EUROGRAPHICS 2025 / A. Bousseau and A. Dai
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 2

InterFaceRays: Interaction-Oriented Furniture Surface Representation for Human Pose Retargeting

Taeil Jin[1] Yewon Lee[2] and Sung-Hee Lee[2]

[1]Korea Institute of Science and Technology, Republic of Korea
[2]Korea Advanced Institute of Science and Technology, Republic of Korea

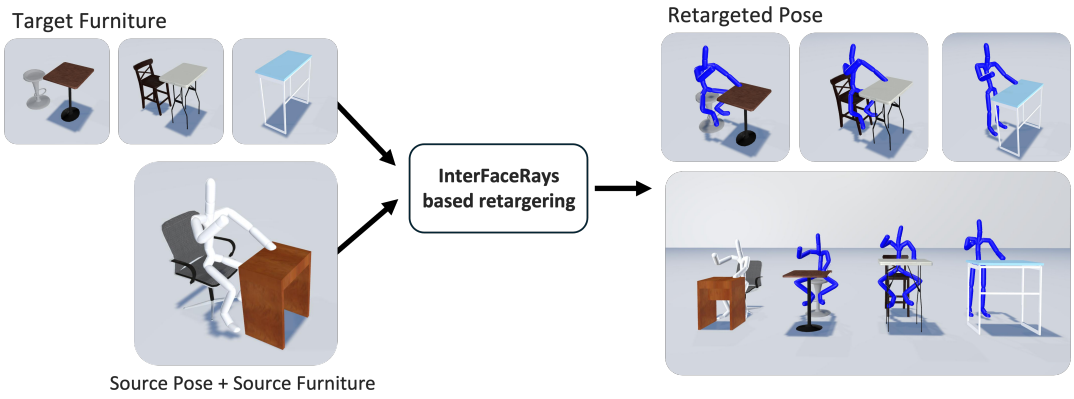

Figure 1: *We propose a novel human-furniture pose retargeting framework for retargeting source pose to significant changes of target furniture.*

Abstract

Motion retargeting is a well-established technique in computer animation that adapts source motion to fit characters with different sizes, morphologies, or environments. Recent deep learning methods have shown promising results in retargeting character motion. However, retargeting human-object interactions to new environments, especially when furniture shapes differ significantly, remains a challenging problem. In this work, we propose a novel retargeting framework to address this challenge by combining motion generative models with optimization-based pose adaptation. Our framework operates in two stages: first, a key pose generator generates the pose of key joints that preserves the interaction state relative to the new furniture; second, final whole-body pose is determined by accommodating the key joints' poses through optimization. A crucial step in our framework is generating key poses that maintain the interaction state of the source motion. To achieve this, we introduce the Interaction Intensity Weight (IIW) and structural rays, called InterFaceRays, which together capture the interaction intensity between body parts and furniture surfaces. The IIW generator, a trained MoE-based decoder from the conditional variational autoencoder (cVAE) model, infers IIWs for the target furniture based on the source motion's interaction state. Extensive experiments demonstrate that our framework effectively retargets continuous character motion across diverse furniture configurations, with the IIW generator significantly enhancing key pose consistency. This hybrid approach offers a robust solution for motion retargeting across dissimilar furniture environments.

CCS Concepts
• ***Computing methodologies*** → ***Motion processing;***

1. Introduction

Motion retargeting is a technique used to adapt a source motion to fit a character of different size, morphology, or an altered environ-

ment, and has long been a focus of research in computer animation. Optimization methods have been widely applied to motion retargeting by setting the target character or environment as constraints while preserving the source motion as the optimization objective. Recently, deep learning-based approaches have harnessed strong regression capabilities to learn the mapping between the source and target characters, enabling adaptation to different bone structures and body surfaces [AMN24], as well as to different species or humanoid forms [LWC*23, YML23].

