal personal information through linkage attacks [5, 6], behavioral biometrics [25], gait signatures [27, 32], and attribute inference [18]. Direct anonymization approaches include adversarial training [22, 33, 34], deep motion masking in VR [23], explanation-based joint perturbation [8], differential

[4] Code and trained models are available in our public code repository.

privacy [28], and ciphertext-based GCN inference [11]. Motion retargeting offers a complementary strategy: PMR [7] and Moon *et al.* [22] transfer motion to target skeletons with adversarial learning, but their single-encoder architectures force a trade-off between action preservation and identity removal. Our dual-encoder design sidesteps this conflict through architectural disentanglement.

### 2.2 Disentangled Representations

Variational approaches ($\beta$-VAE [13], FactorVAE [16], $\beta$-TCVAE [9]) encourage disentanglement through modified ELBO objectives, though Locatello *et al.* [20] showed unsupervised disentanglement is impossible without inductive biases. We take a supervised approach: dedicated encoders with distinct inductive biases (temporal vs. spatial) combined with explicit action and identity supervision.

### 2.3 Motion Retargeting

Neural retargeting maps motions between different skeletons [1, 2, 31]. Recent methods such as R2ET [38] and MoMa [21] advance retargeting on synthetic animation data (*e.g.*, Mixamo), but target character animation rather than real-world sensor data and do not consider privacy. PMR [7] extends retargeting with adversarial learning for privacy, using a single shared CNN-based encoder paired with a CNN-based decoder. Our approach differs fundamentally: we replace the shared encoder with dedicated transformer-based encoders having complementary inductive biases (temporal convolutions for action, spatial GCN for identity) and introduce a factorized cross-attention decoder with adaptive fusion gating.

## 3 Methodology

*Threat model.* We define PII in skeleton data as any signal, whether static (*e.g.*, bone lengths) or dynamic (*e.g.*, gait), that enables re-identification or attribute inference (age, gender, race). We evaluate privacy against re-identification models with SGN [39] and Skeleton-MixFormer [35] backbones.
*Problem Statement.* Given a 3D skeleton sequence $\mathbf{s} \in \mathbb{R}^{J\times T\times 3}$ with $J$=25 joints over $T$=64 frames, **our goal** is to transfer the action from a source sequence $\mathbf{s}_{\text{src}}$ (person $P_1$, action $A_1$) onto a target skeleton $\mathbf{s}_{\text{tgt}}$ (person $P_2$) while suppressing identity-revealing signals.
*Core mechanism.* Disentanglement is enforced primarily by a bandwidth asymmetry between the encoders, not by losses alone: action is high-bandwidth and time-varying, identity low-bandwidth and time-invariant. The action encoder (temporal convolution and attention) captures dynamics but not frame-invariant bone ratios, while the identity encoder sees only a time-averaged pose through a narrow ∼256-dim bottleneck and cannot represent trajectories. Each encoder is thus biased toward the factor it keeps and away from the one it discards, making entanglement structurally difficult; the auxiliary losses remove residual leakage.

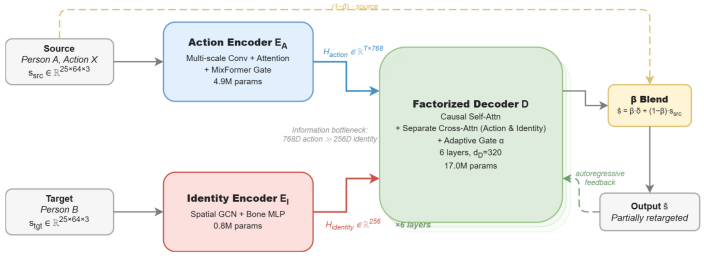


**Fig. 1:** DisentangledTMR overview, including the optional $\beta$-blend (Eq. (16)) of retargeted output with the source. Components: Figs. 2 to 4.

To achieve this goal, we propose a transformer-based motion retargeting model with dual encoders $E_A$ (action) and $E_I$ (identity) whose outputs are fused at a decoder $D$ to generate anonymized skeleton $\hat{\mathbf{s}} = D(E_A(\mathbf{s}_{\text{src}}), E_I(\mathbf{s}_{\text{tgt}}))$.

Encoders $E_A$ and $E_I$ produce representations $\mathbf{H}_{\text{action}} = E_A(\mathbf{s})$ and $\mathbf{H}_{\text{identity}} = E_I(\mathbf{s})$. We say the representations are disentangled if: (i) *exclusion*: mutual information $MI(\mathbf{H}_{\text{action}}; P) \approx 0$ and $MI(\mathbf{H}_{\text{identity}}; A) \approx 0$; (ii) *invariance*: $\mathbf{H}_{\text{action}}$ is invariant to structural changes, $\mathbf{H}_{\text{identity}}$ invariant to temporal re-timing; and (iii) *independence*: $\mathbf{H}_{\text{action}} \perp \mathbf{H}_{\text{identity}}$. We enforce these with orthogonality penalties, adversarial classifiers, contrastive clustering, and cross-correlation decorrelation.

### 3.1 Architecture

DisentangledTMR consists of three components (Fig. 1): (1) an **Action Encoder** ($E_A$) that extracts temporal motion dynamics, (2) an **Identity Encoder** ($E_I$) that extracts spatial skeletal structure, and (3) a **Factorized Decoder** ($D$) that combines these representations through adaptive fusion:

$$\mathbf{H}_{\text{action}} = E_A(\mathbf{s}_{\text{src}}), \quad \mathbf{H}_{\text{identity}} = E_I(\mathbf{s}_{\text{tgt}}), \quad \hat{\mathbf{s}} = D(\mathbf{H}_{\text{action}}, \mathbf{H}_{\text{identity}}), \quad (1)$$

where $\mathbf{H}_{\text{action}} \in \mathbb{R}^{T \times d_A}$ ($d_A$=768) captures motion dynamics and $\mathbf{H}_{\text{identity}} \in \mathbb{R}^{d_I}$ ($d_I$=256) is a single vector summarizing skeletal structure (broadcast to all $T$ time steps in the decoder). The asymmetric total capacity ($T \times d_A = 49{,}152$ vs. $d_I = 256$, a $\sim$192$\times$ ratio) implements an *information bottleneck* [29]: identity is structurally low-dimensional (bone lengths and joint offsets span only $\sim$24 scalar degrees of freedom for 25 joints), so a 256-dimensional code suffices for faithful skeletal reconstruction while limiting the bandwidth available to encode action-correlated identity signatures.

