REFM: SEMANTIC-AWARE REFINEMENT FLOW MODEL FOR MOTION RETARGETING

Anonymous authors
Paper under double-blind review

ABSTRACT

Motion retargeting transfers motion across characters with different skeletal structures while preserving semantic intent and physical plausibility. Despite recent progress, two fundamental questions remain: *(i) how can reliable source-motion semantics be learned without high-quality paired retargeting data, and (ii) how should retargeting be formulated when no reliable paired motion can serve as a definitive regression objective?* Existing methods commonly preserve semantics by constraining predictions toward copied motions. However, such initializations entangle useful articulation cues with artifacts caused by mismatched skeletal proportions and body geometry. Moreover, directly regressing a final motion in one forward pass is restrictive because retargeting is inherently underdetermined, and the desired solution must balance semantic fidelity with target-specific physical and temporal constraints rather than match a unique paired target. Motivated by these limitations, we propose **ReFM**, a source-mesh-agnostic, energy-guided model that reformulates motion retargeting as progressive refinement. First, an $\mathrm{SO}(3)$ canonicalizer removes redundant global-orientation variations. Second, a cross-character semantic encoder, pretrained through contrastive learning, provides a character-invariant representation for both optimization guidance and semantic evaluation. ReFM then progressively refines an initialized target motion through a learned flow guided by semantic consistency, physical plausibility, temporal coherence, and minimal motion modification. The framework is compatible with different initialization strategies, including both direct motion copying and Autodesk HumanIK (Autodesk, Inc., 2026), an industry-standard full-body inverse-kinematics retargeting system. Experiments show that ReFM can further refine HumanIK-transferred motions while consistently reducing self-penetration and maintaining strong semantic consistency. Extensive evaluations demonstrate that ReFM achieves a favorable balance between semantic preservation and physical plausibility. The project website is at `https://anonymous-user-tq.github.io/ReFM-Anonymous/`.

1 INTRODUCTION

Motion retargeting transfers motion from a source character to a target with different skeletal proportions and topologies while aiming to preserve semantic intent and physical plausibility. It is a fundamental problem in character animation, digital humans, and robotics, where the same motion must often be reused across substantially different body structures (Tak & Ko, 2005; Reda et al., 2023). Recent learning-based methods (Villegas et al., 2021; Zhang et al., 2023; Yang et al., 2025) directly predict target motions from source motions and have achieved substantial improvements in efficiency and generalization over traditional optimization-based pipelines. Meanwhile, simultaneously preserving motion semantics (Zhang et al., 2024a), satisfying physical constraints (Ayusawa & Yoshida, 2017), and generalizing across diverse characters remains challenging.

We revisit this problem through two fundamental questions. **First, how can the semantic meaning of a source motion be represented reliably when high-quality paired source–target retargeting supervision is sparse?** Limited cross-character correspondences do not provide dense supervision for arbitrary source–target pairs, so existing methods (Zhang et al., 2023; Yang et al., 2025; Ye et al., 2024) commonly rely on self-reconstruction and copied-motion consistency, constraining predictions toward source rotations replayed on the target skeleton. Such copied motion provides a useful initialization because it preserves much of the source articulation, but it is not a reliable

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

target. Applying identical joint rotations to characters with different bone lengths, body proportions, and surface geometry can introduce spatial misalignment, missing contacts, and self-penetration. Consequently, low-level similarity to this surrogate may propagate its errors into the retargeted result. Although Zhang et al. (2024a) introduce visual-language guidance, explicit semantic modeling directly in motion space, where skeletal geometry and temporal articulation are jointly represented, remains underexplored.

Second, how should motion retargeting be formulated when no unique target motion provides a definitive regression objective? Most existing neural methods directly regress a final target motion in a single forward pass. However, motion retargeting is inherently underdetermined: a satisfactory result is defined by jointly satisfying semantic fidelity, geometric compatibility, temporal coherence, and target-specific physical constraints, rather than by matching a unique paired target. Compressing these competing objectives into a one-step prediction provides no mechanism to reject detrimental modifications or continue improving the initial output. *In contrast, professional animators typically begin with a coarse retargeted motion and progressively correct semantic and physical artifacts until further editing no longer improves the result.* This motivates progressive refinement, in which an initialized motion is iteratively evaluated and improved under explicit quality criteria.

Based on these insights, we propose **ReFM**, a source-mesh-agnostic model that combines motion canonicalization, explicit semantic modeling, and energy-guided progressive refinement. ReFM first applies a parameter-free $\mathrm{SO}(3)$ canonicalizer to remove redundant global-heading variations and reduce the effective motion space. It then learns a cross-character semantic encoder through contrastive pretraining, providing a character-invariant representation for both semantic guidance and evaluation. Finally, rather than directly regressing a final target motion, ReFM progressively refines an initialized motion through a learned flow guided by semantic fidelity, physical plausibility, and temporal coherence. This formulation enables ReFM to preserve source-motion semantics while adapting to target-specific skeletal and geometric constraints without requiring the source character mesh, and consistently improves both naive copying and industry-standard initializations.

2 Related Work

Skinned Motion Retargeting. Skinned motion retargeting uses source surface geometry to measure contact, proximity, and penetration beyond sparse skeletal joints. Classical optimization methods preserve salient kinematic or spatial constraints across characters with different proportions (Choi & Ko, 2000; Tak & Ko, 2005), while geometry-aware formulations further exploit surface relationships to preserve self-contact and near-body interactions (Jin et al., 2018; Liu et al., 2018; Basset et al., 2020). Recent methods extend this geometric reasoning: CAR preserves detected self-contacts and suppresses interpenetration through geometry-conditioned optimization (Villegas et al., 2021); MeshRet aligns dense mesh-interaction fields to model both contact and non-contact body-part relationships (Ye et al., 2024); STaR introduces dense shape representations, limb penetration constraints, and temporal consistency to jointly improve geometric plausibility and motion smoothness (Yang et al., 2025); ReConForM uses rigged key vertices and adaptive weighting for real-time contact-aware retargeting (Cheynel et al., 2025a); and spatially adaptive interaction guidance is utilized to handle exaggerated target morphologies (Choi et al., 2026).

Skin-Agnostic Motion Retargeting. In practical pipelines, motion data and character assets are often acquired independently: motion-capture systems recover skeletal sequences, while meshes, rigs, and skinning weights are authored separately, and motion libraries provide skeletal animations for transfer to new characters (Chen et al., 2021; Mourot et al., 2023). *Therefore, the source animation is frequently available only as skeletal motion without its original mesh or skinning information.* Skin-agnostic retargeting methods address this limitation by preserving motion semantics using skeletal structure and motion dynamics. For instance, NKN uses forward kinematics and cycle consistency (Villegas et al., 2018), PMnet disentangles pose and global movement (Lim et al., 2019), SAN handles different skeleton topologies through skeleton-aware operators (Aberman et al., 2020), SAME learns a skeleton-agnostic motion embedding (Lee et al., 2023), and PAN performs body-part-level retargeting with pose-aware attention (Hu et al., 2024). More recent target-geometry-aware but source-mesh-agnostic methods, such as R2ET and M-R2ET, combine skeleton-aware semantic losses with target-shape-aware geometric correction (Zhang et al., 2023; 2024b). MoCaNet (Zhu et al., 2022) performs in-the-wild motion retargeting by disentangling motion, body structure, and canonicalized camera views from 2D skeletal sequences, without relying on source-character mesh or skinning information. Our work follows this source-mesh-agnostic setting, but explicitly compresses

the motion space through canonicalization, learns a semantic motion embedding from cross-character positives, and refines the copied motion through a progressive refinement flow.

3 PRELIMINARIES

3.1 SKIN-AGNOSTIC MOTION RETARGETING

Let $\mathcal{S}$, $\mathcal{X}$, and $\mathcal{M}$ denote the spaces of skeletons, character meshes, and motion sequences, respectively. A motion sequence is represented as

$$\mathbf{m} = \{\mathbf{q}_{1:T}, \mathbf{r}_{1:T}\} \in \mathcal{M}, \tag{1}$$

where $\mathbf{q}_{1:T}$ denotes local joint rotations and $\mathbf{r}_{1:T}$ denotes root motion over T frames.

Given a source motion $\mathbf{m}_s$, its source skeleton $\mathbf{s}_s$, a target skeleton $\mathbf{s}_t$, and the target mesh $\mathbf{x}_t$, skin-agnostic retargeting aims to predict

$$\hat{\mathbf{m}}_t = f(\mathbf{m}_s, \mathbf{s}_s, \mathbf{s}_t, \mathbf{x}_t), \quad f : \mathcal{M} \times \mathcal{S} \times \mathcal{S} \times \mathcal{X} \to \mathcal{M}. \tag{2}$$

In practice, the source mesh $\mathbf{x}_s$ is typically unavailable, while the target mesh $\mathbf{x}_t$ is available. Importantly, the source motion does not in general specify a unique target motion. Instead, we denote by $\mathcal{M}_t^\star \subseteq \mathcal{M}$ the set of admissible target motions that preserve the semantic intent of $\mathbf{m}_s$ while satisfying the structural and geometric requirements of the target character. The retargeting objective is therefore to obtain $\hat{\mathbf{m}}_t \in \mathcal{M}_t^\star$ rather than to regress toward a uniquely defined paired target.

3.2 GROUP CANONICALIZATION

Let $\mathcal{G}$ be a transformation group acting on an input space $\mathcal{Z}$, where $\mathbf{z} \in \mathcal{Z}$ denotes a motion-related input, e.g., $(\mathbf{m}, \mathbf{s})$ or $(\mathbf{m}, \mathbf{s}, \mathbf{x})$. In motion retargeting, the global character orientation is represented in $\mathrm{SO}(3)$. However, different rotational components need not have the same semantic role. When the ground plane is fixed, changes in global facing direction correspond to rotations around the global up axis and generally do not alter the underlying motion semantics, whereas pitch and roll may encode meaningful pose or motion characteristics. We therefore treat the yaw subgroup of $\mathrm{SO}(3)$ as the nuisance transformation group in ReFM.

A canonicalizer estimates a group element

$$\kappa : \mathcal{Z} \to \mathcal{G}, \qquad \bar{\mathbf{z}} = \kappa(\mathbf{z})^{-1} \cdot \mathbf{z}, \tag{3}$$

where $\bar{\mathbf{z}}$ is the canonicalized representation. Ideally, κ is equivariant to the group action,

$$\kappa(g \cdot \mathbf{z}) = g\kappa(\mathbf{z}), \qquad \forall g \in \mathcal{G}, \tag{4}$$

which directly yields invariance:

$$\overline{g \cdot \mathbf{z}} = \kappa(g \cdot \mathbf{z})^{-1} \cdot (g \cdot \mathbf{z}) = \kappa(\mathbf{z})^{-1} \cdot \mathbf{z} = \bar{\mathbf{z}}. \tag{5}$$

Thus, canonicalization removes nuisance group variations before learning. In our setting, motions that differ only in their global facing direction are mapped to a shared canonical space, while pitch, roll, and parent-relative articulation are preserved. This removes redundant global-orientation variations without discarding rotational information that may correlate with motion semantics, allowing the model to focus on semantic preservation and target-character adaptation.

4 METHODOLOGY

As discussed in Sec. 2, practical motion retargeting often starts from skeletal animation alone, while the source-character mesh and skinning information are unavailable. We therefore follow this source-mesh-agnostic setting and require no source mesh geometry throughout the retargeting process. Given a source motion $\mathbf{m}_s$, source skeleton $\mathbf{s}_s$, target skeleton $\mathbf{s}_t$, and target mesh $\mathbf{x}_t$, our goal is to generate a retargeted motion $\hat{\mathbf{m}}_t$ that preserves the semantic intent of $\mathbf{m}_s$ while remaining temporally coherent and physically plausible on the target character. As illustrated in Fig. 1, ReFM is designed to address the two questions introduced in Sec. 1. Before addressing them, we apply a parameter-free $\mathrm{SO}(3)$ canonicalizer that estimates the initial global-heading component g and transforms the source motion into a unified facing direction. By removing orientation-dependent variations that are irrelevant to motion semantics, canonicalization reduces the effective learning space for subsequent semantic representation learning and motion refinement.