Compared to retargeting motion for a new character, retargeting human-object interaction motion to a new environment—such as different furniture—poses a greater challenge, as it requires consideration of interaction semantics and, in some cases, object affordance. If the source and target furniture have similar shapes, allowing for a continuous one-to-one correspondence between the surfaces, interaction motion can be directly retargeted by adjusting the motion to fit the geometry of the target furniture within optimization frameworks [KPBL16, CHC*23]. However, the difficulty of motion retargeting increases as the dissimilarity between the source and target furniture grows, often exceeding the capabilities of geometric optimization.

For example, if a furniture component involved in the interaction in the source motion is absent in the target furniture—such as retargeting a chair-sitting pose with arms on the armrests to a chair without armrests—the resulting motion must be creatively adapted to accommodate these differences. A more challenging scenario arises when the source and target furniture share the same affordance but differ significantly. For instance, if the source motion involves writing at a desk while seated, and the target environment consists of a standing board, a plausible retargeting solution that preserves the original task would involve standing and writing on the board.

Another line of animation research has focused on generating motions for given environmental conditions using generative deep learning models. Previous works have allowed high-level conditions, such as action labels [HCV*21, SZKS19] or hand interaction labels [ZBS*22], to serve as inputs for generating appropriate poses in given environments.

In this paper, we propose a novel framework for retargeting furniture interaction motions to furniture with significantly different shapes, structures, or types by leveraging motion generative models and optimization-based pose adaptation approaches. We assume a retargeting problem for static furniture configurations, focusing on the precise mapping of body-part interactions to the target environment, specifically for office furniture. Our framework consists of two stages. First, given the interaction state of the source motion, our key pose generator produces the position and orientation of the key joints in the output motion while maintaining the interaction state relative to the new furniture. We represent the interaction state through the presence of contact and the associated type of furniture for the effector parts, such as the pelvis and hands. Next, the final whole-body pose is determined by accommodating the key joints' poses through optimization.

A crucial step in our framework is generating key poses that preserve the interaction state of the source motion. To achieve this, we introduce the Interaction Intensity Weight (IIW), which represents the interaction intensity between a sample point on a furniture surface and the effector parts. Given the source motion's interaction state and the target furniture environment, a conditional variational autoencoder (cVAE) model, referred to as the IIW generator, infers IIWs for sample points on the target furniture. The IIWs of a dense set of sample points provide strong cues for determining the pose of individual body parts to maintain the interaction state with the target furniture. Specifically, we obtain surface sample points by projecting structural rays, termed *InterFaceRays*, from the character's current root position to the surrounding furniture surface. In this way, we represent the interaction characteristics of the furniture surface through the IIW values of the InterFaceRays.

The inferred IIWs are then passed to the key pose generator, along with other inputs, to generate a key pose. The key pose generator produces the key pose autoregressively, allowing for continuous motion, such as engaging with the target furniture. Once the engagement is complete, hand pose retargeting is additionally applied to ensure the task is performed as required.

The effectiveness of our motion retargeting technique is validated through extensive experiments. We demonstrate that our framework successfully retargets continuous character motion across diverse furniture combinations. Additionally, quantitative experiments are conducted to evaluate the contribution of our IIW generator in improving the quality and consistency of key pose generation.

Our work makes the following contributions:

- Our framework uses InterFaceRays to retarget continuous character motion across various furniture configurations, including dissimilar types.
- The IIW generator improves the pose consistency of key-pose generation with the corresponding interaction labels
- Our framework introduces a hybrid approach combining a conditional pose generative model with an optimization process, improving motion retargeting for dissimilar furniture configurations.

2. Related Work

2.1. Deep learning-based skeletal motion retargeting

Motion retargeting networks using deep learning have shown success in retargeting to target characters with diverse topologies [ALL*20, VYCL18, VCH*21, ZWK*23] and to non-human characters [LWC*23, YML23]. Additionally, recent research has focused on retargeting motions to humanoid robots [YML23, AMN24] to generate synthetic human-like motions for robot motion learning.

Early motion retargeting methods for different bone structures primarily relied on supervised learning with ground truth pairs of source and target poses [JKY*18, VYCL18]. Specifically, graph convolutional networks [ZLH*24, DM24] encode poses with different topologies into a single graph node, representing structure-invariant pose semantics. To alleviate the need for paired training datasets in supervised learning, unsupervised approaches, such as cycleGAN, [ZPIE17] have been proposed [AMN24, ZWZZ24, LWC*23]. More recent studies have considered not only bone structure but also the character's surface in retargeting [ZWK*23,

ZWZZ24]. One such study introduced a method that separates the latent values of the character's bones and surface using cross-attention.