**Action Encoder** The action encoder emphasizes *how* joints move rather than *where* they are (Fig. 2).

*Input processing.* At each frame $t$, joint positions $\mathbf{s}_t \in \mathbb{R}^{J \times 3}$, velocity ($\mathbf{V}_t = \mathbf{s}_t - \mathbf{s}_{t-1}$), and acceleration ($\mathbf{A}_t = \mathbf{V}_t - \mathbf{V}_{t-1}$) are concatenated to form $\mathbf{X} \in \mathbb{R}^{J \times T \times 9}$, emphasizing temporal change over static position.

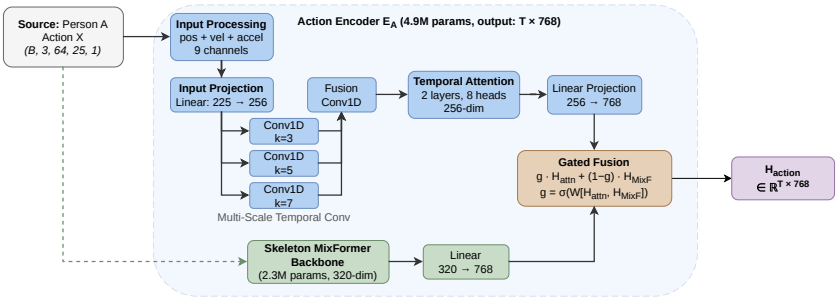


**Fig. 2:** Action encoder architecture. Two parallel pathways (multi-scale temporal convolutions + MixFormer backbone) are fused via a learned gate to produce $\mathbf{H}_{\text{action}} \in \mathbb{R}^{T\times 768}$.

*Multi-scale temporal convolutions.* Parallel 1D convolutions with kernels $k \in \{3,5,7\}$ capture motion at different temporal scales. Let $d_{\text{int}} = 256$ denote the intermediate feature dimension:

$$\mathbf{H}_{\text{conv}} = \mathbf{W}_{\text{proj}}\big[\text{ReLU}(\text{Conv1D}_k(\mathbf{X}))\big]_{k\in\{3,5,7\}} \in \mathbb{R}^{T\times d_{\text{int}}}, \tag{2}$$

*Temporal attention.* Two layers of 8-head self-attention [30] along the temporal dimension capture long-range dependencies:

$$\mathbf{H}_{\text{attn}} = \text{TemporalAttention}(\mathbf{H}_{\text{conv}}) \in \mathbb{R}^{T\times d_{\text{int}}}. \tag{3}$$

A final linear projection yields $\mathbf{H}'_{\text{attn}} = \mathbf{W}_{\text{up}}\mathbf{H}_{\text{attn}} \in \mathbb{R}^{T\times d_A}$ (with $d_A$=768).
*MixFormer gate.* A modified skeleton MixFormer [35] backbone produces per-frame features $\mathbf{H}_{\text{MixF}} \in \mathbb{R}^{T\times d_A}$ that are fused with $\mathbf{H}'_{\text{attn}}$ via an element-wise learned gate:

$$\mathbf{H}_{\text{action}} = \boldsymbol{g} \odot \mathbf{H}'_{\text{attn}} + (1-\boldsymbol{g}) \odot \mathbf{H}_{\text{MixF}}, \tag{4}$$

where $\boldsymbol{g} \in [0,1]^{T\times d_A}$ is produced by a linear layer with sigmoid over $[\mathbf{H}'_{\text{attn}}; \mathbf{H}_{\text{MixF}}]$. The MixFormer substream is randomly initialized (Kaiming normal [12]) and trained jointly from scratch; no pretrained weights are loaded.

**Identity Encoder** The identity encoder captures spatial skeletal structure while discarding temporal dynamics (Fig. 3).
*Static pose and bone lengths.* A static pose $\mathbf{s}_{\text{static}} = \frac{1}{T}\sum_t \mathbf{s}_t \in \mathbb{R}^{J\times 3}$ removes action-specific motion. Let $\mathcal{E}$ denote the set of 24 bone edges defined by the skeleton topology (connecting each non-root joint to its parent). Bone lengths $B_i = \|\mathbf{s}^{(j_a)}_{\text{static}} - \mathbf{s}^{(j_b)}_{\text{static}}\|_2$ for each edge $(j_a, j_b) \in \mathcal{E}$ provide explicit identity-specific features, collected into $\mathbf{B} = [B_1, \ldots, B_{|\mathcal{E}|}] \in \mathbb{R}^{24}$. Prior to feature extraction, skeletons are root-centered by subtracting the pelvis joint and de-rotated to a canonical orientation [39].

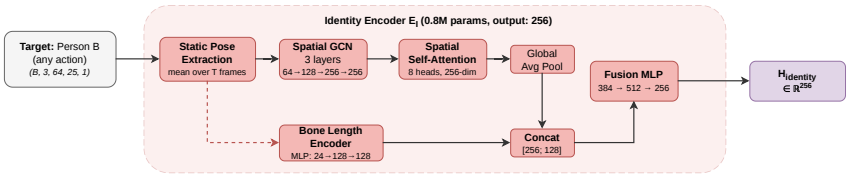


**Fig. 3:** Identity encoder architecture. A spatial GCN on the skeleton graph and a bone-length MLP are fused to produce $\mathbf{H}_{\text{identity}} \in \mathbb{R}^{256}$.