4.1 SO(3) CANONICALIZER

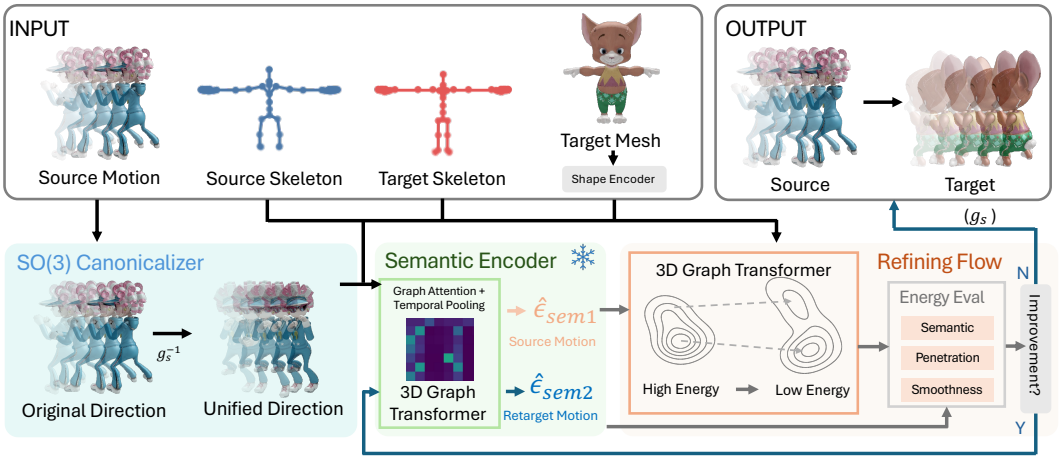
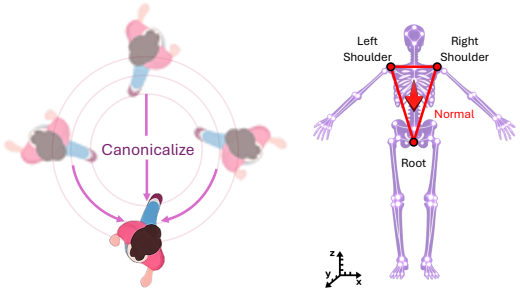


Figure 1: Overview of the proposed ReFM framework. The $\mathrm{SO}(3)$ canonicalizer first removes redundant global orientation from the source motion. A pretrained cross-character semantic encoder then extracts character-invariant representations from the source and candidate target motions, while a shape encoder provides geometric conditioning from the target mesh. Starting from an initialized motion, the semantic-aware refinement flow progressively applies local corrections that reduce a joint energy including semantic deviation, self-penetration, and temporal inconsistency.

Following the group canonicalization formulation in Sec. 3.2, we consider the global facing direction as a nuisance component of the character's $\mathrm{SO}(3)$ orientation. In practical motion libraries, the same action may be captured or authored under different stage layouts, coordinate systems, or initial headings. For example, walking motions facing different directions, as illustrated in Fig. 2(a), exhibit different global orientations while preserving the same semantic content and parent-relative articulation. Importantly, we do not canonicalize the complete global rotation. With a fixed ground plane, pitch and roll may encode meaningful pose or motion characteristics, whereas yaw primarily determines the character's global facing direction. We therefore remove only this redundant heading component while preserving the remaining rotational information. Such canonicalization before semantic encoding and motion refinement reduces redundant variations in the learning space.



(a) Canonicalizer (b) Body-orientation triangle

Figure 2: Illustration of the proposed canonicalizer. (a) Motions with identical semantics but different global headings are aligned to a shared canonical orientation. (b) Intuitive geometric interpretation of the character's facing direction, illustrated by the normal of the body-orientation plane formed by the root and shoulder joints.

Geometrically, the facing direction can be interpreted as the horizontal normal of the body-orientation plane formed by the root and two shoulder joints, as illustrated in Fig. 2(b). Given a source motion $\mathbf{m}_s = \{\mathbf{q}_{1:T}, \mathbf{r}_{1:T}\}$, let $\mathbf{q}_1^{\mathrm{root}}$ denote its root orientation in the first frame. We obtain the corresponding facing direction by rotating a predefined canonical forward axis $\mathbf{e}_{\mathrm{fwd}}$ and projecting it onto the horizontal plane:

$$\tilde{\mathbf{f}}_s = \Pi_{\mathrm{hor}} R\big(\mathbf{q}_1^{\mathrm{root}}\big)\, \mathbf{e}_{\mathrm{fwd}}, \qquad \mathbf{f}_s = \frac{\tilde{\mathbf{f}}_s}{\|\tilde{\mathbf{f}}_s\|_2}, \qquad \Pi_{\mathrm{hor}} = I - \mathbf{e}_{\mathrm{up}}\mathbf{e}_{\mathrm{up}}^{\top}, \tag{6}$$

where $R(\cdot)$ converts a quaternion into its corresponding rotation matrix and $\mathbf{e}_{\mathrm{up}}$ denotes the fixed global up axis. The horizontal projection isolates the heading component used for canonicalization while leaving pitch- and roll-related information unconstrained.

We then define $g_s \in \mathrm{SO}(3)$ as the yaw rotation around $\mathbf{e}_{\mathrm{up}}$ that maps the canonical forward direction to the estimated facing direction:

$$g_s \mathbf{e}_{\mathrm{fwd}} = \mathbf{f}_s, \qquad g_s \mathbf{e}_{\mathrm{up}} = \mathbf{e}_{\mathrm{up}}. \tag{7}$$

Although g_s is represented as an element of $\mathrm{SO}(3)$, it belongs specifically to the yaw subgroup introduced in Sec. 3.2. For the quaternion motion representation used by the semantic encoder and refinement model, canonicalization removes this common heading component from the root rotation

of every frame:

$$\bar{\mathbf{q}}_t^{\text{root}} = q(g_s^{-1}) \otimes \mathbf{q}_t^{\text{root}}, \qquad \bar{\mathbf{q}}_t^{j} = \mathbf{q}_t^{j}, \quad j \neq \text{root}, \qquad t = 1, \ldots, T, \tag{8}$$

where $q(g_s^{-1})$ denotes the quaternion corresponding to g_s^{-1} and $\otimes$ denotes quaternion composition. The root translation is also transformed accordingly. Thus, the same first-frame yaw correction is applied uniformly across the sequence, while pitch, roll, and all parent-relative joint rotations remain unchanged.

The canonicalizer is parameter-free and exactly invertible: the original global heading can be restored by reapplying g_s to the canonicalized root rotations after retargeting. By removing only the semantically redundant facing-direction component of global $\mathrm{SO}(3)$ orientation, the canonicalizer reduces the effective motion space without discarding rotational information that may characterize the underlying action, enabling the subsequent semantic encoder and refinement flow to focus on motion semantics and target-character adaptation.

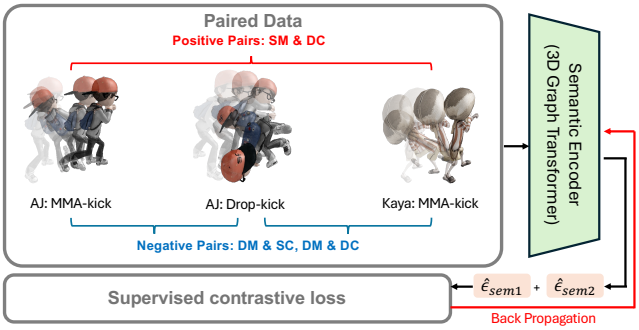


Figure 3: Supervised contrastive pretraining of the cross-character semantic encoder. Cross-character positives contain the same motion performed by different characters (SM+DC), while negatives contain different motions performed by either the same character (DM+SC) or different characters (DM+DC). The encoder is pretrained and subsequently frozen in ReFM training.

4.2 Cross-Character Semantic Encoder

We next address the first question: *how can source-motion semantics be represented reliably when high-quality paired source–target supervision is sparse?* Although copied motion retains useful articulation cues, skeletal and geometric differences can introduce severe geometric misalignment, e.g., self-penetration, making it an unreliable semantic target.

We therefore exploit limited cross-character correspondences to learn a transferable representation. Motions sharing the same motion identity across different characters provide positive pairs, while different motion identities provide negatives. In motion retargeting, semantic consistency means preserving motion identity and characteristic articulation across characters, rather than merely matching a broad action category. The canonicalization in Sec. 4.1 removes redundant global-heading variations, allowing the encoder to focus on character-independent motion patterns. Given a canonicalized motion $\bar{\mathbf{m}}^{(i)}$ and its rest-pose skeleton $\mathbf{s}^{(i)}$, the semantic encoder E_{sem} produces an L_2-normalized clip-level embedding

$$\hat{\boldsymbol{\epsilon}}_{\text{sem}}^{(i)} = \frac{E_{\text{sem}}\left(\bar{\mathbf{m}}^{(i)}, \mathbf{s}^{(i)}\right)}{\left\| E_{\text{sem}}\left(\bar{\mathbf{m}}^{(i)}, \mathbf{s}^{(i)}\right) \right\|_2} \in \mathbb{R}^d. \tag{9}$$

We instantiate E_{sem} with a 3D graph Transformer that jointly encodes the canonicalized motion and rest-pose skeleton into a clip-level representation.

As illustrated in Fig. 3, positives correspond to SM+DC, while negatives include DM+SC and DM+DC. During training, all samples sharing the anchor's motion identity are treated as positives. Let $\mathcal{B}$ denote the batch index set, $\mathcal{P}(i) \subseteq \mathcal{B} \setminus \{i\}$ the positive indices for anchor i, and s_{ia} the cosine similarity between normalized embeddings. The training objective is

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \log \frac{|\mathcal{P}(i)|^{-1} \sum_{p \in \mathcal{P}(i)} \exp(s_{ip}/\tau)}{\sum_{a \in \mathcal{B}} \exp(s_{ia}/\tau)}, \tag{10}$$

where $\mathcal{I}$ contains anchors with at least one positive and τ is the temperature. We use a variant of the supervised contrastive objective (Khosla et al., 2020), averaging positives inside the logarithm while retaining the anchor's self-similarity in the denominator. This encourages motion-dependent representations while suppressing character-specific variation.

After pretraining, E_{sem} is frozen throughout ReFM training. Cosine distance between source and candidate embeddings defines the semantic energy $\mathcal{E}_{\text{sem}}$, while cosine similarity is reported as the

semantic-consistency metric. Thus, semantic guidance is independent of copied-motion similarity. We further evaluate cross-character invariance and motion discrimination on held-out motion groups in Appendix A.

4.3 Energy-Guided Refinement Flow

We now address the second question: *how should motion retargeting be formulated when no unique target motion provides a definitive regression objective?* Most learning-based methods directly predict a final retargeted motion in a single forward pass. However, motion retargeting is inherently underdetermined, and a satisfactory solution must jointly satisfy multiple constraints rather than match a uniquely defined paired target. Professional animators instead begin with a coarse retargeted motion, progressively correct its semantic and physical artifacts, and stop when further editing no longer improves the result. Following this refinement workflow, we formulate motion retargeting as iterative energy reduction. Starting from an initialized target motion $\mathbf{m}_t^{(0)}$, obtained by motion copying or an inverse-kinematics solver, ReFM progressively predicts local corrections and evaluates their quality under explicit motion-energy criteria rather than directly regressing a complete target motion in one step. Importantly, ReFM does not aim to model or sample the full distribution of admissible retargeted motions; the underdetermined nature of retargeting instead motivates a refinement formulation that searches for an improved solution from a given initialization.

Motion energy. As illustrated in Fig. 1, each candidate motion $\mathbf{m}$ is evaluated through five energy terms that jointly enforce semantic fidelity, physical plausibility, temporal coherence, and minimal modification from the initialization. The overall motion energy is formulated as

$$\begin{aligned}\mathcal{E}(\mathbf{m}) = \lambda_{\text{sem}}\mathcal{E}_{\text{sem}}(\mathbf{m}, \bar{\mathbf{m}}_s) + \lambda_{\text{init}}\mathcal{E}_{\text{init}}(\mathbf{m}, \mathbf{m}_t^{(0)}) + \lambda_{\text{pen}}\mathcal{E}_{\text{pen}}(\mathbf{m}) \\ + \lambda_{\text{smo}}\mathcal{E}_{\text{smo}}(\mathbf{m}) + \lambda_{\text{curv}}\mathcal{E}_{\text{curv}}(\mathbf{m}, \mathbf{m}_t^{(0)}),\end{aligned} \tag{11}$$

where $\mathcal{E}_{\text{sem}}$ preserves motion semantics by measuring the discrepancy between the frozen semantic representations of the canonicalized source motion $\bar{\mathbf{m}}_s$ and the candidate target motion $\mathbf{m}$, while $\mathcal{E}_{\text{init}}$ discourages unnecessary deviation from the initialized motion $\mathbf{m}_t^{(0)}$ and thereby preserves its useful articulation. For physical plausibility, $\mathcal{E}_{\text{pen}}$ penalizes self-penetration on the target character. For temporal coherence, $\mathcal{E}_{\text{smo}}$ suppresses local temporal inconsistency, while $\mathcal{E}_{\text{curv}}$ regularizes excessive deformation of the motion trajectories relative to the initialization. While $\mathcal{E}_{\text{pen}}$ and $\mathcal{E}_{\text{init}}$ build upon conventional retargeting objectives (Yang et al., 2025), $\mathcal{E}_{\text{sem}}$, $\mathcal{E}_{\text{smo}}$, and $\mathcal{E}_{\text{curv}}$ are introduced in ReFM. Their formulations and weights are provided in Appendix B. The semantic-weight study is described in Appendix C. Together, these terms guide ReFM to improve semantic fidelity and physical plausibility while making modest refinements.

Conditional refinement field. We refer to the dynamics induced by iteratively applying the time-conditioned vector field v_θ as the refinement flow, which progressively transports the initialized motion toward lower-energy states. Given the current motion state $\mathbf{m}$, refinement time t, and condition $\mathbf{c}$, a spatio-temporal graph transformer predicts a per-frame, per-joint velocity $v_\theta(\mathbf{m}, t, \mathbf{c})$ with the same dimensionality as $\mathbf{m}$. The condition $\mathbf{c}$ comprises the source semantic representation, source and target skeleton features, and target-mesh shape feature. Spatial attention propagates information along the skeletal hierarchy, while temporal attention models each joint trajectory across frames.

