STaR: Seamless Spatial-Temporal Aware Motion Retargeting with Penetration and Consistency Constraints

Supplementary Material

We include the following sections in the supplemental materials:

- Further detail of limb penetration loss.
- Further detail of the evaluation metrics.
- Real-world applications: retargeting motion from real humans.
- User study.
- Ablation study on joint orientation loss.
- Ablation study on global motion prediction.
- Single-pass motion retargeting and separate motion retargeting.
- Ablation study on Dense Shape Representation.
- Efficiency of limb penetration constraint module.
- Demo videos.

A. Further Detail of Limb Penetration Loss

Figure 7 illustrates penetration detection using Signed Distance Fields (SDF) in motion retargeting. The large sphere represents the body surface, while the curved trajectory represents the motion path of a limb or body part. Along this trajectory, multiple vertices exist, some penetrating the body (inside the sphere) and others remaining outside. The penetration loss computation involves finding the nearest reference vertex on the body surface for each motion path vertex, computing the vector from the query vertex to its reference vertex, and multiplying this vector by the normal vector of the reference vertex to estimate penetration depth. The black and red arrows represent surface normals, which guide penetration correction. Blue arrows indicate vectors from penetrating vertices to their nearest reference points, while cyan arrows represent similar vectors for non-penetrating vertices. This visualization highlights how full-body geometric correction is applied across the motion trajectory, ensuring that motion retargeting maintains geometric plausibility by preventing unnatural interpenetration.

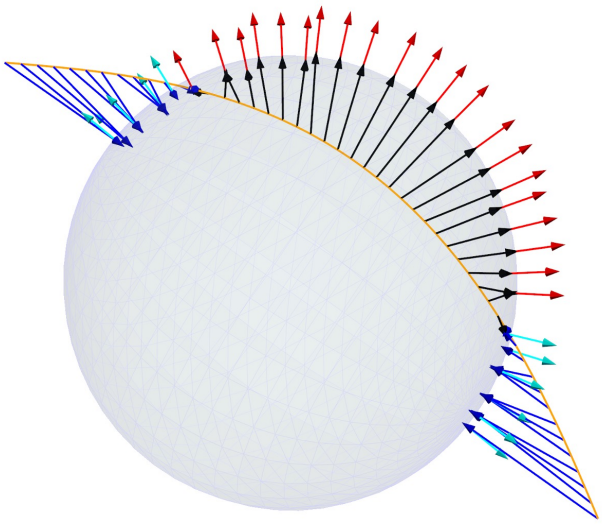

Figure 7. The detail of limb penetration loss computation for penetrated vertices and non-penetrating vertices.

B. Evaluation Metrics

We evaluate the retargeted motion primarily from three perspectives: semantics, geometry, and motion smoothness. These are measured using MSE, penetration rate, and curvature, respectively.

Mean Squared Error. The Mean Squared Error (MSE) evaluates semantic preservation by assessing how closely the retargeted skeleton joints $\hat{\mathbf{X}}$, align with the ground truth joints, $\mathbf{X}_{gt}$. Although the ground truth suffers severe penetration, we investigate the motion sequences, and they can still work as an auxiliary evaluation of how semantics is maintained. The squared error is normalized by the character's height h. The metric is formulated as:

$$\mathbf{MSE} = \frac{1}{h}\|\mathbf{X}_{gt} - \hat{\mathbf{X}}\|_2^2. \tag{8}$$

Penetration Rate. The penetration rate is calculated as the ratio of interpenetrating points to the total number of limb vertices. Unlike [20], our approach considers all limbs:

$$\mathbf{Pen\ Rate} = \frac{\text{Number of penetrated limb vertices}}{\text{Number of all limb vertices}} \times 100\%. \tag{9}$$

Curvature. To address the discontinuity in the retargeted motion path, we compute the curvature of the motion path for each joint based on acceleration. Let $\mathbf{r}$ represent the motion vector, the curvature is defined as:

$$\mathbf{Curv} = \left\| \frac{\mathrm{d}^2\mathbf{r}}{\mathrm{d}t^2} \right\|_2^2. \tag{10}$$

C. Real-world Applications

Figure 8 presents the motion retargeting results of our STaR on real-human motion data from the ScanRet dataset [18]. The left column shows real-human motions with texture

maps removed for anonymity, while the right columns display the retargeted motions on diverse target characters. STaR effectively preserves motion semantics, transferring poses and movement dynamics while adapting to different body shapes, including a wrestler, a stylized boy, and a cartoon-like figure. Despite variations in skeletal structures, STaR maintains spatial and temporal coherence, ensuring natural adaptation without excessive limb stretching or severe interpenetration. The results demonstrate STaR's ability to generalize human motion to stylized characters while preserving geometric plausibility and temporal consistency.