*Spatial GCN.* Following graph-based skeleton modeling [36], a pointwise input projection ($3 \to 64$ channels) followed by three spatial GCN layers ($64 \to 128 \to 256 \to 256$ channels, each with residual connections) operates on the skeleton adjacency graph to aggregate neighborhood structure:

$$\mathbf{H}_{\text{GCN}}^{(\ell)} = \text{ReLU}(\mathbf{A}\,\mathbf{H}_{\text{GCN}}^{(\ell-1)}\,\mathbf{W}^{(\ell)}), \quad \ell = 1, 2, 3, \tag{5}$$

where $\mathbf{H}_{\text{GCN}}^{(\ell-1)} \in \mathbb{R}^{J \times d_\ell}$ is the input node feature matrix at layer $\ell$, $\mathbf{W}^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell+1}}$ is a learnable weight matrix, and $\mathbf{A} \in \mathbb{R}^{J \times J}$ is the normalized adjacency matrix. Each layer includes batch normalization before the activation and a learnable residual connection when input and output dimensions differ. A multi-head (8-head) spatial self-attention layer then captures non-local joint relationships (*e.g.*, left–right hand coordination):

$$\mathbf{H}_{\text{spatial}} = \text{softmax}\left(\frac{\mathbf{H}_{\text{GCN}}^{(3)}\,\mathbf{W}_Q\,(\mathbf{H}_{\text{GCN}}^{(3)}\,\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{H}_{\text{GCN}}^{(3)}\,\mathbf{W}_V \;\in \mathbb{R}^{J \times 256}, \tag{6}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learned projections.
*Feature fusion.* Bone lengths are encoded via a 2-layer MLP ($24 \to 128 \to 128$) and concatenated with globally-pooled spatial features, then projected through a fusion MLP:

$$\mathbf{H}_{\text{identity}} = \text{MLP}\big([\text{GlobalAvgPool}(\mathbf{H}_{\text{spatial}});\ \text{MLP}_{\text{bone}}(\mathbf{B})]\big) \in \mathbb{R}^{d_I}. \tag{7}$$

**Factorized Decoder** The decoder generates motion autoregressively (Fig. 4). Let $\mathbf{Z}_D \in \mathbb{R}^{T \times d_D}$ denote the hidden state at a given layer. Each of the 6 decoder layers applies: (1) causal self-attention over $\mathbf{Z}_D$, (2) *separate* cross-attention to $\mathbf{H}_{\text{action}}$ and $\mathbf{H}_{\text{identity}}$, and (3) adaptive fusion via a learned gate:

$$\mathbf{Z}_A = \text{Attn}(\mathbf{Z}_D, \mathbf{H}_{\text{action}}, \mathbf{H}_{\text{action}}), \quad \mathbf{Z}_I = \text{Attn}(\mathbf{Z}_D, \mathbf{H}_{\text{identity}}, \mathbf{H}_{\text{identity}}), \tag{8}$$

$$\mathbf{Z} = \boldsymbol{\alpha} \odot \mathbf{Z}_A + (\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{Z}_I, \quad \boldsymbol{\alpha} = \sigma\big(\mathbf{W}_g\,[\mathbf{Z}_A; \mathbf{Z}_I]\big), \tag{9}$$

where Attn denotes self-attention, $[\mathbf{Z}_A; \mathbf{Z}_I] \in \mathbb{R}^{T \times 2d_D}$ denotes concatenation along the feature dimension, $\mathbf{W}_g \in \mathbb{R}^{2d_D \times d_D}$ is a learnable projection, and

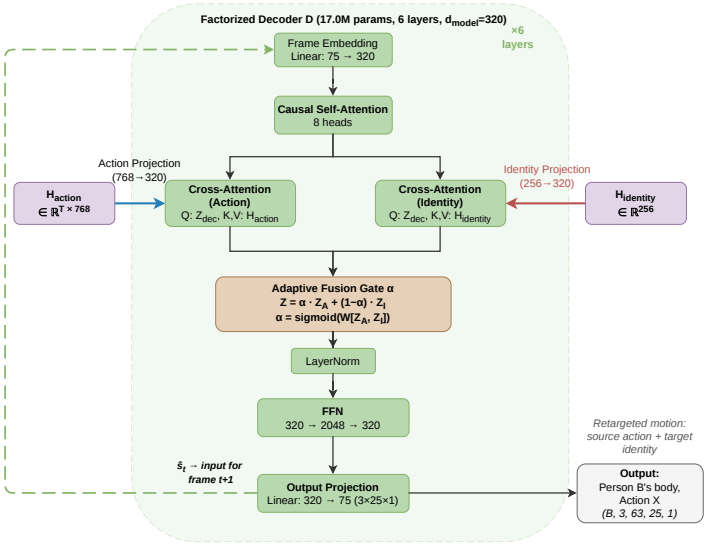


**Fig. 4:** Factorized decoder. Separate cross-attention to action and identity features are fused via an adaptive gate $\boldsymbol{\alpha}$; output feeds back autoregressively.

$\boldsymbol{\alpha} \in (0,1)^{T \times d_D}$ is a per-element gate vector. Since $\mathbf{H}_{\text{identity}}$ is a single global token, identity cross-attention effectively broadcasts the skeletal descriptor to all temporal positions, functioning as a learned conditioning signal analogous to FiLM [24]. This is intentional: identity is a global, time-invariant property and does not require per-frame key-value diversity.

### 3.2 Three-Stage Training

End-to-end training from random initialization risks reconstruction gradients dominating the disentanglement objectives, because the decoder can exploit entangled shortcuts before factor separation is established. The three-stage protocol addresses this by first enforcing disentanglement, then introducing reconstruction pressure under already-separated representations.

**Stage 1: Encoder Pretraining** The decoder is frozen. Encoders and auxiliary heads (AR classifier, RI classifier, identity discriminator) are trained with six losses:

$$\mathcal{L}_{\text{S1}} = \lambda_{\text{AR}}\mathcal{L}_{\text{AR}} + \lambda_{\text{RI}}\mathcal{L}_{\text{RI}} + \lambda_{\text{ctr}}\mathcal{L}_{\text{ctr}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{orth}}\mathcal{L}_{\text{orth}} + \lambda_{\text{CC}}\mathcal{L}_{\text{CC}}. \quad (10)$$

*AR and RI losses.* Standard cross-entropy: $\mathcal{L}_{\text{AR}}$ trains a 3-layer MLP to classify action from $\mathbf{H}_{\text{action}}$ (temporal-average pooled); $\mathcal{L}_{\text{RI}}$ classifies identity from $\mathbf{H}_{\text{identity}}$.