Unlike motion retargeting in free space, our research focuses on finding appropriate poses that can adapt to various target furniture configurations. Existing methods typically require corresponding target motion data for source motions to enable end-to-end learning, but in cases involving environmental interactions, determining such corresponding pose data becomes challenging. For example, a character might be sitting in a chair, but in the target environment, the character should stand when the chair is absent. In contrast, our environmental motion retargeting approach addresses this retargeting problem in situations where defining paired pose correspondences is difficult.

2.2. Shape Correspondence-based Pose Adaptation

Research on generating target poses by adapting the source pose to satisfy given constraints, such as target hand position, through optimization is a long-established field [Gle98, GFK*23]. Previous furniture retargeting works adapt the source pose to the target furniture by constructing a spatial correspondence map between the surrounding spaces of the source and target furniture surfaces.

Object representation research can identify surface-level correspondences between source and target objects. Recent work by Huang et al. [HXH*24] employs a self-supervised approach to learn the signed distance field (SDF) of furniture. To represent the spatial relationships between two objects, the Neural Interaction Field [HXD*23] is inferred using the spherical harmonics of spatial points between the source and target objects, obtained through IBS [ZWK14], a spatial relationship representation. These methods can establish correspondences even when there are significant shape differences between objects. However, as these methods focus on representing object surfaces, they can only identify contact points on the surface. Furthermore, these methods are hard to establish correspondences between highly dissimilar furniture setups, such as when some parts of the target furniture are absent. In contrast, our retargeting framework adapts motion performed not only on the object surface but also near it. At each time step, the IIW is extracted from the InterFaceRays projected onto the object, guiding the generation of a reference key joint pose for pose adaptation, which allows retargeting across significantly different furniture setups.

For retargeting motions performed around objects, Kim et al. [KPBL16] created a correspondence map between the surrounding spaces of two objects by modeling the free space around the source object as a tetrahedral volume mesh and deforming it to fit the target object. Choi et al. [CHC*23] improved upon this work by applying the thin-plate spline method, which deforms the uniform point cloud around the source furniture to maintain manually specified corresponding key points on the surface of the target furniture, without relying on watertight surface meshes. However, these methods are effective only for objects with similar shapes. When two objects have different shapes—for example, one chair has a backrest while the other does not—defining spatial correspondences near these unmatched surfaces becomes impossible. Tonneau et al. [TAAP*16] manually selected corresponding sample points between the object surfaces to allow for efficient and robust motion adaptation. However, this method requires an appropriate trunk pose for the target furniture, which is not always easy for all furniture types. Our framework addresses this problem by learning a key pose generator for various target furniture types, automatically obtaining corresponding sample points by emitting rays from the character to the furniture.

2.3. Human-Object Interaction in Indoor Scene

The relationship between humans and objects has been a consistent focus of research in computer graphics and vision. Recent works have developed methods to effectively overlap 3D objects and human meshes onto 2D image of human-object interactions [XBPM22, XBLPM24]. These approaches focus on preserving realistic 3D contact interactions, particularly for key regions like hands and feet, between the human mesh and the objects in the scene. Although these works primarily focus on generating static poses, our key pose generator addresses the challenge of generating coherent motions that align with the scene.

Deep-learning-based research for generating interaction motions with furniture has been actively explored. Starke et al. [SZKS19] proposed the Neural State Machine (NSM), which generates motions that interact with the environment using input action labels. Zhang et al. [ZBS*22] added action label inputs for the hands , and recent diffusion model-based works have focused on predicting contact areas for predefined interactions, such as holding and sitting, based on input objects and text prompts [PXW*23, DD24]. Hassan et al. [HCV*21] utilized the autoregressive structure of the NSM to train a pose generator that can produce various poses for the target furniture based on given action inputs. Recently, Jin et al. [JL23] proposed an unsupervised action predictor capable of generating actions without requiring input action type labels, even in complex furniture configurations that include both a chair and a desk. In these methods, the environment is approximately represented by voxels [HCV*21, ZBS*22] or sphere-shaped colliders around the character to calculate the occupancy by the furniture [JL23]. In contrast, our work explores an efficient surface-level representation for human-object interaction.