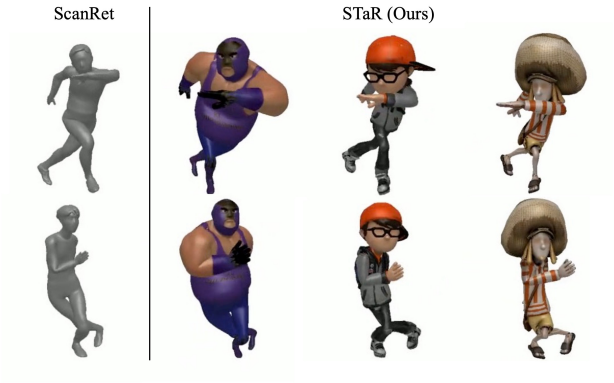

Figure 8. Motion retargeting from real human actors.

D. User Study

We conducted a user study to compare the retargeted motion sequences of STaR with three methods: SAN [1], R^{2}ET [20], and MeshRet [18]. Participants were presented with 10 randomly selected motion sequences, each containing the source motion and retargeted results from all four methods, shown in a randomized order to prevent bias.

Participants were asked to evaluate the results based on four key aspects:

1. Which one better preserves the semantics of the original motion?
2. Which one is physically more reasonable (fewer penetrations and distortions)?
3. Which movement appears smoother and more natural?
4. Which one do you prefer overall?

To ensure a diverse set of responses, we distributed questionnaires and collected feedback from 21 participants. The results, presented in Tab. 3, show that STaR significantly outperforms all baseline methods across all criteria. Notably, 77.14% of participants favored STaR in terms of motion semantics preservation, and 76.19% preferred our results overall. Our model also received 70.48% approval for motion coherence, demonstrating its ability to produce smooth and temporally stable motion trajectories, while achieving 73.33% in geometric correctness, highlighting its effectiveness in preventing interpenetration and ensuring spatial plausibility.

Compared to other methods, SAN [1] and R^{2}ET [20] struggle with geometric plausibility, while MeshRet [18] fails on characters with diverse body shapes, leading to excessive penetration issues. In contrast, STaR consistently maintains a balance between motion semantics, geometric correctness, and temporal consistency, making it the preferred choice for high-quality motion retargeting.

E. Ablation Study on Joint Orientation Loss

Figure 9 illustrates an ablation study on the joint orientation loss, comparing results without the loss (left) and with the loss (right). In the absence of this constraint, our STaR model, which operates in a large search space, occasionally produces unnatural poses, such as flipped arms or misaligned limb orientations. These artifacts arise due to the increased flexibility of the model, which, without explicit regularization, may lead to implausible joint rotations.

By incorporating the joint orientation loss, as shown on the right, the model effectively regulates joint rotations, ensuring physically plausible limb orientations while preserving motion semantics. This demonstrates the importance of enforcing orientation constraints to prevent extreme limb deviations and enhance the stability of motion retargeting.

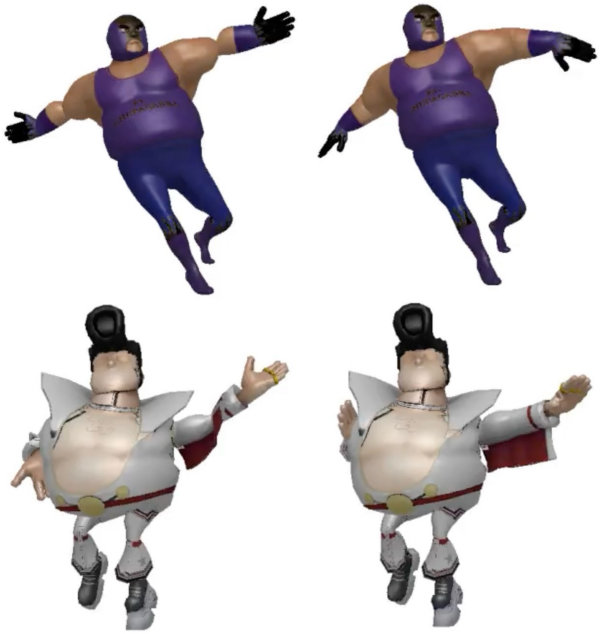

Figure 9. Ablation study on joint orientation loss. Without this loss (left), STaR's large search space may lead to unnatural poses, such as flipped arms. Adding the loss (right) ensures physically plausible joint orientations.

Table 3. Human evaluation results between our STaR and baseline methods.