*Contrastive loss.* Clusters action features by action class across identities using supervised contrastive learning [15]:

$$\mathcal{L}_{\text{ctr}} = -\frac{1}{|\mathcal{P}|}\sum_{i=1}^{N}\sum_{j\in\mathcal{P}(i)}\log\frac{\exp(\mathbf{h}_i^\top\mathbf{h}_j/\tau)}{\sum_{k=1}^{N}\exp(\mathbf{h}_i^\top\mathbf{h}_k/\tau)}, \tag{11}$$

where $\mathbf{h}_i = \mathbf{H}^{(i)}_{\text{action}}/\|\mathbf{H}^{(i)}_{\text{action}}\|$, $\mathcal{P}(i)$ indexes positives (same action, different identity), and $\tau$=0.07 is a fixed temperature hyperparameter.
*Adversarial loss.* A discriminator $D_{\text{adv}}$ is trained with standard cross-entropy to predict identity from action features stopgradient($\mathbf{H}_{\text{action}}$), where stopgradient denotes *stop-gradient* (gradients are not back-propagated into $E_A$ during the discriminator update). The encoder is then trained to maximize the discriminator's entropy by minimizing the KL divergence between the discriminator's output distribution and a uniform distribution over identities:

$$\mathcal{L}_{\text{adv}} = \text{KL}(D_{\text{adv}}(\cdot \mid \mathbf{H}_{\text{action}}) \parallel \mathcal{U}(1, N_{\text{id}}))\,, \tag{12}$$

where $D_{\text{adv}}(\cdot \mid \mathbf{H}_{\text{action}})$ denotes the discriminator's predicted probability distribution over all $N_{\text{id}}$ identities given action features $\mathbf{H}_{\text{action}}$, and $\mathcal{U}(1, N_{\text{id}})$ is the discrete uniform distribution over identities. The stop-gradient decouples the two objectives: the discriminator learns to detect identity from a frozen snapshot of $\mathbf{H}_{\text{action}}$, while $E_A$ receives gradients only from $\mathcal{L}_{\text{adv}}$, which drives it to erase identity. This is analogous to domain-adversarial training [10] but implemented via separate optimization steps rather than a gradient-reversal layer.
*Orthogonality loss.* Penalizes linear correlation between factor representations:

$$\mathcal{L}_{\text{orth}} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\mathbf{h}}_a^{(i)}\cdot\hat{\mathbf{h}}_{id}^{(i)}\right|, \tag{13}$$

where $\hat{\mathbf{h}}_a^{(i)}$ and $\hat{\mathbf{h}}_{id}^{(i)}$ are L2-normalized projections of $\mathbf{H}^{(i)}_{\text{action}}$ and $\mathbf{H}^{(i)}_{\text{identity}}$.
*Cross-correlation loss* ($\mathcal{L}_{\text{CC}}$, inspired by Barlow Twins). Minimizes the Frobenius norm of the cross-correlation matrix, a decorrelation objective inspired by Barlow Twins [37] that encourages statistical independence:

$$\mathcal{L}_{\text{CC}} = \frac{1}{d^2}\left\|\frac{1}{N}\tilde{\mathbf{H}}_a^\top\tilde{\mathbf{H}}_{id}\right\|_F^2, \quad d = \min(d_A, d_I), \tag{14}$$

where $\tilde{\mathbf{H}}_a$ and $\tilde{\mathbf{H}}_{id}$ are zero-mean unit-variance standardized projections of $\mathbf{H}_{\text{action}}$ and $\mathbf{H}_{\text{identity}}$. Since $d_A \neq d_I$, both losses operate on representations truncated to a common dimensionality $d$=$\min(d_A, d_I)$=256. For $\mathcal{L}_{\text{orth}}$, action features are temporally averaged before truncation; for $\mathcal{L}_{\text{CC}}$, the temporal dimension is flattened and the first $d$ elements are retained.

**Stage 2: Decoder Training** Encoders are trained at a reduced learning rate (1% of the base rate) while the decoder trains at the full learning rate. The

training objective combines reconstruction and physical plausibility losses:

$$\begin{aligned}\mathcal{L}_{\text{S2}} = \lambda_{\text{MSE}}\mathcal{L}_{\text{MSE}} + \lambda_{\text{bone}}\mathcal{L}_{\text{bone}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} \\ + \lambda_{\text{EE}}\mathcal{L}_{\text{EE}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}}.\end{aligned} \tag{15}$$

$\mathcal{L}_{\text{MSE}}$ is framewise joint-position MSE. $\mathcal{L}_{\text{bone}}$ penalizes deviations between output and target bone lengths: $\frac{1}{NT|\mathcal{E}|}\sum_{n,t,b}(|\hat{b}_{n,t,b}|-|b_{n,t,b}|)^2$. $\mathcal{L}_{\text{smooth}}$ penalizes joint acceleration magnitude to suppress jitter, $\mathcal{L}_{\text{vel}}$ enforces framewise velocity matching between generated and target joints, and $\mathcal{L}_{\text{EE}}$ applies position and velocity MSE on 7 end-effectors (hands, feet, head). $\mathcal{L}_{\text{joint}}$ penalizes anatomical range violations with a hinge loss (limits in Section S10), and $\mathcal{L}_{\text{foot}}$ suppresses velocity at ground-contact frames detected from near-zero ground-truth foot speeds.

*Teacher forcing.* During training, each batch uses teacher forcing (feeding ground-truth frames as decoder input) with probability $p_{\text{TF}}$, otherwise the decoder runs fully autoregressively. Stage 2 decays $p_{\text{TF}}$ linearly from 1.0 to 0.5; Stage 3 from 0.5 to 0.3. At inference, $p_{\text{TF}} = 0$ (fully autoregressive).

**Stage 3: End-to-End Fine-Tuning** All components are unfrozen and jointly optimized under the combined objective $\mathcal{L}_{\text{S3}} = \mathcal{L}_{\text{S1}} + \mathcal{L}_{\text{S2}} + \mathcal{L}_{\text{out}}$ at one-tenth of the base learning rate ($5\times10^{-5}$), where $\mathcal{L}_{\text{out}}$ denotes the output-level supervision losses described below. This lets encoders and decoder co-adapt while the disentanglement and physical losses prevent re-entanglement and artifact accumulation.