Our surface-level representation is similar to that in [LSY*23], where dynamic environments are represented using collision points from rays cast vertically downward from the character. However, a major difference is that we infer the IIW of the sample surface points, while [LSY*23] uses these sample points solely to represent the environment geometry. In addition to vertical rays, our method projects radial rays. Vertical rays cannot differentiate between vertically occluded objects, such as a chair beneath a desk. In contrast, our InterFaceRays can represent complex furniture combinations, allowing for an accurate representation of furniture surfaces. Thus, our interaction-oriented furniture surface representation enhances the quality and consistency of key poses for interaction motion retargeting.

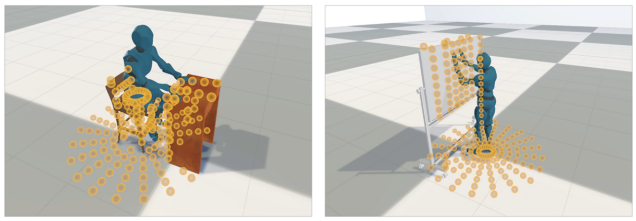

Figure 2: *Visualization of ray points: orange circles indicate the intersected ray points.*

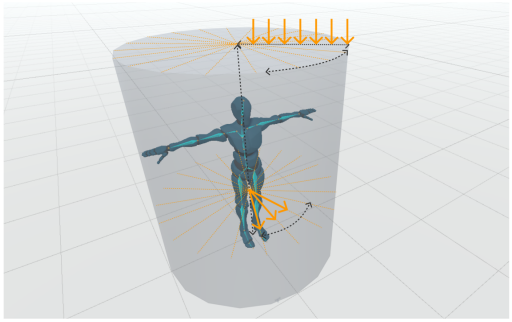

Figure 3: *Visualization of InterFaceRays. InterFaceRays are projected vertically from the top and radially along the vertical axis from the character root to identify the points on nearby furniture surfaces.*

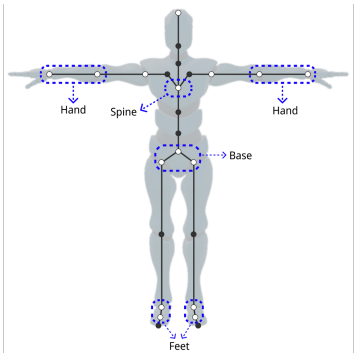

Figure 4: *Visualization of indicator joints (enclosed in dashed circles) from the key joints (white dots).*

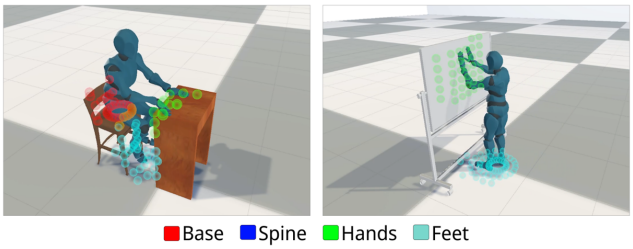

Figure 5: *Visualization of InterFaceRays. The color of the spheres on the furniture represents the IIW for each body part: red for the base, blue for the spine, green for the hands, and cyan for the feet. While there are originally six IIW values, the two hands and two feet are visualized with the same color in this figure.*

3. Method

3.1. Overview

The human-furniture interaction scene S in our retargeting framework is defined as follows:

$$S = [p, R, E], \tag{1}$$

where $p \in \mathbb{R}^{9j}$ denotes a pose of j (= 25) number of joints with respect to $R \in \mathbb{R}^{4\times4}$, the transformation matrix of the root. We define the orientation of a single joint as a 9-dimensional vector consisting of the joint's position, an up-vector, and a forward direction. The root's position is determined as the ground-projection of the pelvis joint, and its orientation aligns with the up-vector and the forward direction of the pelvis.