Criteria	SAN[1]	R^{2}ET [20]	MeshRet[18]	STaR (Ours)
Semantics Preservation	11.43%	8.57%	2.86%	77.14%
Geometry Correctness	12.38%	9.52%	4.76%	73.33%
Motion Smoothness	20.00%	4.76%	4.76%	70.48%
Overall Quality	12.38%	8.57%	2.86%	76.19%

F. Ablation Study on Global Motion Prediction

After examining the relationship between global motion and character height in the Mixamo dataset [2], we observed no statistically significant correlation. After introducing a compact decoder to enhance global motion prediction, we observed rapid overfitting to the training set, resulting in poor generalization. As shown in Tab. 4, its performance does not exceed the baseline, which simply normalizes and denormalizes global motion relative to character height.

Table 4. Ablation study on global motion prediction. Since the global motion prediction module will only affect the **MSE** metric, the other three metrics are omitted for clarity. Please note that this study evaluates a preliminary variant—not our final model configuration.

Methods	MSE ↓	MSE^{lc} ↓	Pen% ↓	Curv ↳
Global Motion Prediction				
Baseline	1.7770	-	-	-
Global Motion Decoder	1.8708	-	-	-

G. Single-pass Motion Retargeting and Separate Motion Retargeting

As presented in Tab. 5, our well-designed spatio-temporal model supports motion retargeting of varying sequence lengths within a single forward pass. The results show no significant difference between separate retargeting and single-pass retargeting.

Table 5. The results of single-pass motion retargeting and separate motion retargeting

Methods	MSE ↓	MSE^{lc} ↓	Pen% ↓	Curv ↳
Separated Inference V.S. Inference Once				
Model 1 Separate Inference	0.0369	0.0175	7.99	10.61
Model 1 Inference Once	0.0368	0.0174	7.99	10.66
Model 2 Separate Inference	0.0355	0.0162	8.41	8.69
Model 2 Inference Once	0.0355	0.0162	8.41	8.66

H. Ablation study on Dense Shape Representation

For shape representation, we exclude skeleton data and skeleton-based shape information due to their limited utility and potential drawbacks. The skeleton bounding box's dimensions [20] provide limited shape details, which are insufficient for effective motion retargeting that minimizes penetration. This limitation arises because motion is influenced by the shapes of limbs and other body parts. We show 3 results as proof in Tab. 6: (1) skeleton data only; (2) our DSR; (3) their combination. Comparing (1) and (2), skeleton data is inadequate for preventing penetration. Additionally, (3) shows that integrating skeleton data with point clouds does not enhance results.

In DSR, the point cloud is NOT separated before being fed into the point cloud transformer [7], and the K-channel shape representation is derived via a compact MLP from the comprehensive geometric feature produced by the transformer. Unlike frame-by-frame methods [19, 20] that rely on spatially aligned skeleton data, our approach extracts both global and local information for each joint. This enables the spatial and temporal transformers to access comprehensive shape information, particularly benefiting the temporal network. In Tab. 6 (4), we show the result of extracting a separate point cloud for each joint as proof. Comparing (2) and (4), separating the points does not benefit the retargeting pipeline.

Table 6. Ablation study on Dense Shape Representation.

Methods	MSE ↓	MSE^{lc} ↓	Pen% ↓	Curv ↳
Global Motion Prediction				
(1) Skeleton Info	0.0367	0.0175	9.42	10.11
(2) DSR (Ours)	0.0368	0.0174	**7.99**	10.66
(3) Skeleton Info + DSR	0.0366	0.0174	8.22	10.05
(4) Separate Point Cloud	0.0934	0.0783	13.45	**8.70**

I. Efficiency of the Limb Penetration Constraint Module

In Tab. 7, we compare the training speeds of (1) the SDF loss $\mathcal{L}_{sdf}$ from R^{2}ET [20], (2) modified SDF loss [20] including limb-limb penetration, and (3) our limb penetration

constraint $\mathcal{L}_{lp}$. Using the same number of vertices, we measured the time per iteration on an RTX 4090 graphics card. Our limb penetration loss is approximately 8 times faster compared with the SDF loss from [20].

Table 7. Loss Efficiency.

Methods	**Speed (s/iter)** $\downarrow$
$\mathcal{L}_{sdf}$ [20]	127.66
$\mathcal{L}_{sdf}$ [20] w/ limb	133.77
$\mathcal{L}_{lp}$ (Ours)	**16.32**

J. Demo Videos

We provide demo videos to showcase the performance of our STaR method. These videos show the motion retargeting from real-human datasets, ScanRet [18]. We include the demo videos in the supplementary materials.