*Output-level supervision.* Encoder- and reconstruction-level losses do not directly supervise the retargeted output, so the decoder can re-entangle information through the adaptive gate. We therefore add four losses on the decoded output $\hat{\mathbf{s}}$ in Stage 3: two cooperative (an *output action classifier* $\mathcal{L}_{\text{out-AR}}$ and an *enhanced end-effector* loss) and two adversarial (a *distribution discriminator* pushing $\hat{\mathbf{s}}$ toward the raw-data distribution and an *output identity adversary* suppressing identity in $\hat{\mathbf{s}}$). The adversarial pair drives the privacy gain while the cooperative pair preserves utility (Section S16); the lightweight auxiliary heads are discarded at inference.

*Why Stage 3 dominates.* Stage 3 trains all components together under every loss and does most of the optimization: the stage ablation (Sec. 4.5) shows configurations including Stage 3 all reach ∼82% pre-trained AR while Stage 1 alone collapses to 57.3%. The earlier stages act as staged initialization—Stage 1 pre-separates the encoders, Stage 2 stabilizes reconstruction—so Stage 3 refines rather than discovers.

*Inference procedure.* At inference, the target skeleton is drawn uniformly at random from a pool of reference subjects distinct from the source. The identity encoder extracts $\mathbf{H}_{\text{identity}}$ from a single reference sequence of the target subject (any action suffices, since the identity encoder discards temporal dynamics). The decoder then generates the retargeted motion autoregressively. When source and target are the same person, the model performs self-reconstruction; we use this setting during training but not at inference, where the goal is identity transfer.

### 3.3 Partial Retargeting

Full retargeting ($\hat{\mathbf{s}} = D(\mathbf{H}_{\text{action}}, \mathbf{H}_{\text{identity}})$) replaces the source identity entirely, maximizing privacy but producing output whose joint statistics differ from raw data, so *pre-trained* downstream models (trained on raw skeletons, applied without retraining) fail on it. To enable compatibility with such models, we introduce *partial retargeting*, a residual blend applied after autoregressive generation:

$$\hat{\mathbf{s}}_\beta = \beta \cdot D(\mathbf{H}_{\text{action}}, \mathbf{H}_{\text{identity}}) + (1 - \beta) \cdot \mathbf{s}_{\text{src}}, \tag{16}$$

where the *retargeting ratio* $\beta \in [0, 1]$ is fixed before training and held constant. At $\beta$=1.0 this is full retargeting (maximum privacy); at $\beta$=0.2 the residual preserves 80% of the source structure and the decoder supplies 20%. Training end-to-end with the skip connection active makes the decoder learn output *complementary* to the source; each $\beta$ is a separate run. We use $\beta$=0.2 as the recommended operating point (Sec. 4).

### 3.4 Implementation Details

The action encoder produces $d_A$=768-dimensional per-frame features and the identity encoder produces $d_I$=256-dimensional global features; the factorized decoder has 6 layers with 8-head attention ($d_D$=320). The full model contains 22.7M parameters: 4.9M in the action encoder (including 2.3M for the MixFormer backbone), 0.8M in the identity encoder, and 17.0M in the decoder. The asymmetric encoder capacity (∼6× ratio) reflects the intentional information bottleneck on identity. Full training details are reported in Sec. 4.

*Loss weights.* An Optuna [3] TPE study informed the relative importance of the loss terms, but its best trial underperformed at full scale, so the final weights were set manually from those rankings (details in Section S8). The Stage-1 weights (Eq. (10)) are $\lambda_{\text{AR}}$=3.0, $\lambda_{\text{RI}}$=1.0, $\lambda_{\text{ctr}}$=1.0, $\lambda_{\text{adv}}$=0.5, $\lambda_{\text{orth}}$=0.1, $\lambda_{\text{CC}}$=0.01; the Stage-2 weights (Eq. (15)) are $\lambda_{\text{MSE}}$=1.0, $\lambda_{\text{EE}}$=2.0, $\lambda_{\text{bone}}$=1.0, $\lambda_{\text{vel}}$=0.2, $\lambda_{\text{smooth}}$=0.1, $\lambda_{\text{foot}}$=1.0, $\lambda_{\text{joint}}$=1.0 (Stage 3 reuses both sets plus the output-level losses).

## 4 Experiments

We evaluated DisentangledTMR on three skeleton action recognition benchmarks: NTU RGB+D 60 [26], NTU RGB+D 120 [19], and ETRI-Activity3D [14]. All experiments use cross-view protocols, measuring utility via action recognition accuracy (AR) and privacy via re-identification accuracy (RI). Unless otherwise stated, results refer to NTU60.

### 4.1 Datasets and Setup

All three benchmarks use Kinect v2 skeletons (25 joints) with cross-view protocols; multi-person actions were excluded to match our single-person design.

**NTU60** [26]: 56,880 sequences, 60 actions, 40 subjects; cameras 2 & 3 for training, camera 1 for testing (49 single-person classes after filtering). Cross-subject splits are unsuitable because re-identification requires all actors in both the train and test sets. **NTU120** [19]: 114,480 sequences, 120 actions, 106 subjects; 94 single-person classes. **ETRI** [14]: 112,620 sequences, 55 daily activities, 100 subjects, 8 viewpoints; cameras $\{2, 3, 5, 7\}$ train, $\{1, 4, 6, 8\}$ test. Paired training sets use 2×2 actor-action quadruplets $(P_a, P_b) \times (A_i, A_j)$ sampled for diversity (50k quadruplets for NTU60). All sequences are normalized to $T$=64 frames, root-centered, de-rotated, and translation-removed following SGN [39]; global scale is preserved and enforced via bone lengths in decoding.

### 4.2 Implementation Details

We trained with Adam [17], batch size 128, and loss weights set manually with guidance from an Optuna [3] importance study (values in Sec. 3.4; study details in Section S8). Training runs 20/15/20 epochs for Stages 1/2/3, totaling ∼28 GPU-hours on a single GPU. A detailed computational cost comparison with baselines is provided in the supplementary (Section S12).

### 4.3 Baselines and Evaluation

We compare against **Raw Skeleton** (no protection), **DMR** [7] (retargeting without adversarial training), and **PMR** [7] (retargeting with adversarial learning) as the most directly comparable retargeting baselines; a reference **Gaussian noise** perturbation ($\sigma$=0.05) is reported only in the reconstruction (MSE) analysis (Section S4) as an input-corruption sanity check. Other skeleton anonymization methods [8, 11, 22, 33, 34] differ along several protocol axes at once (output form, target-identity conditioning, privacy metric), so off-the-shelf numerical comparison is not meaningful; Section S14 tabulates these incompatibilities. The key distinction is that in-place perturbation methods keep the original skeletal identity, whereas retargeting replaces it, so their RI numbers measure different quantities; the closest aligned comparators are DMR and PMR.