We define InterFaceRays to represent the environment surface E through ray points, as illustrated in Figure 2.

$$E = [E^{pos}, E^{label}, E^{IIW}] \tag{2}$$

InterFaceRays are obtained from n (=540) ray points, as shown in Figure 3. For the vertical rays, the top circular surface of the cylinder is divided into 20 slices, each 18 degrees apart, and each slice's radius is split into seven sections, creating 140 points for ray emission. For the horizontal rays, the cylinder's height is divided into 20 segments, with 20 rays emitted per segment, resulting in 400 rays. Combining these vertical and horizontal rays, a total of 540 rays are cast in each frame.

For each ray, we check for collision with the environment, identifying the 3D coordinates of the collision points $E^{pos} \in R^{3n}$ and the type of collided furniture $E^{label} \in 6n$, where the furniture type is encoded in a binary vector with six categories: chair, desk, bed, board, ground, and null. Non-colliding rays are assigned a maximum collision distance of 10 meters. As a result, the InterFaceRays capture the furniture surface shape in both source and target environments. Figure 2 shows the obtained surface points for example environments. The Interaction Intensity Weight (IIW) matrix $E^{IIW} \in \mathbb{R}^{n\times6}$ will be detailed in Sec. 3.2.

Body Part Contact Description From the given source scene, we extract its interaction state in terms of the body contact status label c^b and the furniture contact status label c^f. These contact status labels are used to maintain body part-level interaction consistency between the source and target scenes.

For c^b, we check whether the effector parts are in contact with environment. The effector parts include six body regions: pelvis, spine, right hand, left hand, right foot, and left foot. c^b is a binary vector indicating the contact status of each effector. Contact for each effector is determined by the environment contact of the corresponding indicator joints, as shown in Figure 4. For example, [1,0,1,1,1,1] represents the body contact status in interaction scene of Figure 5(left).

In addition, we identify the furniture type each effector is in con-

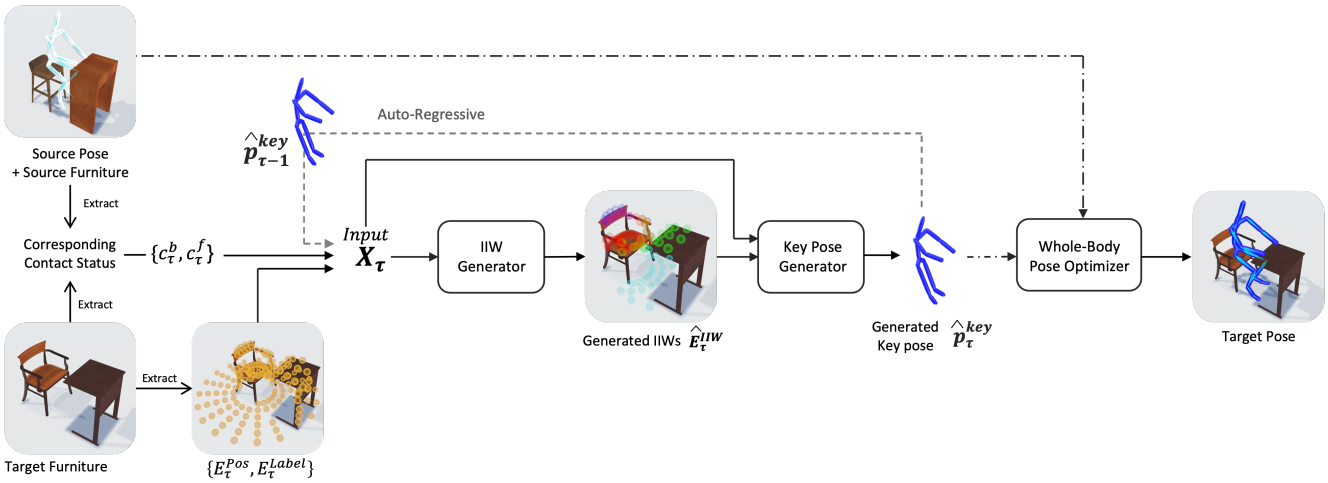

Figure 6: *An overview of our retargeting framework: Given a target furniture configuration and a scene where the character interacts with the source furniture, our framework adjusts poses to adapt to the diverse shapes and configurations of different furniture.*