Energy-guided training. Because retargeting admits multiple valid solutions, there is no unique ground-truth trajectory from an initialization to a satisfactory target motion. We therefore derive local refinement supervision directly from the differentiable motion energy. For a candidate state $\mathbf{m}$, we define an energy-decreasing target direction as

$$\mathbf{u}(\mathbf{m}) = \text{Norm}\left(\mathcal{P}_{\mathbf{m}}\left[-\nabla_{\mathbf{m}}\mathcal{E}(\mathbf{m})\right]\right), \tag{12}$$

where $\mathcal{P}_{\mathbf{m}}$ projects the negative energy gradient onto the valid rotation tangent space and $\text{Norm}(\cdot)$ controls its magnitude. We train the refinement field to approximate this local descent direction:

$$\mathcal{L}_{\text{ref}} = \mathbb{E}_{(\mathbf{m},t)\sim\mathcal{D}_{\text{state}}}\left[\left\|\mathbf{M} \odot \left(v_\theta(\mathbf{m}, t, \mathbf{c}) - \mathbf{u}(\mathbf{m})\right)\right\|_2^2\right], \tag{13}$$

where $\mathcal{D}_{\text{state}}$ denotes the distribution of intermediate refinement states and $\mathbf{M}$ specifies the editable joints. Therefore, from the optimization perspective, the refinement field in ReFM can be understood as a learned, condition-dependent descent field that progressively updates the current motion toward lower-energy states. Unlike existing one-step regression methods that directly predict a final retargeted

motion, ReFM models local corrections over intermediate states. Meanwhile, the resulting refinement flow is not intended to model a probability distribution, as in normalizing flows or flow-matching methods, but rather describes the iterative dynamics induced by repeatedly applying the learned field. During inference, the explicit motion energy further evaluates candidate updates and accepts only those that decrease the energy.

Energy-guided inference. At inference, ReFM iteratively applies the predicted refinement direction to the current motion. Candidate updates with different step sizes are evaluated using Eq. 11, and only an update that decreases the energy is accepted. Refinement terminates when none of the candidate updates yields further energy reduction, and the last accepted state is returned as the final motion. The complete training and inference procedures are introduced in Appendix B.

5 EXPERIMENTS

We evaluate ReFM quantitatively on the Mixamo test dataset and provide additional qualitative results on ScanRet. All Mixamo comparisons use a unified 65-joint evaluation protocol. Detailed experimental settings are provided in Appendix D.

5.1 COMPARISON WITH EXISTING RETARGETING METHODS

Table 1: Quantitative comparison on the Mixamo test dataset under the unified 65-joint evaluation setting. ↓ and ↑ indicate that lower and higher values are better, respectively.

Method	Pen ↓	$\mathrm{Sem}_{\mathrm{sim}}$ ↑	Curv ↓	MSE ↓
Reference motions				
Ground truth	0.141	–	0.654	–
Naive copy	0.140	0.999	0.657	0.051
HumanIK (Autodesk, Inc., 2026)	0.141	0.999	0.657	0.049
Skinned retargeting methods				
STaR (Yang et al., 2025)	0.137	0.998	0.702	0.074
MeshRet (Ye et al., 2024)	0.135	0.977	0.702	0.102
Skin-agnostic retargeting methods				
SAN (Aberman et al., 2020)	0.137	0.955	1.497	0.161
R2ET (Zhang et al., 2023)	0.136	0.998	0.854	0.082
Ours: skin-agnostic refinement				
ReFM-Copy	0.117	0.998	0.772	0.086
ReFM-HumanIK	0.117	0.996	0.772	0.091

Note: Pen and $\mathrm{Sem}_{\mathrm{sim}}$ are the primary metrics for physical plausibility and semantic preservation, respectively. Curv and MSE are auxiliary measures whose limitations are discussed in Appendix E. It is noted that, although the reference motions largely preserve the source-motion semantics by construction, they may violate other constraints such as physical plausibility.

Quantitative comparison. We consider two ReFM initializations: direct motion copying and HumanIK (Autodesk, Inc., 2026). Copying directly applies source joint rotations to the target skeleton and may cause geometric misalignment and self-penetration, whereas HumanIK provides a stronger full-body IK initialization but does not explicitly model target surface geometry. We compare with representative skinned methods, STaR (Yang et al., 2025) and MeshRet (Ye et al., 2024), and source-mesh-agnostic methods, SAN (Aberman et al., 2020) and R2ET (Zhang et al., 2023). All outputs are evaluated using the same 65-joint topology, height-normalized skeletons, target meshes, and metric implementations. STaR outputs are transferred from its native 22-joint topology to the corresponding core joints of the unified topology.

We use Pen to assess self-penetration and $\mathrm{Sem}_{\mathrm{sim}}$ as an encoder-space measure of source-motion consistency. Since $\mathrm{Sem}_{\mathrm{sim}}$ uses the same frozen encoder that defines $\mathcal{E}_{\mathrm{sem}}$, it is not fully independent of ReFM training and should not be interpreted as a direct measure of human semantic judgment. As shown in Table 1, ReFM-Copy reduces Pen from 0.140 to 0.117 ($\sim 16\%$), while achieving $\mathrm{Sem}_{\mathrm{sim}} = 0.998$. It also reduces penetration by approximately 15%, 13%, and 14% relative to STaR, MeshRet, and R2ET, respectively. Starting from HumanIK, ReFM-HumanIK reduces Pen from 0.141 to 0.117 ($\sim 17\%$), with $\mathrm{Sem}_{\mathrm{sim}} = 0.996$. This improvement does not uniformly extend to the auxiliary Curv and MSE metrics, which increase relative to the corresponding initializations, indicating a trade-off between target-specific geometric correction and low-level motion correspondence. We therefore interpret these metrics jointly rather than claiming uniform improvement across all criteria.

Overall, ReFM reduces self-penetration from both copied-motion and HumanIK initializations while retaining high encoder-space source-motion consistency. The pretrained encoder additionally provides a cross-character motion-consistency measure complementary to conventional geometric and kinematic metrics. Sec. 5.2 further analyzes its performance and relation to human judgments.

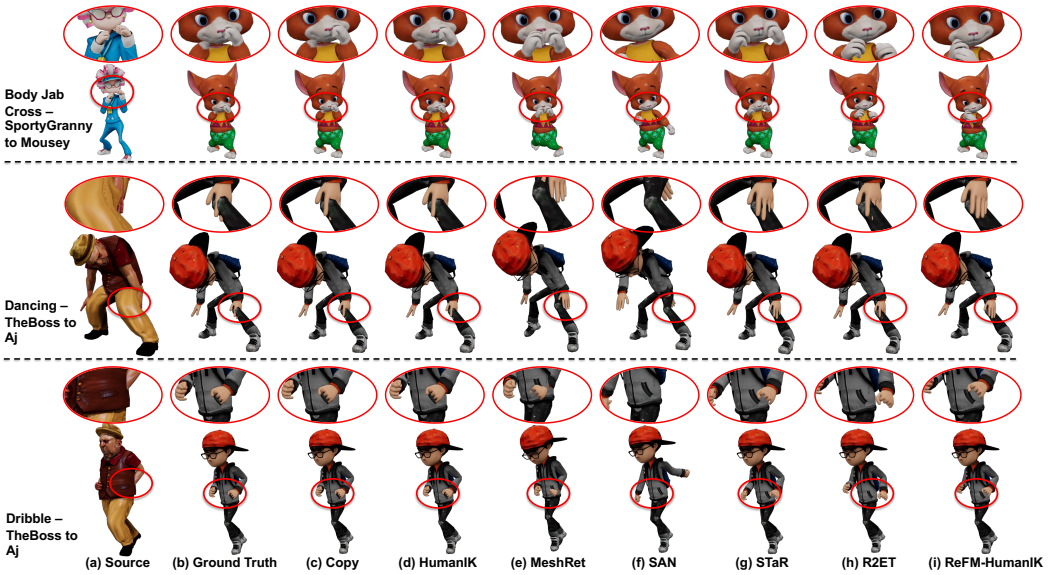


Figure 4: Qualitative comparison on the Mixamo dataset. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. Red circles highlight representative regions with self-penetration or geometry-sensitive interactions.

Qualitative comparison. Figure 4 further demonstrates the advantage of ReFM under imperfect retargeting supervision. Notably, the recorded ground-truth motions are not necessarily physically valid and can themselves contain visible self-penetration, as highlighted by the red circles, consistent with prior observations on Mixamo (Yang et al., 2025). Therefore, paired target motions should not be regarded as physically clean regression targets. ReFM instead preserves source-motion semantics while explicitly refining target-specific geometric artifacts, consistently reducing penetration without sacrificing characteristic poses. For example, in *Body Jab Cross*, several baselines exhibit interference between the hands and upper body, whereas ReFM-HumanIK maintains the characteristic defensive hand configuration with cleaner spatial separation. Similarly, in *Dancing*, ReFM reduces the arm–leg intersection while largely preserving the source-derived upper-body configuration. These examples support that rather than regressing toward an imperfect target motion, ReFM applies minimal target-aware corrections to improve physical plausibility while maintaining semantic intent. Additional qualitative results and user studies are provided in Appendix F.

5.2 Effectiveness of Semantic Encoder

Beyond comparing the retargeting results with baseline methods, we further investigate whether the proposed semantic encoder (SE) can effectively distinguish different motions and whether its semantic evaluation is consistent with human perception.

Can the pretrained SE distinguish different motions? As detailed in Appendix A, we independently evaluate the frozen encoder on held-out motion groups that are never observed during semantic pretraining. Cross-character positive pairs, consisting of the same motion performed by different characters, achieve a cosine similarity of 0.9969 ± 0.0202, whereas negative pairs consisting of different motions obtain only 0.0859 ± 0.2911. The learned embedding space further exhibits compact intra-motion clusters and clear inter-motion separation, demonstrating that the pretrained SE captures motion-dependent semantics while remaining largely invariant to character identity.

Is the pretrained SE aligned with human perception? We further examine how $\mathrm{Sem}_{\mathrm{sim}}$ relates to perceived semantic preservation through the user study in Appendix F.6. Although naive copy and HumanIK achieve near-ceiling $\mathrm{Sem}_{\mathrm{sim}}$ values, participants rank both below ReFM-HumanIK and R2ET for semantic preservation. Because participants judge rendered motions, visible physical artifacts may affect whether they perceive the source action and its characteristic gestures as preserved, even when the encoder assigns high similarity. The user study provides complementary evidence, but we do not claim a clip-level correlation between its rankings and $\mathrm{Sem}_{\mathrm{sim}}$. We therefore interpret $\mathrm{Sem}_{\mathrm{sim}}$ as a motion-space measure rather than a measure of human semantic judgment.

5.3 Ablation Study

We ablate the $\mathrm{SO}(3)$ canonicalizer, semantic component, and progressive refinement flow under both copy and HumanIK initialization. All variants use the same test set and otherwise share the same settings.

For *w/o Canon.*, the framework is retrained without global-heading canonicalization. For *w/o Sem.*, we remove both the semantic representation supplied to the refinement model and the semantic energy $\mathcal{E}_{\mathrm{sem}}$, retaining the frozen encoder only for evaluation; thus, this variant evaluates the semantic component as a whole rather than isolating either mechanism. For *w/o Flow*, we replace iterative refinement with a one-step regressor using the same backbone, training data, initialization, and motion energy, enabling a controlled comparison with energy-trained one-step prediction.

Table 2: Ablation of the canonicalizer, semantic component, and refinement flow under copy and HumanIK initialization. The *w/o Flow* variant uses one-step prediction. Lower is better for Pen, Curv, and MSE; higher is better for $\mathrm{Sem}_{\mathrm{sim}}$.

		Component			Metric			
Initialization	Method	Canon.	Sem.	Flow	$\mathrm{Pen}\downarrow$	$\mathrm{Sem}_{\mathrm{sim}}\uparrow$	$\mathrm{Curv}\downarrow$	$\mathrm{MSE}\downarrow$
Copy	w/o Canon.	✗	✓	✓	0.130	0.997	0.718	0.074
	w/o Sem. Enc.	✓	✗	✓	0.129	0.992	0.729	0.074
	w/o Flow	✓	✓	✗	0.137	0.998	0.689	0.058
	ReFM	✓	✓	✓	0.117	0.998	0.772	0.086
HumanIK	w/o Canon.	✗	✓	✓	0.131	0.997	0.730	0.073
	w/o Sem. Enc.	✓	✗	✓	0.129	0.996	0.737	0.080
	w/o Flow	✓	✓	✗	0.136	0.998	0.698	0.049
	ReFM	✓	✓	✓	0.117	0.996	0.772	0.091

As shown in Table 2, the components contribute differently to the refinement objective. Removing canonicalization increases Pen from 0.117 to 0.130 under copy initialization and to 0.131 under HumanIK, suggesting that removing redundant global-heading variation facilitates geometric correction. Removing the semantic component reduces $\mathrm{Sem}_{\mathrm{sim}}$ from 0.998 to 0.992 for copy initialization and increases penetration for both initializations, while yielding lower Curv and MSE. Since this ablation jointly removes semantic conditioning and $\mathcal{E}_{\mathrm{sem}}$, these effects should be attributed to the combined semantic component rather than either mechanism individually. Replacing progressive refinement with one-step prediction yields lower Curv and MSE but higher Pen: 0.137 versus 0.117 for copy and 0.136 versus 0.117 for HumanIK. Thus, under the same backbone and energy formulation, progressive refinement primarily improves target-specific geometric correction rather than uniformly improving all motion-similarity metrics.