**Privacy** is measured by re-identification accuracy (RI, ↓; random chance ≈2.5% for 40 subjects) and **utility** by action recognition accuracy (AR, ↑), evaluated under two complementary protocols. *Pre-trained* (primary): SGN [39] classifiers trained to convergence on raw data are applied directly to retargeted output *without retraining*, matching the common deployment in which downstream models cannot be retrained on anonymized data; both AR and RI use this protocol. *Re-trained AR* (secondary): a fresh SGN action classifier is trained from scratch on retargeted data, a more conservative utility estimate. We report re-trained AR but not re-trained RI, since retraining an identity classifier on anonymized data assumes labeled identities in the anonymized domain, the very thing anonymization prevents. All retargeting methods reduce AR vs. raw data; the relevant comparison is among anonymization methods.

**Table 1:** Main results on NTU60 (X-View) at $\beta$=0.2. *Pre-trained* AR/RI: raw-trained evaluators applied directly to retargeted output (ours: mean ± std over 5 runs). *Re-trained* AR: classifier retrained on retargeted data. Physical-plausibility metrics are computed against the ground-truth target. Random chance: AR ≈ 2%, RI ≈ 2.5%. Best value per column among anonymization methods in **bold** (pre-trained RI excluded, as PMR reaches a lower value only by collapsing AR).

| | Pre-trained SGN | | Re-trained Recon. | | Physical Plausibility | | | |
|---|---|---|---|---|---|---|---|---|
| Method | AR ↑ | RI ↓ | AR ↑ | MSE ↓ | MPJPE ↓ | Bone ↓ | Smooth ↓ | Vel ↓ |
| Raw Skeleton | 89.1 | 75.4 | 89.1 | — | — | — | — | — |
| DMR [7] | 49.1 | 25.7 | 43.1 | 0.047 | 0.202 | 0.029 | **0.023** | 0.001 |
| PMR [7] | 35.7 | 7.8 | 19.9 | 0.050 | 0.272 | 0.051 | 0.094 | 0.002 |
| Ours ($\beta$=0.2) | **75.8±1.1** | 18.1±1.8 | **87.1** | **0.030** | **0.113** | **0.016** | 0.029 | **0.0005** |

### 4.4 Main Results

Table 1 compares all methods. DMR achieves moderate AR but poor privacy; identity leaks through motion dynamics despite structural anonymization. PMR's adversarial training improves privacy but, with a single-encoder pipeline, collapses utility (35.7% pre-trained AR), over-anonymizing rather than disentangling.

At the recommended operating point $\beta$=0.2, DisentangledTMR achieves 75.8 ± 1.1% pre-trained SGN AR (mean ± std over 5 runs), more than double PMR's 35.7% under the same protocol, while reducing pre-trained RI from raw 75.4% to 18.1 ± 1.8%. Under the re-trained protocol it retains 87.1% AR, and it attains the lowest reconstruction MSE (0.030 vs. 0.047–0.050). PMR reaches a lower pre-trained RI (7.8%) only because its output is so degraded that even action becomes unreadable (19.9% re-trained AR). Dedicated encoders sidestep the action-identity conflict that limits single-encoder approaches; ablations (Sec. 4.5) show architectural separation underlies the low-RI regime, while output-level supervision and $\beta$ set the operating point. The full $\beta$ sweep (Sec. 4.7) reveals a sharp utility threshold at $\beta \approx 0.2$.

*Physical plausibility.* Among retargeting methods, DisentangledTMR achieves the lowest MPJPE (0.113) and roughly 2× better bone-length consistency than DMR (0.016 vs. 0.029), a direct consequence of the identity encoder's explicit bone-length modeling through the spatial GCN and bone MLP (Sec. 3.1). Its velocity error is also lowest, indicating physically plausible trajectories without jitter.

*Qualitative.* Across methods, DisentangledTMR preserves source pose trajectories while adopting different skeletal proportions and achieves the lowest MSE against ground truth, whereas DMR and PMR show visible motion distortion; DMR visually retains more source-identity characteristics while PMR and ours shift more clearly toward the target. Side-by-side comparisons and partial retargeting visualizations are in the supplement (Sections S13, S15).

### 4.5 Ablation Study

*Component ablations.* We validated the architectural choices through component ablations: *Static Identity* (GCN replaced by a fixed bone-length vector), *Full-Seq Identity* (temporal sequence instead of the time-averaged pose), *No Adversarial*/*No Orthogonality* (removing the corresponding Stage-1 losses), and *No Action Backbone*/*No Temporal Convs* (removing one of the two action-encoder pathways). Each degrades the privacy–utility trade-off: removing either action pathway causes the largest utility drops (the two pathways are complementary), *Static Identity* worsens privacy (the learned GCN captures richer structure than raw bone lengths), and *Full-Seq Identity* preserves privacy but loses utility (temporal redundancy in the identity stream disrupts the decoder's gating). Architectural separation keeps every variant within a low-RI regime that no single encoder component is responsible for; output-level supervision (Section S16) and the ratio $\beta$ then set the operating point within that regime.
*Stage ablation.* Ablating the curriculum at $\beta$=0.2 (Section S3) shows that all configurations including Stage 3 cluster at ∼82% pre-trained AR, so end-to-end fine-tuning is the dominant contributor; Stage 1 alone collapses to 57.3% AR (though it reaches the strongest privacy, 15.3% RI), confirming that encoder pre-training alone cannot produce action-preserving output, while Stage 2 stabilizes the decoder before joint optimization.

### 4.6 Additional Analyses

The supplement provides extensive additional analyses. On privacy, **dedicated adversarial re-identification classifiers** (Section S1; full-retargeting configuration), including attackers trained directly on retargeted data and MixFormer coverage on NTU120/ETRI, stay within ±0.3pp of the standard evaluator, indicating residual identity is genuinely removed rather than merely hidden from a weak evaluator; and a **sensitivity analysis** (Section S2) shows graceful degradation under input noise ($\sigma$ up to 0.20), joint dropout (up to 30%), and reduced temporal resolution (down to 8 frames). On the model, a **reconstruction analysis** (Section S4) confirms balanced source/target transfer, a **decoder gate analysis** (Section S5) shows end-effectors prefer action features while torso joints prefer identity, and **embedding visualizations** (Section S7) show tight action clusters with uniformly scattered identities.

We further provide **HPO details** (Section S8), **dataset statistics** (Section S9), **joint angle limits** (Section S10), **extended limitations and broader impact** (Section S11), a **computational cost comparison** (Section S12), more **visualizations** (Section S13), a **protocol-alignment table** for excluded anonymization methods (Section S14), and the **full $\beta$ sweep** (Section S15).

### 4.7 Partial Retargeting Results

At full retargeting ($\beta$=1.0), pre-trained evaluators collapse to 2.0% AR because the body proportions change too much; partial retargeting (Sec. 3.3) makes this

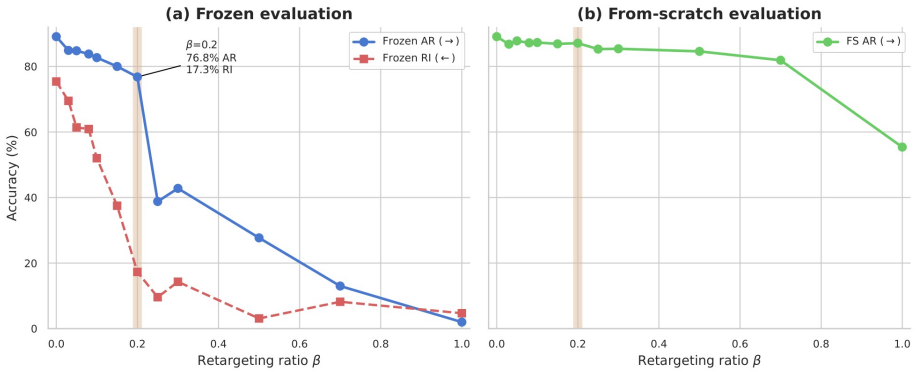


**Fig. 5:** Partial-retargeting trade-off vs. $\beta$ on NTU60. **(a)** Pre-trained: AR (blue) drops sharply between $\beta$=0.2 and 0.25, RI (red) falls rapidly; shaded region marks $\beta$=0.2. **(b)** Re-trained AR (green) degrades gradually, confirming action is preserved across $\beta$.

**Table 2:** Cross-dataset partial-retargeting sweep (pre-trained SGN AR/RI, %; single run). NTU120/ETRI use earlier checkpoints without the full output-level supervision suite; the NTU60 $\beta$=0.2 row matches Tab. 1. Chance RI: ≈2.5% (NTU60), ≈0.9% (NTU120), ≈1% (ETRI).

| | NTU60 | | NTU120 | | ETRI | |
|---|---|---|---|---|---|---|
| $\beta$ | AR ↑ | RI ↓ | AR ↑ | RI ↓ | AR ↑ | RI ↓ |
| 0.00 (raw) | 89.1 | 75.4 | 88.2 | 67.6 | 90.8 | 45.6 |
| 0.10 | 82.7 | 52.0 | 85.3 | 62.4 | 90.0 | 43.0 |
| 0.20 | 76.8 | 17.3 | 81.4 | 39.4 | 87.8 | 35.7 |
| 0.30 | 42.8 | 14.3 | 73.5 | 22.8 | 84.0 | 26.0 |
| 0.50 | 27.7 | 3.1 | 47.1 | 9.9 | 67.3 | 10.4 |
| 1.00 | 2.0 | 4.7 | 2.6 | 0.9 | 3.7 | 1.0 |

tunable. Figure 5 sweeps $\beta$ on NTU60 and reveals a **sharp threshold at** $\beta \approx 0.2$: below it, pre-trained SGN AR stays above 76% (80% of the source structure is retained, so pre-trained models still read the action); between $\beta$=0.20 and 0.25 it collapses (76.8%→38.8%) as the retargeted proportions dominate. Re-trained AR degrades only gradually (to 55.4% at $\beta$=1.0), confirming action information survives across the full range. The full numeric sweep is in Section S15.

### 4.8 Cross-Dataset and Generalization

Table 2 sweeps $\beta$ across all three benchmarks. RI decreases smoothly with $\beta$ on every dataset, confirming the disentanglement generalizes. NTU60 (trained with the full output-level supervision suite) shows a sharp AR cliff between $\beta$=0.2 and 0.3 (76.8%→42.8%), whereas NTU120 and ETRI (earlier checkpoints without that suite) decline gradually, indicating that output-level supervision sharpens the pre-trained-compatibility window. At $\beta$=0.2 all datasets retain

high pre-trained AR (76.8/81.4/87.8%) while cutting RI well below raw data (75.4/67.6/45.6% → 17.3/39.4/35.7%).

## 5 Limitations

Our privacy metric captures empirical risk reduction but does not constitute a formal guarantee; dedicated metric-learning or model-inversion attacks could potentially achieve higher RI. At $\beta$=0.2 the pre-trained utility gap from raw data (75.8% vs. 89.1% AR, ∼13pp) reflects the residual distribution shift from changing body proportions; full retargeting ($\beta$=1.0) maximizes privacy but drops pre-trained AR to 2.0%, requiring downstream models to be retrained. Closing the pre-trained gap (*e.g.*, via domain adaptation) without re-introducing identity leakage remains open. All experiments use 25-joint Kinect v2 skeletons; the GCN is topology-agnostic in principle (it consumes the skeleton adjacency), but cross-topology transfer (*e.g.*, OpenPose, COCO, SMPL) remains an open limitation we do not claim. Training at batch size 128 requires ∼40 GB GPU memory, and all three datasets share the same sensor. Extended limitations and broader impact discussion are provided in the supplementary material (Section S11).