6 Conclusion and Limitations

We proposed ReFM, a source-mesh-agnostic motion retargeting framework that combines $\mathrm{SO}(3)$ canonicalization, cross-character semantic representation learning, and energy-guided progressive refinement. Rather than regressing a final motion in a single pass, ReFM progressively improves different retargeting initializations under semantic, physical, and temporal constraints. Experiments show that ReFM consistently reduces self-penetration while preserving motion semantics, and can further refine motions initialized by the industry-standard HumanIK system.

Several limitations suggest directions for future work. First, the current semantic encoder is pretrained from limited cross-character correspondences whose motion quality can be imperfect, making the learned representation relatively insensitive to geometric artifacts (as introduced in Sec. 5.2). Moreover, although the evaluation test motions do not overlap with semantic-encoder pretraining, $\mathrm{Sem}_{\mathrm{sim}}$ is computed using the same frozen encoder that defines $\mathcal{E}_{\mathrm{sem}}$ during ReFM training. Thus, this metric is not fully isolated from the training objective and should be interpreted jointly with human evaluation and physical metrics. A promising direction is to pretrain the semantic encoder on substantially larger and more diverse motion corpora, yielding a general motion-semantic prior that can remain fixed across downstream retargeting tasks and character sets, while also serving as a more reliable and task-independent metric for semantic preservation. Second, our current canonicalizer considers only the yaw subgroup of $\mathrm{SO}(3)$. The proposed canonicalization principle is more general with a broader application space: for tasks involving additional ground, contact, or environment constraints, the task-relevant transformation group can be redefined accordingly, allowing the same framework to preserve constraint-relevant components while canonicalizing only nuisance transformations.

AI USE STATEMENT

Generative AI tools were used during the development of this work to provide assistance on paper polishing, code debugging and project organization. All AI-assisted content was independently reviewed and verified by the authors. Experimental results reported in this paper were obtained from our implemented methods and evaluation pipelines rather than generated by AI tools. The authors take full responsibility for the methodology, experiments, claims, and final content of the paper.

REPRODUCIBILITY STATEMENT

We will open-source the complete ReFM implementation on GitHub upon publication. To facilitate reproducibility, the appendix provides detailed implementation information, including model architectures, training and inference configurations, dataset construction and preprocessing, hyperparameter settings, and the definitions and computation procedures of all evaluation metrics.

REFERENCES

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (ToG)*, 39(4):62–1, 2020.

Adobe. Mixamo. `https://www.mixamo.com/`.

Autodesk, Inc. Autodesk maya. `https://www.autodesk.com/products/maya/overview`, 2026. 3D animation, modeling, simulation, and rendering software.

Ko Ayusawa and Eiichi Yoshida. Motion retargeting for humanoid robots based on simultaneous morphing parameter identification and motion optimization. *IEEE Transactions on Robotics*, 33 (6):1343–1357, 2017.

Jean Basset, Stefanie Wuhrer, Edmond Boyer, and Franck Multon. Contact preserving shape transfer: Retargeting motion from one shape to another. *Computers & Graphics*, 89:11–23, 2020.

Kang Chen, Yupan Wang, Song-Hai Zhang, Sen-Zhe Xu, Weidong Zhang, and Shi-Min Hu. Mocapsolver: A neural solver for optical motion capture data. *ACM Transactions on Graphics (TOG)*, 40 (4):1–11, 2021.

Théo Cheynel, Thomas Rossi, Baptiste Bellot-Gurlet, Damien Rohmer, and Marie-Paule Cani. Reconform: Real-time contact-aware motion retargeting for more diverse character morphologies. In *Computer Graphics Forum*, volume 44, pp. e70028. Wiley Online Library, 2025a.

Théo Cheynel, Thomas Rossi, Omar El Khalifi, Oscar Fossey, Damien Rohmer, and Marie-Paule Cani. Mirrored-anims: Motion inversion for rig-space retargeting to obtain a reliable enlarged dataset of character animations. In *Proceedings of the 2025 18th ACM SIGGRAPH Conference on Motion, Interaction, and Games*, pp. 1–12, 2025b.

Kwang-Jin Choi and Hyeong-Seok Ko. Online motion retargetting. *The Journal of Visualization and Computer Animation*, 11(5):223–235, 2000.

Soojin Choi, Seokhyeon Hong, Chaelin Kim, Junghyun Nam, Junhyuk Jeon, and Junyong Noh. Skinned motion retargeting with spatially adaptive interaction guidance. *ACM Transactions on Graphics (TOG)*, 45(4):1–17, 2026.

Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational visual media*, 7(2):187–199, 2021.

Lei Hu, Zihao Zhang, Chongyang Zhong, Boyuan Jiang, and Shihong Xia. Pose-aware attention network for flexible motion retargeting by body part. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):4792–4808, 2024.

Taeil Jin, Meekyoung Kim, and Sung-Hee Lee. Aura mesh: Motion retargeting to preserve the spatial relationships between skinned characters. In *Computer Graphics Forum*, volume 37, pp. 311–320. Wiley Online Library, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Sunmin Lee, Taeho Kang, Jungnam Park, Jehee Lee, and Jungdam Won. Same: Skeleton-agnostic motion embedding for character animation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Jongin Lim, Hyung Jin Chang, and Jin Young Choi. Pmnet: Learning of disentangled pose and movement for unsupervised motion retargeting. In *30th British Machine Vision Conference (BMVC 2019)*. British Machine Vision Association, BMVA, 2019.

Zhiguang Liu, Antonio Mucherino, Ludovic Hoyet, and Franck Multon. Surface based motion retargeting by preserving spatial relationship. In *Proceedings of the 11th ACM SIGGRAPH Conference on Motion, Interaction and Games*, pp. 1–11, 2018.

Lucas Mourot, Ludovic Hoyet, François Le Clerc, and Pierre Hellier. Humot: human motion representation using topology-agnostic transformers for character animation retargeting. *arXiv preprint arXiv:2305.18897*, 2023.

Daniele Reda, Jungdam Won, Yuting Ye, Michiel Van De Panne, and Alexander Winkler. Physics-based motion retargeting from sparse inputs. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 6(3):1–19, 2023.

Seyoon Tak and Hyeong-Seok Ko. A physically-based motion retargeting filter. *ACM Transactions on Graphics (ToG)*, 24(1):98–117, 2005.

Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8639–8648, 2018.

Ruben Villegas, Duygu Ceylan, Aaron Hertzmann, Jimei Yang, and Jun Saito. Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9720–9729, 2021.

Xiaohang Yang, Qing Wang, Jiahao Yang, Gregory Slabaugh, and Shanxin Yuan. Star: Seamless spatial-temporal aware motion retargeting with penetration and consistency constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12947–12955, 2025.

Zijie Ye, Jia-Wei Liu, Jia Jia, Shikun Sun, and Mike Zheng Shou. Skinned motion retargeting with dense geometric interaction perception. *Advances in Neural Information Processing Systems*, 37: 125907–125934, 2024.

Haodong Zhang, Zhike Chen, Haocheng Xu, Lei Hao, Xiaofei Wu, Songcen Xu, Zhensong Zhang, Yue Wang, and Rong Xiong. Semantics-aware motion retargeting with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2155–2164, 2024a.

Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Jue Wang, and Zhigang Tu. 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.

Jiaxu Zhang, Zhigang Tu, Junwu Weng, Junsong Yuan, and Bo Du. A modular neural motion retargeting system decoupling skeleton and shape perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6889–6904, 2024b.

Wentao Zhu, Zhuoqian Yang, Ziang Di, Wayne Wu, Yizhou Wang, and Chen Change Loy. Mocanet: Motion retargeting in-the-wild via canonicalization networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 3617–3625, 2022.

Appendix

A Evaluation of the Cross-Character Semantic Encoder

To characterize the pretrained cross-character semantic encoder, we evaluate it on motion groups held out from encoder pretraining. Here, a *motion identity* denotes a particular underlying motion performed by multiple characters, whereas an *action category* denotes a broader label that may contain distinct executions. For example, *Dancing (1)* and *Dancing (2)* share the category *Dancing* but represent different motion identities. We combine clips from Mixamo and ScanRet, retain motion groups containing multiple characters, and partition the data by motion identity so that no validation identity appears during pretraining.

In both contrastive pretraining and held-out evaluation, positive pairs contain the same motion identity performed by different characters. Negative pairs contain different motion identities performed by either the same or different characters, *including identities within the same action category*. Thus, *Dancing (1)* and *Dancing (2)* are treated as negatives despite their shared action label. This construction evaluates whether the encoder distinguishes different executions within a category as well as different categories, while matching the same motion across characters. We report cosine similarity for these pairs and examine the resulting motion-group embedding structure. Because separately performed but semantically equivalent clips are not labeled as positives in this protocol, the results can serve as the evidence of cross-character matching and fine-grained motion discrimination, rather than a direct test of broader semantic equivalence.

Table 3: Evaluation of the pretrained cross-character semantic encoder on held-out motion groups. Positive pairs contain the same motion performed by different characters, while negative pairs contain different motions. The standard deviations are provided after $\pm$.

Metric	**Positive**	**Negative**
Cosine similarity	0.9969 ± 0.0202	0.0859 ± 0.2911
Motion-group embedding structure		
Mean intra-class distance $\downarrow$	0.0441	
Mean inter-class distance $\uparrow$	0.8768	
Separation ratio $\uparrow$	0.9497	

As shown in Table 3, the learned representation exhibits strong cross-character semantic discrimination. For example, motions with the same identity but performed by different characters achieve a cosine similarity of 0.9969 ± 0.0202, whereas motions with different identities have substantially lower similarity. The corresponding motion-group statistics further show compact intra-motion clusters and clear separation across different motions. These results demonstrate that the pretrained encoder learns a representation that is highly invariant to character identity while remaining discriminative across held-out motion identities. Since motion identity serves as the semantic supervision in our contrastive construction, this cross-character separation supports its use as an operational representation of motion semantics for both refinement guidance and semantic-consistency evaluation.

Clarification on semantic pretraining and data leakage. We emphasize that pretraining the semantic encoder on a broad collection of motion sequences should not be interpreted as exposing ReFM to ground-truth retargeting targets. Motion retargeting is inherently ill-posed, and existing datasets, particularly Mixamo, do not provide a unique or physically reliable paired target for a given source–target character pair. Indeed, as also observed in our qualitative results and prior work (Zhang et al., 2023; Yang et al., 2025), the recorded "ground-truth" motions can themselves contain substantial self-penetration and other geometric artifacts. ReFM therefore never treats these target motions as supervision to be reproduced; its objective is instead to preserve the semantic intent of the source motion while refining an initialization toward improved target-specific physical plausibility, and its output may consequently be preferable to the recorded target motion under these criteria. The semantic encoder serves only as a pretrained representation of motion meaning rather than as a predictor of any target motion. Importantly, its semantic discrimination ability is evaluated independently on motion groups that are held out from encoder pretraining, as reported in Table 3. These held-out results show strong separation between same-motion cross-character positives and different-motion negatives, including fine-grained negatives with closely related action categories, such as different kicking motions (e.g., *Kick-1* and *Kick-2*) that share the same coarse action category but differ in their detailed gestures. Thus, the observed semantic discrimination

cannot be explained by memorization of the evaluated motion identities, while ReFM itself receives no paired target-motion supervision from the evaluation set.

B Implementation Details

This section provides the implementation details directly related to the proposed refinement formulation. Additional engineering details, including data loading, optimization, numerical stabilization, and checkpoint management, are provided in our open-source project.

Quaternion-space refinement. ReFM represents each joint rotation using a unit quaternion. Since the motion energy is differentiated in the ambient Euclidean space, we project its gradient onto the tangent space of the unit quaternion manifold before constructing the refinement target. For a quaternion $\mathbf{q} \in \mathbb{S}^3$ and gradient $\mathbf{g} \in \mathbb{R}^4$, the projection is

$$\mathcal{P}_{\mathbf{q}}[\mathbf{g}] = \mathbf{g} - \langle \mathbf{g}, \mathbf{q} \rangle \mathbf{q}. \tag{14}$$

This removes the radial component of the Euclidean gradient and satisfies $\langle \mathcal{P}_{\mathbf{q}}[\mathbf{g}], \mathbf{q} \rangle = 0$. For a motion sequence, the projection is independently applied to every joint quaternion at every frame.

Following the limb-focused geometric correction setting adopted in prior retargeting methods (Yang et al., 2025), we further use a binary refinement mask $\mathbf{M}$ to restrict the editable joints. Our final configuration uses the `core22_limbs` mask, which allows ReFM to modify the major arm and leg chains while keeping the remaining joints fixed to the initialization. After each update, the motion is projected back to the feasible motion space by

$$\Pi_{\mathcal{M}} \left(\mathbf{m}; \mathbf{m}_t^{(0)} \right) = \mathbf{M} \odot \mathrm{QuatNorm}(\mathbf{m}) + (\mathbf{1} - \mathbf{M}) \odot \mathbf{m}_t^{(0)}, \tag{15}$$

where $\mathrm{QuatNorm}(\cdot)$ performs per-joint quaternion normalization.

Gradient-matching training. Because paired progressive-refinement trajectories are unavailable, we construct local supervision directly from the motion energy. Given an intermediate state $\mathbf{m}^{(i)}$, its target refinement direction is defined as

$$\mathbf{u}^{(i)} = \mathrm{Norm} \left(\mathcal{P}_{\mathbf{m}^{(i)}} \left[-\nabla_{\mathbf{m}^{(i)}} \mathcal{E} \left(\mathbf{m}^{(i)}; \bar{\mathbf{m}}_s, \mathbf{m}_t^{(0)}, \mathbf{s}_t, \mathbf{x}_t \right) \right] \right), \tag{16}$$

where $\mathrm{Norm}(\cdot)$ normalizes the direction magnitude on a per-sample basis. The refinement field v_θ is then trained to predict this local energy-descent direction:

$$\mathcal{L}_{\mathrm{ref}} = \frac{1}{S} \sum_{i=0}^{S-1} \left\| \mathbf{M} \odot \left(v_\theta(\mathbf{m}^{(i)}, \tau_i, \mathbf{c}) - \mathbf{u}^{(i)} \right) \right\|_2^2, \qquad \tau_i = \frac{i}{S}. \tag{17}$$

The intermediate states are generated through a short energy-descent trajectory,

$$\mathbf{m}^{(i+1)} = \Pi_{\mathcal{M}} \left(\mathbf{m}^{(i)} + \eta \mathbf{u}^{(i)}; \mathbf{m}_t^{(0)} \right). \tag{18}$$

We use $S = 4$ refinement states and $\eta = 0.05$. This formulation trains the network to approximate locally useful refinement directions rather than directly regressing a unique final target motion.

Model configuration. The semantic encoder contains four graph Transformer layers with hidden dimension 128 and produces a 128-dimensional clip-level embedding. It is pretrained with supervised contrastive learning using temperature $\tau = 0.1$, optimized by Adam with learning rate 10^{-3} annealed to 10^{-6} by a cosine schedule over 50 epochs at batch size 16; we retain the checkpoint with the lowest validation loss and keep the encoder frozen during ReFM training. The refinement field consists of six spatio-temporal graph Transformer blocks with hidden dimension 320, four attention

heads, and MLP dimension 768. Moreover, the target-mesh condition is extracted by a frozen shape encoder (Guo et al., 2021).

The refinement field is trained for 40 epochs with Adam at a constant learning rate of 10^{-3}, without warmup or weight decay, and with gradient-norm clipping at 1.0. We use data-parallel training over 8 NVIDIA T4 GPUs with a per-GPU batch of 4, giving an effective batch size of 32. On Mixamo this amounts to 68 iterations per epoch over 2,175 training clips, and takes roughly 14 GPU-hours per device ($\sim$21 minutes per epoch); ScanRet uses identical optimization settings.

Energy Terms. The penetration energy $\mathcal{E}_{\text{pen}}$ and initialization-preserving energy $\mathcal{E}_{\text{init}}$ build upon the penetration and minor-modification objectives of STaR (Yang et al., 2025). We additionally introduce semantic, temporal-smoothness, and trajectory-curvature energies. Let $\mathbf{q}_{t,j}$ denote the quaternion of editable joint j at frame t, and let $\mathbf{p}_{t,j}(\mathbf{m})$ denote its global position obtained through forward kinematics. The temporal-smoothness energy is

$$\mathcal{E}_{\text{smo}}(\mathbf{m}) = \frac{1}{4(T-2)|\mathcal{J}_e|} \sum_{t=2}^{T-1} \sum_{j \in \mathcal{J}_e} \|\mathbf{q}_{t+1,j} - 2\mathbf{q}_{t,j} + \mathbf{q}_{t-1,j}\|_2^2 , \tag{19}$$

where $\mathcal{J}_e$ denotes the set of editable joints. The temporal sign continuity is enforced before computing the smoothness energy. To prevent refinement-induced trajectory distortion, we further define

$$\begin{aligned} \kappa(\mathbf{m}) &= \frac{1}{(T-2)J} \sum_{t=2}^{T-1} \sum_{j=1}^{J} \left\| \frac{\mathbf{p}_{t+1,j}(\mathbf{m}) - 2\mathbf{p}_{t,j}(\mathbf{m}) + \mathbf{p}_{t-1,j}(\mathbf{m})}{\Delta t^2} \right\|_2 , \\ \mathcal{E}_{\text{curv}}\big(\mathbf{m}, \mathbf{m}_t^{(0)}\big) &= \Big[\kappa(\mathbf{m}) - \kappa\big(\mathbf{m}_t^{(0)}\big) \Big]_+ , \end{aligned} \tag{20}$$

where $\Delta t = 1/60\,\text{s}$ and $[\cdot]_+$ denotes the ReLU operation. This one-sided penalty suppresses only curvature introduced beyond that of the initial motion. Finally, semantic preservation is measured by

$$\mathcal{E}_{\text{sem}}(\mathbf{m}, \bar{\mathbf{m}}_s) = 1 - \widehat{\boldsymbol{\epsilon}}(\mathbf{m}, \mathbf{s}_t)^\top \widehat{\boldsymbol{\epsilon}}(\bar{\mathbf{m}}_s, \mathbf{s}_s), \tag{21}$$

where $\widehat{\boldsymbol{\epsilon}}(\cdot)$ denotes the normalized output of the frozen semantic encoder after yaw canonicalization. We use $\lambda_{\text{sem}} = 0.5$, $\lambda_{\text{init}} = 0.1$, $\lambda_{\text{pen}} = 2.0$, $\lambda_{\text{smo}} = 0.1$, and $\lambda_{\text{curv}} = 0.1$.

Energy-monitored inference. ReFM can start from either a directly copied motion or a HumanIK-retargeted motion. At each refinement step, the network predicts a direction $\mathbf{v}^{(k)}$, and we evaluate several candidate update magnitudes using the complete motion energy:

$$\mathbf{m}_a^{(k)} = \Pi_{\mathcal{M}} \left(\mathbf{m}^{(k)} + a\mathbf{v}^{(k)}; \mathbf{m}_t^{(0)} \right), \qquad a \in \mathcal{A}. \tag{22}$$

The accepted step is

$$a^\star = \arg\min_{a \in \mathcal{A}} \mathcal{E} \left(\mathbf{m}_a^{(k)}; \bar{\mathbf{m}}_s, \mathbf{m}_t^{(0)}, \mathbf{s}_t, \mathbf{x}_t \right). \tag{23}$$

We use $\mathcal{A} = \{0, 0.02, 0.05, 0.1\}$ and perform at most $K = 4$ refinement steps. Since $0 \in \mathcal{A}$, the current motion is always a valid candidate. Therefore, if none of the nonzero updates decreases the total energy, $a^\star = 0$ is selected and refinement terminates. This design allows ReFM to adapt both the magnitude and number of refinements according to the quality of the current motion, while guaranteeing a non-increasing sequence of accepted motion energies.

C Sensitivity to the Semantic Energy Weight

The semantic energy $\mathcal{E}_{\text{sem}}$ encourages semantic consistency during progressive refinement. We examine the sensitivity of ReFM-HumanIK to its weight by varying $\lambda_{\text{sem}} \in \{0.1, 0.2, 0.5\}$ while keeping all other energy weights fixed following Yang et al. (2025). We build up the comparison experiments to isolate the effect of semantic guidance on the balance between semantic preservation, physical plausibility, and temporal consistency.

Table 4: Sensitivity of ReFM-HumanIK to λ_{sem} on the test dataset. All other energy weights and training configurations are fixed. Lower Pen, Curv, and MSE are better; higher Sem_{sim} is better.

λ_{sem}	Pen $\downarrow$	Curv $\downarrow$	MSE $\downarrow$	Sem_{sim} $\uparrow$
0.1	0.118	0.788	0.093	0.996
0.2	0.119	0.783	0.095	0.997
0.5	0.117	0.772	0.091	0.996

Table 4 shows that ReFM is relatively insensitive to the semantic-energy weight over the tested range. In particular, Sem_{sim} remains consistently high, varying only from 0.996 to 0.997. Meanwhile, the physical and temporal metrics remain stable as λ_{sem} increases. This experiment shows that ReFM is relatively insensitive to the semantic energy weight, i.e., a modest setting can preserve nearly identical results.

D Experimental Settings

Datasets. We conduct experiments on the Mixamo (Adobe) and ScanRet (Ye et al., 2024) datasets, which jointly contain 113 characters and approximately $11{,}973$ motion sequences. All characters are converted to a shared 65-joint topology comprising 22 major body joints, 40 finger joints, and 3 extra leaf/end-effector joints. Each training sequence is randomly cropped or padded to 60 frames. For semantic-encoder pretraining, motion identities are partitioned into training and validation sets such that the same motion does not appear in both splits. For motion retargeting, each source motion is paired with a target character of different skeletal proportions. For the Mixamo dataset, we evaluate on 222 ground-truth-paired retargeting cases (11 characters, 91 distinct motion clips) drawn from a split that is disjoint from training at the (character, motion)-instance level: none of the test (character, motion) instances occur among the training instances, and 3 of the 11 characters (Kaya, Ortiz, XBot) are never animated during training. Moreover, the test dataset is not used in semantic encoder pretraining. Additionally, we use ScanRet to examine the retargeting ability on real-human motions. The split of the ScanRet dataset follows Ye et al. (2024).

Evaluation Metrics. We evaluate retargeting quality using semantic similarity (Sem_{sim}), penetration rate (Pen), trajectory curvature (Curv), and joint-position error to the ground truth (MSE). Sem_{sim} measures the cosine similarity between the frozen semantic embeddings of the source and retargeted motions, quantifying semantic consistency under the learned representation introduced in Sec. 4.2. Since the same frozen encoder is also used to construct the semantic energy during ReFM refinement, Sem_{sim} should be interpreted primarily as measuring whether semantic consistency is retained while improving other aspects of the motion, rather than as an entirely independent evaluation criterion. Pen measures the proportion of target-mesh vertices that penetrate other body parts and evaluates the physical plausibility of the retargeted motion. The validity of the pretrained semantic representation is evaluated independently in Appendix A, while detailed definitions and computation procedures of the evaluation metrics are provided in Appendix E.

E Introduction of Evaluation Metrics

We evaluate all methods under a unified 65-joint representation. Each predicted motion is applied to the same target skeleton and mesh, and all metrics are computed after forward kinematics or linear blend skinning as required. We report semantic similarity, penetration rate, MSE, and trajectory curvature. Among them, semantic similarity and penetration rate serve as our primary metrics because they directly evaluate semantic preservation and physical plausibility without assuming a unique ground-truth retargeting. MSE and curvature are included as auxiliary metrics for compatibility with prior work, but should be interpreted with their limitations discussed below.

Semantic Similarity. We measure semantic preservation using the frozen cross-character semantic encoder introduced in Sec. 4.2. Given a source motion $\mathbf{m}_s$ on source skeleton $\mathbf{s}_s$ and a retargeted motion $\hat{\mathbf{m}}_t$ on target skeleton $\mathbf{s}_t$, we first canonicalize both motions and extract their normalized embeddings:

$$\hat{\boldsymbol{\epsilon}}_s = \frac{E_{\text{sem}}\left(\bar{\mathbf{m}}_s, \mathbf{s}_s\right)}{\left\|E_{\text{sem}}\left(\bar{\mathbf{m}}_s, \mathbf{s}_s\right)\right\|_2}, \qquad \hat{\boldsymbol{\epsilon}}_t = \frac{E_{\text{sem}}\left(\bar{\hat{\mathbf{m}}}_t, \mathbf{s}_t\right)}{\left\|E_{\text{sem}}\left(\bar{\hat{\mathbf{m}}}_t, \mathbf{s}_t\right)\right\|_2}. \tag{24}$$

The semantic similarity is defined as their cosine similarity:

$$\mathrm{Sem}_{\mathrm{Sim}}\left(\mathbf{m}_s, \hat{\mathbf{m}}_t\right) = \hat{\boldsymbol{\epsilon}}_s^{\top} \hat{\boldsymbol{\epsilon}}_t. \tag{25}$$

Since the embeddings are L_2-normalized, $\mathrm{Sem}_{\mathrm{Sim}} \in [-1, 1]$, where a larger value indicates stronger preservation of the source action semantics. Unlike position-based errors, this metric compares motions in a character-invariant semantic space and does not require a unique target motion as a reference.

Penetration Rate. Following STaR (Yang et al., 2025), we evaluate geometric plausibility by measuring the proportion of target-mesh vertices that penetrate other body regions. The target rest-pose mesh is first deformed by the complete 65-joint motion using linear blend skinning. Let $\mathbf{v}_i^t$ and $\mathbf{n}_i^t$ denote the position and outward normal of vertex i at frame t, respectively.

We divide the target mesh into query groups $\mathcal{G}$ corresponding to interaction-prone body regions, including the arms, hands, legs, and head. For each group $g \in \mathcal{G}$, let $\mathcal{Q}_g$ denote its query vertices and $\mathcal{R}_g$ denote the reference vertices belonging to the remaining relevant body regions. For each query vertex $i \in \mathcal{Q}_g$, its nearest reference vertex is

$$r_{t,g}^{\star}(i) = \underset{r \in \mathcal{R}_g}{\arg\min} \left\| \mathbf{v}_i^t - \mathbf{v}_r^t \right\|_2. \tag{26}$$

The signed penetration value is estimated by projecting the query-to-reference displacement onto the outward normal of the reference surface:

$$\phi_{t,g}(i) = \left(\mathbf{v}_{r_{t,g}^{\star}(i)}^t - \mathbf{v}_i^t \right)^{\top} \mathbf{n}_{r_{t,g}^{\star}(i)}^t. \tag{27}$$

Under this convention, $\phi_{t,g}(i) > 0$ indicates that the query vertex lies behind the local outward-facing tangent plane and is classified as penetrating. The sequence-level penetration rate is

$$\mathrm{Pen}\left(\hat{\mathbf{m}}_t\right) = \frac{1}{T \sum_{g \in \mathcal{G}} |\mathcal{Q}_g|} \sum_{t=1}^{T} \sum_{g \in \mathcal{G}} \sum_{i \in \mathcal{Q}_g} \mathbb{I}\left[\phi_{t,g}(i) > 0\right], \tag{28}$$

where $\mathbb{I}[\cdot]$ is the indicator function. A lower penetration rate indicates better physical plausibility. Because this metric directly counts geometric violations on the posed target mesh, it provides a task-aligned evaluation independent of the quality of the available ground-truth motion.