## 6 Conclusion

We presented DisentangledTMR, a motion retargeting architecture that separates action and identity through dual encoders with complementary inductive biases (temporal for action, spatial for identity) and a factorized decoder with adaptive per-joint gating. On NTU60 at the recommended operating point $\beta$=0.2, DisentangledTMR reduces pre-trained re-identification from 75.4% to 18.1% while retaining 75.8% pre-trained (87.1% re-trained) action recognition, more than doubling PMR's utility under the same protocol. The retargeting ratio $\beta$ exposes a continuous privacy–compatibility trade-off with a sharp threshold at $\beta \approx 0.2$: below it, pre-trained models retain utility; above it, full retargeting ($\beta$=1.0) maximizes privacy at the cost of requiring downstream retraining. Cross-dataset sweeps on NTU120 and ETRI show the disentanglement generalizes. The central finding is that architectural inductive biases that make entanglement structurally difficult provide a strong privacy *foundation*: even with individual encoder-level disentanglement losses removed, the dual-encoder design stays in a low re-identification regime that single-encoder methods cannot reach. Output-level adversarial supervision and the partial-retargeting ratio $\beta$ then set the operating point within that regime, together yielding a practical privacy–utility trade-off.

## Acknowledgements

This work was supported in part by the U.S. National Science Foundation (1840080, 2103829, and 2348391).

## References


1. Aberman, K., Li, P., Lischinski, D., Sorkine-Hornung, O., Cohen-Or, D., Chen, B.: Skeleton-aware networks for deep motion retargeting. ACM Transactions on Graphics (TOG) **39**(4), 62:1–62:14 (2020)
2. Aberman, K., Wu, R., Lischinski, D., Chen, B., Cohen-Or, D.: Learning character-agnostic motion for motion retargeting in 2d. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019). https://doi.org/10.1145/3306346.3322999
3. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 2623–2631 (2019)
4. Aziz, S., Komogortsev, O.: Exploring the uncoordinated privacy protections of eye tracking and VR motion data for unauthorized user identification. In: 2025 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 217–227 (2025). https://doi.org/10.1109/VR64551.2025.00043
5. Carr, T., Lu, A., Xu, D.: Linkage attack on skeleton-based motion visualization. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp. 3758–3762 (2023)
6. Carr, T., Xu, D., Lu, A.: A review of privacy and utility in skeleton-based data in virtual reality metaverses. In: 2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom). pp. 198–205 (2024). https://doi.org/10.1109/MetaCom62920.2024.00041
7. Carr, T., Xu, D., Yuan, S., Lu, A.: Privacy-centric deep motion retargeting for anonymization of skeleton-based motion visualization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13162–13170 (2025)
8. Carr, T., Zhao, Y., Xu, D., Lu, A.: Explanation-based anonymization methods for motion privacy. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. pp. 52–64. Springer (2025)
9. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. vol. 31 (2018)
10. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016)
11. Gao, X., Li, K., Liu, X., Nie, J., Chen, W., Tian, Y.: Privacy-preserving 3D skeleton-based video action recognition via graph convolution network. IEEE Transactions on Consumer Electronics **71**(2), 6627–6641 (2025). https://doi.org/10.1109/TCE.2024.3476273
12. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
13. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations (2017)
14. Jang, J., Kim, D., Park, C., Jang, M., Lee, J., Kim, J.: Etri-activity3d: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10990–10997. IEEE (2020)

15. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18661–18673 (2020)
16. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning. pp. 2649–2658. PMLR (2018)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
18. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition **98**, 107069 (2020)
19. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence **42**(10), 2684–2701 (2019)
20. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Sch"olkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: International Conference on Machine Learning. pp. 4114–4124. PMLR (2019)
21. Martinelli, G., Garau, N., Bisagno, N., Conci, N.: Moma: Skinned motion retargeting using masked pose modeling. Computer Vision and Image Understanding **249**, 104141 (2024)
22. Moon, S., Kim, M., Qin, Z., Liu, Y., Kim, D.: Anonymization for skeleton action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (2023)
23. Nair, V., Guo, W., O'Brien, J.F., Rosenberg, L., Song, D.: Deep motion masking for secure, usable, and scalable real-time anonymization of ecological virtual reality motion data. In: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp. 493–500. IEEE (2024)
24. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: AAAI Conference on Artificial Intelligence (2018)
25. Pfeuffer, K., Geiger, M.J., Prange, S., Mecke, L., Buschek, D., Alt, F.: Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12 (2019)
26. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1010–1019 (2016)
27. Shen, Y., Wen, H., Luo, C., Xu, W., Zhang, T., Hu, W., Rus, D.: Gaitlock: Protect virtual and augmented reality headsets using gait. IEEE Transactions on Dependable and Secure Computing **16**(3), 484–497 (2018)
28. Sun, R., Wang, H., Xue, M., Chen, H.T.: Privacy in motion: Implementing differential privacy for user motion in VR. In: Proceedings of the 36th Australasian Conference on Human-Computer Interaction (OzCHI). pp. 223–231 (2025). https://doi.org/10.1145/3726986.3727017
29. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)

31. Villegas, R., Yang, J., Ceylan, D., Lee, H.: Neural kinematic networks for unsupervised motion retargetting. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8639–8648 (2018)
32. Weiß, S., Bonenberger, C., Niedermaier, T., Knof, M., Schneider, M.: Privacy beyond the face: Assessing gait privacy through realistic anonymization in industrial monitoring. Sensors **26**(1), 187 (2026). https://doi.org/10.3390/s26010187
33. Wu, Z., Wang, Z., Wang, Z., Jin, H.: Towards privacy-preserving visual recognition via adversarial training: A pilot study. In: Computer Vision – ECCV 2018. Lecture Notes in Computer Science, vol. 11220, pp. 627–645. Springer (2018)
34. Xiao, T., Tsai, Y.H.H., Sohn, K., Chandraker, M., Yang, M.H.: Adversarial learning of privacy-preserving and task-oriented representations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12434–12441 (2020)
35. Xin, W., Miao, Q., Liu, Y., Liu, R., Pun, C.M., Shi, C.: Skeleton mixformer: Multivariate topology representation for skeleton-based action recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 2211–2220 (2023)
36. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. pp. 7444–7452. AAAI Press (2018)
37. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 12310–12320 (2021)
38. Zhang, J., Weng, J., Kang, D., Zhao, F., Huang, S., Zhe, X., Bao, L., Shan, Y., Wang, J., Tu, Z.: Skinned motion retargeting with residual perception of motion semantics & geometry. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13864–13872 (2023)
39. Zhang, P., Lan, C., Zeng, W., Xing, J., Xue, J., Zheng, N.: Semantics-guided neural networks for efficient skeleton-based human action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1112–1121 (2020)