Trajectory Curvature. Following STaR (Yang et al., 2025), temporal smoothness is evaluated through the mean acceleration magnitude of global joint trajectories. Let $\mathbf{p}_{t,j}$ denote the forward-kinematics position of joint j at frame t, and let f be the motion sampling frequency, which is 60 frames per second in our evaluation. The discrete velocity and acceleration are computed as

$$\mathbf{v}_{t,j} = f\left(\mathbf{p}_{t+1,j} - \mathbf{p}_{t,j}\right), \qquad \mathbf{a}_{t,j} = f\left(\mathbf{v}_{t+1,j} - \mathbf{v}_{t,j}\right) = f^2\left(\mathbf{p}_{t+2,j} - 2\mathbf{p}_{t+1,j} + \mathbf{p}_{t,j}\right). \tag{29}$$

The reported curvature is then

$$\mathrm{Curv} = \frac{1}{(T-2)J} \sum_{t=1}^{T-2} \sum_{j=1}^{J} \left\| \mathbf{a}_{t,j} \right\|_2, \qquad J = 65. \tag{30}$$

A smaller value generally indicates fewer abrupt trajectory changes. Nevertheless, this quantity is not geometric curvature in the strict differential-geometric sense because it is neither normalized by trajectory speed nor invariant to temporal reparameterization. Fast but valid motions naturally exhibit larger acceleration magnitudes than slow motions, while an excessively static motion may obtain an artificially low value. It is also sensitive to the frame rate and spatial scale used during evaluation. Curvature is therefore reported only as an auxiliary indicator of temporal jitter and should be interpreted jointly with semantic similarity and qualitative motion speed.

Mean Squared Error. For comparison with prior motion-retargeting methods, we additionally report the height-normalized global joint-position error conventionally denoted as MSE. Let $\hat{\mathbf{p}}_{t,j} \in \mathbb{R}^3$ and $\mathbf{p}_{t,j}^{\mathrm{gt}} \in \mathbb{R}^3$ denote the predicted and ground-truth global positions of joint j at frame t, obtained through forward kinematics on the target skeleton. The metric is computed over all $J = 65$ joints as

$$\mathrm{MSE} = \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \frac{\left\| \hat{\mathbf{p}}_{t,j} - \mathbf{p}_{t,j}^{\mathrm{gt}} \right\|_2}{h_t}, \qquad J = 65, \tag{31}$$

where h_t is the target-character height computed from its rest-pose kinematic chains. We retain the term MSE to remain consistent with prior literature and the benchmark implementation, although the implemented quantity is an averaged height-normalized Euclidean joint error.

A lower MSE indicates closer agreement with the recorded target performance. However, it is not a definitive measure of retargeting quality. Motion retargeting generally admits multiple semantically and physically valid solutions, whereas MSE assumes that the recorded target motion is the unique correct output. Moreover, the available ground-truth motions may themselves contain penetration and contact errors, as also discovered by Cheynel et al. (2025a;b); Yang et al. (2025); Zhang et al. (2024a). Consequently, a method can achieve low MSE by closely reproducing an imperfect reference without necessarily producing the most plausible retargeted motion. We therefore treat MSE as an auxiliary correspondence metric rather than a primary quality metric.

Dataset-Level Aggregation. For a test dataset of N aligned source–target pairs, each metric is first computed independently for every motion pair and then averaged, with the same pair identities used for every compared method:

$$\overline{\mathcal{M}} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{M}^{(n)}, \qquad \mathcal{M} \in \{\mathrm{Sem}_{\mathrm{sim}}, \mathrm{Pen}, \mathrm{MSE}, \mathrm{Curv}\}\,. \tag{32}$$

Higher semantic similarity and lower penetration rate indicate better performance under our primary evaluation. MSE and curvature provide complementary diagnostics but are not used in isolation to determine the overall quality of a retargeted motion.

F Additional Results

F.1 Frame-Level Qualitative Comparisons

We provide additional frame-level qualitative comparisons to complement the results presented in the main paper. We select representative timesteps from the *Body Jab Cross*, *Dribble*, and *Dancing* sequences, where differences in character geometry can induce local self-penetration after motion transfer. For each example, we report both the complete character rendering and a separate zoomed-in visualization of the penetration-sensitive region. The full-frame results facilitate comparison of the overall pose and motion semantics, while the enlarged views make subtle geometric artifacts more directly observable.

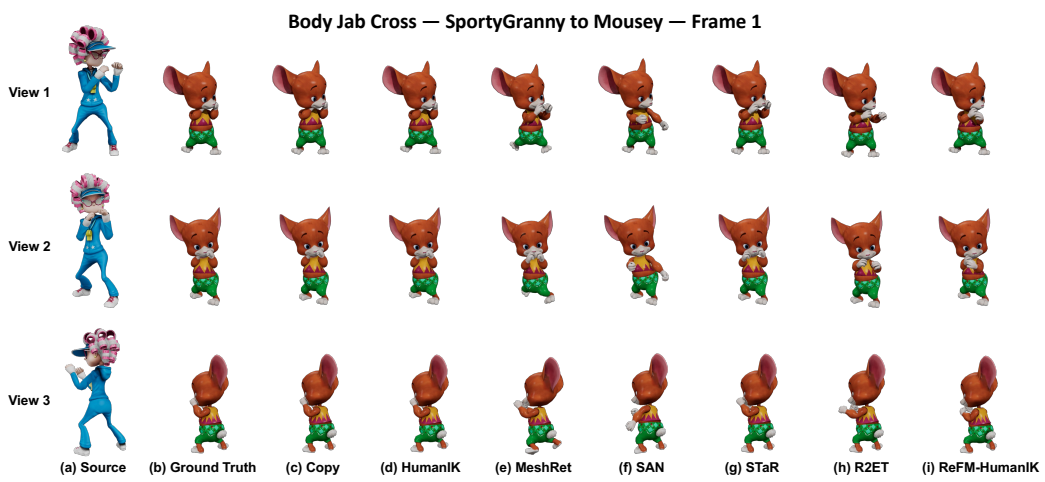


Figure 5: Additional qualitative comparison for *Body Jab Cross* on the Mixamo (Adobe) dataset at frame 1. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. The complete character view illustrates the overall pose and semantic consistency of different retargeting methods.

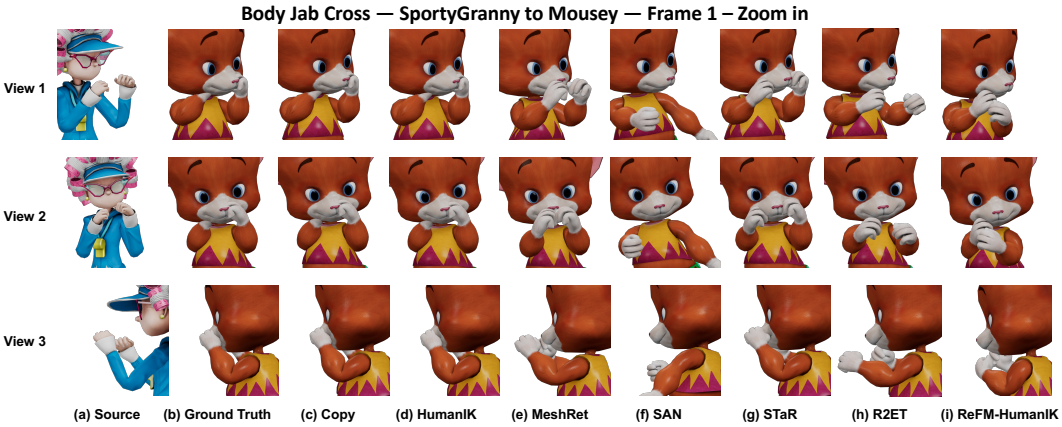


Figure 6: Zoomed-in comparison corresponding to Fig. 5. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. The enlarged view highlights the local region where self-penetration and geometry-sensitive interactions occur.

Figures 5 and 6 show *Body Jab Cross* at frame 1. While the full-frame comparison in Fig. 5 shows that different methods generally retain the characteristic body configuration of the source motion, the enlarged view in Fig. 6 reveals local penetration artifacts that are less apparent at the original scale. ReFM mitigates these geometric conflicts through localized refinement while largely preserving the transferred pose.

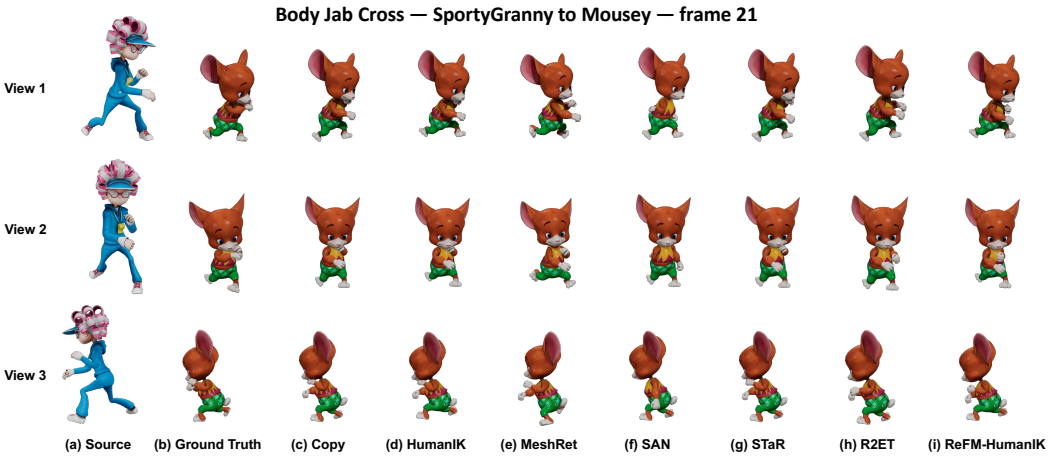


Figure 7: Additional qualitative comparison for *Body Jab Cross* on the Mixamo (Adobe) dataset at frame 21. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. The complete character view compares the overall retargeted configurations across different methods.



Figure 8: Zoomed-in comparison corresponding to Fig. 7. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. The enlarged visualization makes local geometric conflicts and penetration artifacts more directly observable.

Figures 7 and 8 provide another example from *Body Jab Cross* at frame 21. The zoomed-in comparison more clearly exposes differences in the geometric plausibility of the retargeted results around the interacting body regions. Compared with directly transferred or one-shot retargeted motions, ReFM produces a more geometrically compatible configuration without introducing substantial changes to the overall motion semantics.

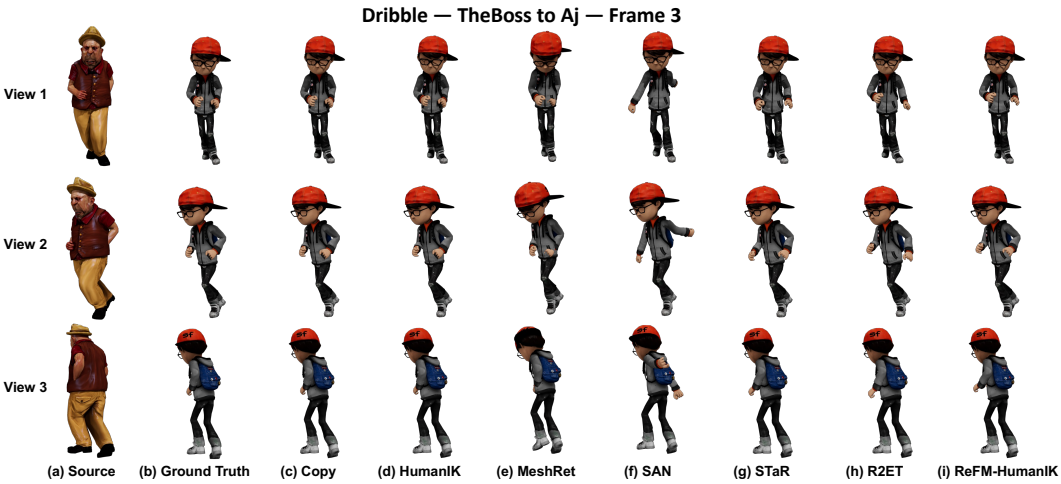


Figure 9: Additional qualitative comparison for *Dribble* on the Mixamo (Adobe) dataset at frame 3. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. The complete character view illustrates the preservation of the characteristic dribbling pose across different retargeting methods.

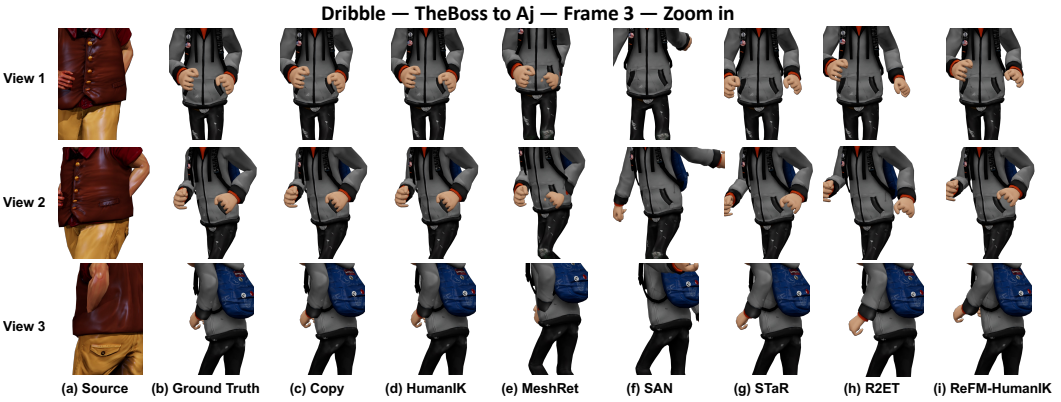


Figure 10: Zoomed-in comparison corresponding to Fig. 9. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. The enlarged view emphasizes the arm–torso interaction and the local self-penetration behavior of different methods.

Figures 9 and 10 further present *Dribble* at frame 3. In this example, the enlarged visualization highlights the interaction between the arm and torso, where mismatched body proportions can easily produce self-intersection. ReFM selectively adjusts the problematic region to reduce penetration while retaining the characteristic pose of the dribbling motion. Together, these examples further demonstrate the refinement principle of ReFM: rather than unnecessarily reconstructing the complete motion, the model focuses its modifications on regions where the initialized retargeting result violates geometric constraints.

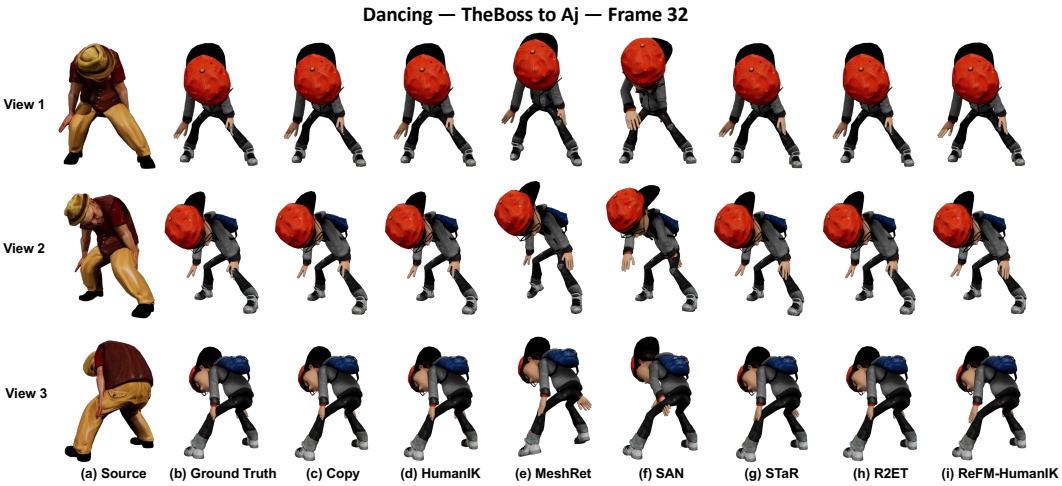


Figure 11: Additional qualitative comparison for *Dancing* on the Mixamo (Adobe) dataset at frame 32. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. The complete character view illustrates the preservation of the characteristic dancing pose across different retargeting methods.

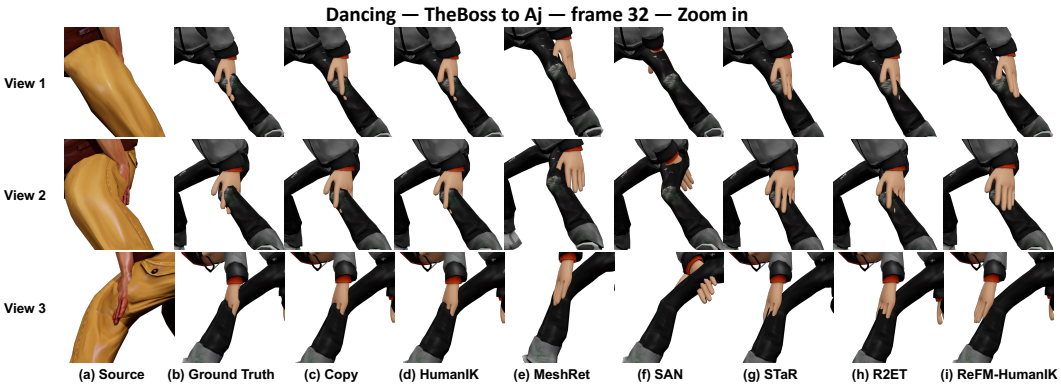


Figure 12: Zoomed-in comparison corresponding to Fig. 11. From left to right: (a) source motion, (b) ground truth, (c) Naive copy, (d) HumanIK (Autodesk, Inc., 2026), (e) MeshRet (Ye et al., 2024), (f) SAN (Aberman et al., 2020), (g) STaR (Yang et al., 2025), (h) R2ET (Zhang et al., 2023), and (i) ReFM initialized from HumanIK. The enlarged view emphasizes the hand–leg interaction and the local self-penetration behavior of different methods.

Figures 11 and 12 jointly illustrate that preserving the overall motion pose does not necessarily guarantee local geometric plausibility. In the full-character views, most methods retain the characteristic crouched posture of the source motion, indicating broadly consistent motion semantics. However, the zoomed-in views reveal substantial differences around the hand–leg interaction: several baselines retain visible intersections between the hand and target body, while others avoid penetration at the cost of larger changes to the local arm configuration. In comparison, ReFM-HumanIK produces cleaner spatial separation while remaining close to the source-derived pose across different viewpoints. This example further demonstrates the benefit of progressive target-aware refinement, which can correct localized geometric artifacts without unnecessarily altering the semantic structure of the initialized motion.

F.2 Results on ScanRet Dataset

We further evaluate ReFM on the ScanRet (Ye et al., 2024) dataset to examine its applicability to real-human motion data beyond Mixamo. Compared with Mixamo, the characters in ScanRet exhibit relatively moderate geometric variation, and direct motion copying and HumanIK already produce motions with limited self-penetration. Consequently, ScanRet requires substantially less target-specific geometric correction and provides a complementary setting for evaluating an important property of ReFM: *refinement should be applied only when necessary*. As summarized in Table 5, the reference motions already exhibit low penetration, and ReFM therefore requires only a small number of accepted refinement steps, with many sequences terminating at zero steps. This behavior follows directly from our energy-monitored inference procedure, which accepts a candidate update only when it decreases the motion energy and otherwise preserves the current motion. Thus, rather than forcing unnecessary modifications to an already satisfactory initialization, ReFM can adaptively retain the reference motion when little geometric correction is required.

Table 5: **Refinement behavior on the ScanRet dataset.** We report the penetration rate before and after ReFM refinement, the average number of accepted refinement steps, and the proportion of sequences requiring zero refinement steps. Statistics are computed over the ScanRet test pairs. The low initial penetration and limited number of accepted updates indicate that ScanRet generally requires substantially less geometric correction than Mixamo, which has a penetration rate of approximately 0.140.

Initialization	Initial Pen ↓	ReFM Pen ↓	Avg. Steps ↓	Zero-Step (%) ↑
Copy	0.092	0.078	0.61	39.0
HumanIK	0.091	0.079	0.60	40.5

As shown in Fig. 13, we uniformly visualize representative frames throughout three motion sequences for the ground truth, ReFM-Copy, and ReFM-HumanIK. Both ReFM variants closely preserve the characteristic poses and their temporal progression exhibited by the ground-truth animations, despite differences in initialization. In particular, the evolution of the major body configuration, limb articulation, and overall movement pattern remains consistent throughout the sequences rather than matching only individual poses. The overlaid temporal visualizations at the bottom of Fig. 13 further show that ReFM preserves the overall motion trajectory and dynamic structure across time. These results indicate that the refinement process does not introduce unnecessary modifications when severe geometric conflicts are absent: instead, it retains the semantic and temporal structure of the initialization while applying only necessary target-aware corrections. The consistent behavior of ReFM-Copy and ReFM-HumanIK demonstrates the effectiveness of ReFM on real-human motion data from ScanRet under different initialization strategies.

F.3 Inference-Time Analysis

ReFM learns to approximate an energy-decreasing refinement field rather than regressing toward a paired target motion. In particular, as shown in Eq. 18, the intermediate supervision states are constructed directly from the gradient of the motion energy. Since this energy depends only on the source motion, initialization, target skeleton, and target mesh, all of which are available at inference, the same objective can also be optimized directly for each test instance without accessing any ground-truth target motion. This provides a natural optimization-based counterpart for evaluating both the effectiveness and computational efficiency of the learned refinement flow.

Specifically, under copy initialization, we construct a *direct optimization* baseline that starts from the same motion as ReFM and follows the projected energy-descent direction in Eq. 16. It performs at most $K = 4$ iterations, selecting a step size from $\mathcal{A}' = \{0, 0.02, 0.05, 0.1\}$ using ReFM's energy accept/reject rule. Direct optimization repeatedly differentiates the complete motion energy, whereas ReFM predicts the refinement direction and evaluates candidate energies. Table 6 reports motion quality and inference time on the test dataset.

As shown in Table 6, ReFM-Copy achieves lower penetration (0.117 versus 0.131), higher semantic similarity (0.998 versus 0.994), and slightly higher MSE (0.086 versus 0.085) than direct optimization. Meanwhile, ReFM-Copy reduces the average inference time per motion from 4.3 s to 0.52 s,

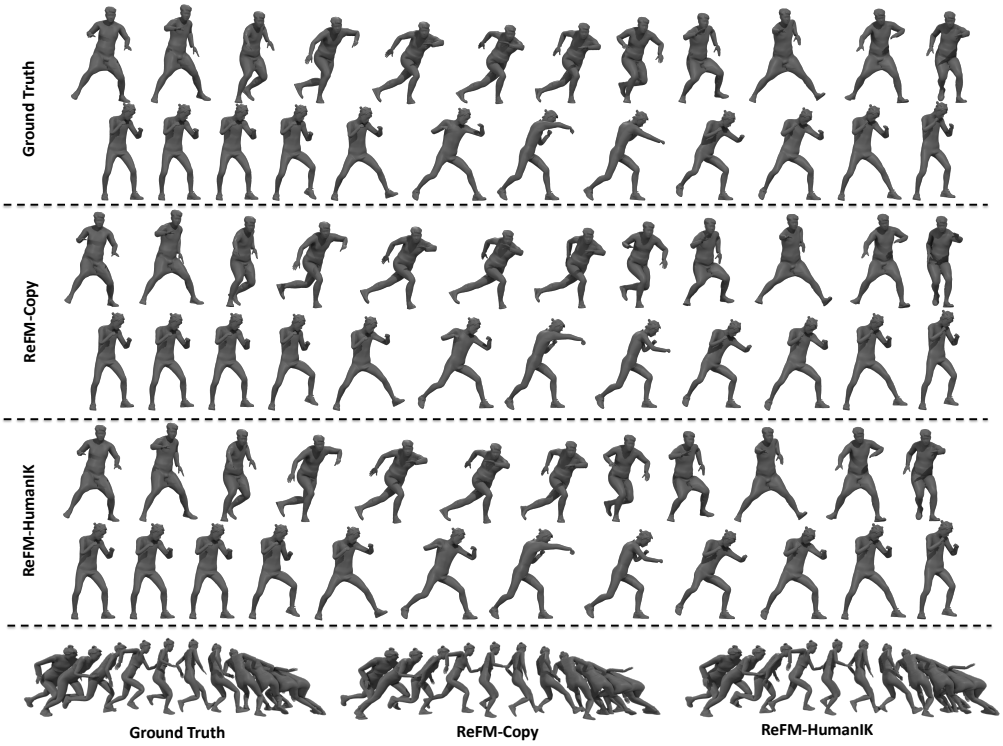


Figure 13: Qualitative temporal results on the ScanRet dataset. Three representative motion sequences are shown by uniformly sampled frames. We compare the paired ground-truth target animations with ReFM initialized from direct motion copying (*ReFM-Copy*) and HumanIK (*ReFM-HumanIK*). Since self-penetration is relatively infrequent in these ScanRet examples, the comparison primarily examines semantic preservation and temporal consistency.

Table 6: Inference-time analysis on the test dataset. Runtime is reported as the average inference time per motion. Direct optimization minimizes the same motion energy used to construct ReFM's refinement supervision, but performs the optimization explicitly at test time. $\downarrow$ and $\uparrow$ indicate that lower and higher values are better, respectively.

Method	Pen $\downarrow$	Sem_{sim} $\uparrow$	Curv $\downarrow$	MSE $\downarrow$	Time (s) $\downarrow$
Direct optimization	0.131	0.994	0.718	0.085	4.3
ReFM-Copy	0.117	0.998	0.772	0.086	0.52

corresponding to an approximately $8.3\times$ speedup. By replacing repeated computation of full-energy gradients with predicted correction directions, the learned refinement field improves the primary quality criteria under the same maximum iteration budget while substantially reducing inference time.

F.4 Statistical Reliability of the Penetration Improvement

We assess the reliability of the penetration improvements using paired measurements from the 222 benchmark cases. For each case, Δ_i denotes the penetration rate before refinement minus that after refinement, so that positive values indicate improvement. We report the mean and median paired reductions and estimate uncertainty in the mean using 95% cluster-bootstrap confidence intervals. To account for dependence among pairs sharing the same target character, bootstrap samples are constructed by resampling target characters and retaining all pairs associated with each sampled character.

As shown in Table 7, ReFM reduces mean penetration by approximately 0.023 under copy initialization and 0.024 under HumanIK initialization, corresponding to relative reductions of 16% and

Table 7: Statistical reliability of the penetration improvement on the Mixamo test set. Positive reductions favor refinement. Mean and relative reductions are reported at the precision used in Table 1. Confidence intervals for the mean reduction are computed from unrounded per-pair measurements using cluster bootstrap resampling over target characters.

	Copy initialization	HumanIK initialization
Mean reduction Δ	0.023	0.024
Relative reduction	16%	17%
Median reduction	0.0051	0.0086
95% cluster-bootstrap CI	[0.0133, 0.0312]	[0.0133, 0.0347]

17%, respectively. Both cluster-bootstrap confidence intervals lie strictly above zero, supporting a positive mean improvement after accounting for dependence among pairs sharing a target character. The positive median reductions further indicate that the gains are not confined to a few sequences with large improvements.

F.5 Generalization to Unseen Characters and Motions

Table 8: **Mixamo results by target-character familiarity.** "Seen" denotes pairs whose target appears in training ($n = 173$), while "unseen" denotes pairs with an unseen target ($n = 49$; Kaya, Ortiz, and XBot). Penetration values across groups should be interpreted cautiously because their ground-truth penetration levels differ substantially.

	Pen ↓		Curv ↓		MSE ↓		$\mathrm{Sem}_{\mathrm{sim}}$ ↑	
Method	seen	unseen	seen	unseen	seen	unseen	seen	unseen
Ground truth	0.154	0.093	0.621	0.771	—	—	0.978	0.963
Copy	0.154	0.090	0.628	0.762	0.046	0.067	0.999	0.998
HumanIK	0.155	0.091	0.628	0.759	0.045	0.065	0.999	0.997
STaR	0.150	0.088	0.673	0.805	0.072	0.083	0.999	0.998
MeshRet	0.148	0.089	0.676	0.792	0.104	0.095	0.989	0.937
SAN	0.150	0.091	1.516	1.428	0.163	0.154	0.961	0.931
R2ET	0.152	0.080	0.810	1.009	0.079	0.093	0.998	0.997
ReFM-Copy	0.128	0.080	0.761	0.810	0.089	0.074	0.998	0.998
ReFM-HumanIK	0.128	0.080	0.760	0.815	0.094	0.080	0.996	0.995

Unseen motions. The Mixamo benchmark already evaluates ReFM predominantly on unseen motions. Specifically, 218 of the 222 evaluation pairs use motion identities that never appear in the training split. Therefore, the results in Table 1 can already be interpreted as predominantly unseen-motion evaluation. We do not report a separate seen/unseen-motion breakdown because only four evaluation pairs contain motions observed during training, making the complementary subgroup too small for meaningful comparison.

Unseen characters. Among the 11 characters appearing in the 222-pair evaluation benchmark, three (Kaya, Ortiz, and XBot) are absent from the training split. This yields 49 pairs with an unseen *target* character and 173 pairs with a seen target. We focus on target-character novelty because the refinement field directly conditions on the target skeleton and mesh. Table 8 reports the corresponding results.

Analysis. ReFM reduces penetration on both seen and unseen target characters under either initialization. On unseen targets, ReFM-Copy lowers Pen from 0.090 to 0.080, while ReFM-HumanIK lowers it from 0.091 to 0.080. On seen targets, the corresponding initial penetration rates of 0.154 and 0.155 are both reduced to 0.128. Although the absolute reductions are smaller on unseen targets, these groups also exhibit substantially lower initial penetration. Therefore, this comparison alone does not isolate the effect of target-character novelty on refinement effectiveness.

Semantic consistency remains stable across target groups: ReFM-Copy achieves $\mathrm{Sem}_{\mathrm{sim}} = 0.998$ on both seen and unseen targets, whereas MeshRet decreases from 0.989 to 0.937. The auxiliary

metrics exhibit mixed changes: ReFM-Copy obtains lower MSE on unseen targets (0.074 versus 0.089), but somewhat higher Curv (0.810 versus 0.761). Together with the predominantly unseen-motion evaluation, these results support ReFM's ability to refine motions for unseen characters while retaining strong encoder-space semantic consistency.

F.6 User Study

Table 9: User-study mean ranks for five motion-retargeting alternatives. Rank 1 denotes the best result and rank 5 the worst. Avg. is the unweighted mean across the three criteria. Values are reported as mean rank $\pm$ standard error.

Method	Overall motion quality ↓	Self-penetration handling ↓	Semantic preservation ↓	Avg. ↓
Reference motions				
Naive copy	3.28 ± 0.10	2.99 ± 0.09	2.71 ± 0.08	2.99 ± 0.05
HumanIK	2.39 ± 0.09	2.66 ± 0.10	3.30 ± 0.10	2.78 ± 0.06
Skin-agnostic Retargeting Methods				
SAN	4.64 ± 0.08	4.41 ± 0.10	4.63 ± 0.07	4.56 ± 0.05
R2ET	2.65 ± 0.10	2.59 ± 0.09	2.13 ± 0.08	2.46 ± 0.05
ReFM-HumanIK (Ours)	2.03 ± 0.10	2.35 ± 0.10	2.02 ± 0.08	2.13 ± 0.05

Although $\mathrm{Sem}_{\mathrm{sim}}$ provides a quantitative measure of semantic consistency, it is computed using our pretrained semantic encoder, which also participates in the formulation and training of ReFM; therefore, it is not fully independent of the proposed framework. To the best of our knowledge, there is currently no established, model-independent metric in motion space that directly evaluates semantic preservation in motion retargeting. To complement the automatic evaluation, following common practice in existing retargeting studies, we therefore additionally perform a user study to provide an independent perceptual evaluation. Specifically, we recruit 20 participants and randomly sample 15 motions from the test set. Each participant evaluates the same subset and compares five alternatives: two reference motions (naive copy and HumanIK (Autodesk, Inc., 2026)), two source-mesh-agnostic baselines (SAN (Aberman et al., 2020) and R2ET (Zhang et al., 2023)), and ReFM-HumanIK. Participants evaluate the results along three dimensions: **overall motion quality**, **self-penetration handling**, and **semantic preservation**, with the last criterion directly assessing whether the semantic meaning of the source motion is retained after retargeting.

For each sampled pair, the five target-motion results are rendered using the same target character, camera viewpoints, and playback settings, with the source motion displayed separately as a semantic reference. Method identities are hidden, and the presentation order is randomized for each participant. Participants independently rank the five alternatives according to three criteria: **overall motion quality**, considering naturalness, temporal smoothness, and visual plausibility; **self-penetration handling**, favoring fewer and less severe visible body-part intersections; and **semantic preservation**, assessing fidelity to the source action and its characteristic gestures. For each criterion, participants assign ranks from 1 (best) to 5 (worst), with ties allowed.

The human evaluation in Table 9 is consistent with the quantitative observations in the main paper. ReFM-HumanIK achieves the lowest mean rank across all three evaluation criteria, with an overall average rank of 2.13. In particular, relative to the HumanIK initialization, its semantic-preservation rank improves substantially from 3.30 to 2.02, while its self-penetration rank improves from 2.66 to 2.35 and its overall motion-quality rank improves from 2.39 to 2.03. These results suggest that the refinement process can correct visible geometric artifacts while simultaneously improving the perceived semantic fidelity and overall quality of the initialized motion. Compared with the learning-based baselines, SAN receives substantially higher ranks across all three criteria, whereas R2ET remains comparatively competitive, particularly in semantic preservation, but obtains a higher average rank of 2.46 than ReFM-HumanIK. Overall, the user study provides complementary perceptual evidence that ReFM achieves a favorable balance between motion quality, geometric plausibility, and semantic preservation. Figures 14 and 15 further illustrate the interface used in the study, including the participant instructions and an example ranking page.

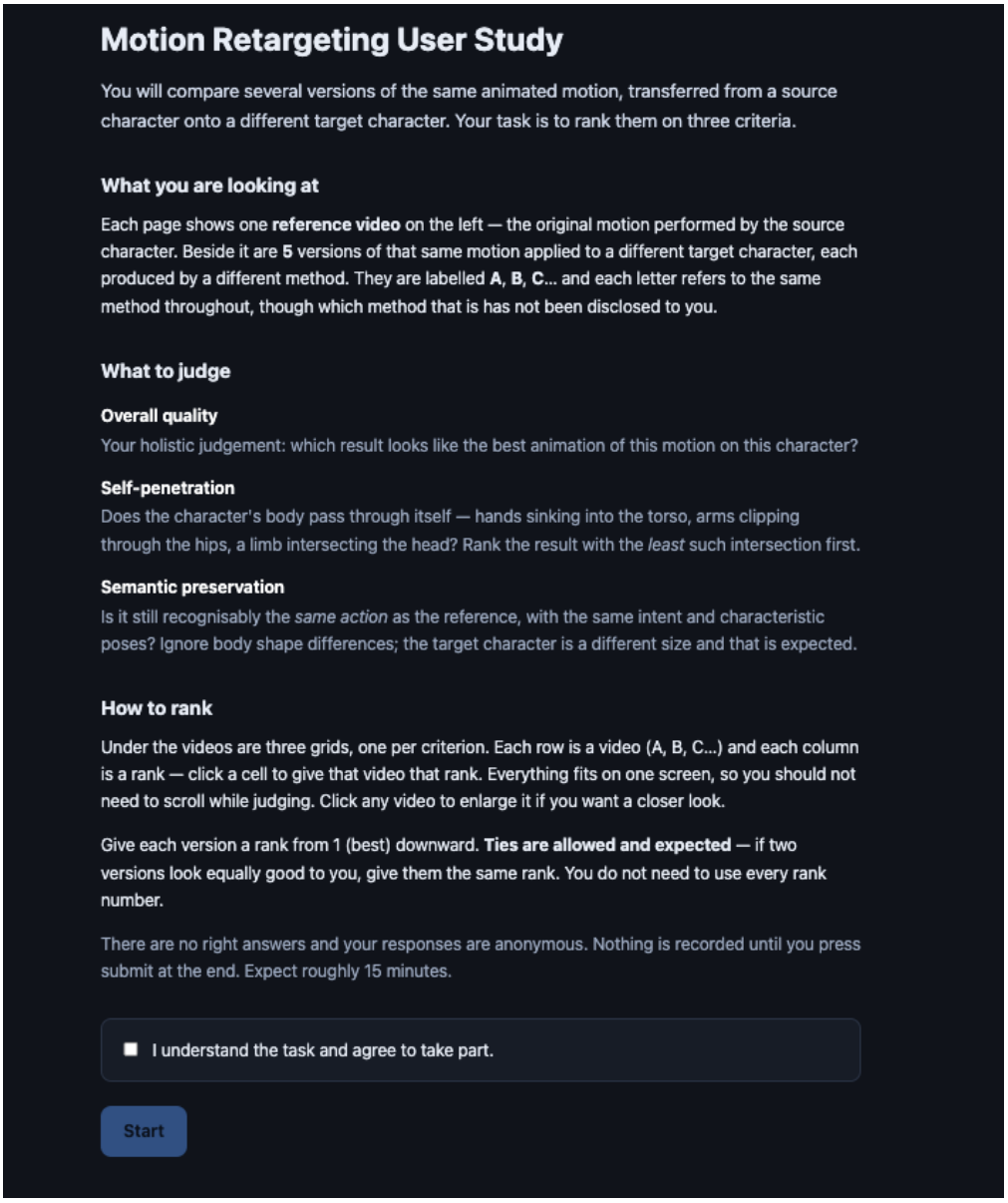


Figure 14: Guideline page of the user-study website. Participants are introduced to the evaluation protocol and the three ranking criteria: overall motion quality, self-penetration handling, and semantic preservation.

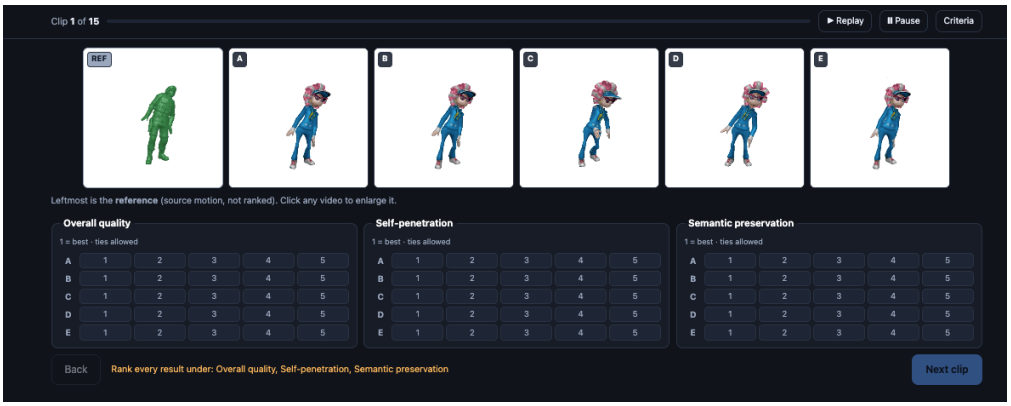


Figure 15: Example ranking page of the user-study website. The source motion is provided as the semantic reference, while the anonymized retargeting results are presented for participants to rank according to the specified evaluation criterion.