tact with using a separate binary vector c^f, similar to E^{label}. c^f is generated for each effector part. For example, the furniture contact status vector of pelvis is [1,0,0,0,0,0] for the sitting posture in Figure 5(left), while it is [0,0,0,0,0,1] for the standing posture in Figure 5(right).

Pose Retargeting Framework Figure 6 illustrates the overall architecture of our framework. Our pose retargeting framework takes the contact status labels extracted from the source scene S_{src} and target furniture as inputs, adapting the source pose to fit the target furniture while maintaining the contact status labels as much as possible. Based on these inputs, the IIW generator infers the IIW for each InterFaceRay in the target environment, corresponding to the contact status labels of the source scene. The IIWs represent the desired intensity of interaction between the effector parts and each sample point on the target furniture.

Next, the Key Pose Generator produces the appropriate key pose based on the input contact status labels and the inferred IIWs. The key pose, denoted as $p^{key} \in \mathbb{R}^{9\times k}$, refers to the pose of k(=15) joints, including the trunk (pelvis, left hip, right hip, spine, head), arms (shoulder, elbow, hand), and feet (foot, toe-base). The key-pose generator is trained as a conditional pose generative model to accurately generate key poses for interacting with the target furniture.

Hand movements, such as pointing, are crucial for preserving the source's semantics. To enhance the interaction consistency, the whole-body pose optimizer adapts the target hand movements based on the key pose, using the corresponding InterFaceRays.

3.2. Interaction Intensity Weight

We denote $E^{IIW} \in \mathbb{R}^{n\times 6}$ as the Interaction Intensity Weight (IIW) matrix, where n represents the number of InterFaceRay points, and each column corresponds to one of the six effector parts. IIW represents the intensity values of each InterFaceRay for the six body part interactions. As a part of the InterFaceRays representation, the IIW matrix reflects the interaction strength of each ray with the environment, based on the proximity of the surface to the body parts. Figure 5 shows the results of E^{IIW} for various furniture configurations, such as a chair-desk setup or a whiteboard.

We build the training dataset for IIW from multi-contact motions collected across various human-furniture interaction scenes. First, we obtain the InterFaceRays at the current time step τ, with E^{IIW} initialized to zero for all. We assume that an InterFaceRay point closer to a effector body will have a stronger interaction intensity than those farther away. Based on this assumption, we compute the IIW in terms of the distance between each InterFaceRay's intersecting point and all associated indicator joints, where each indicator joint corresponds to a specific effector part.

The IIW for the i-th body part is computed using its indicator joints. The weight $w[a,j]$, representing the a-th point of E^{pos} at the j-th indicator joint, is defined as follows:

$$w[a,j] = \begin{cases} 0.0 & \text{if } d_{a,j} > l_b + u_b \\ 1.0 & \text{if } d_{a,j} < l_b \\ \frac{l_b+u_b}{d_{a,j}+u_b} & \text{otherwise,} \end{cases} \quad (3)$$

where $d_{a,j}$ denotes the distance between the position of a-th point of E^{pos} and the j-th indicator joint. The lower and upper bounds of distance, l_b and u_b, are set to 0.1 meters and to a quarter of the character's height (1.8 meters) in our experiment. We finally obtain the IIW for each body part, $E^{IIW}[a,i]$, by taking the mean of all weights $w[a,j]$ associated with the effector part i. As a result, $E^{IIW}[a]$ contains the interaction intensity weights of InterFaceRay a for all effector parts. For example, when standing, the InterFaceRays near the ground will have a higher IIW for the feet, as shown in Figure 5.

E^{IIW} is updated at each time step based on the positions of the InterFaceRays. For example, as shown in Figure 7(bottom), when a person is walking from a distance, the IIW for the left hand is 0,

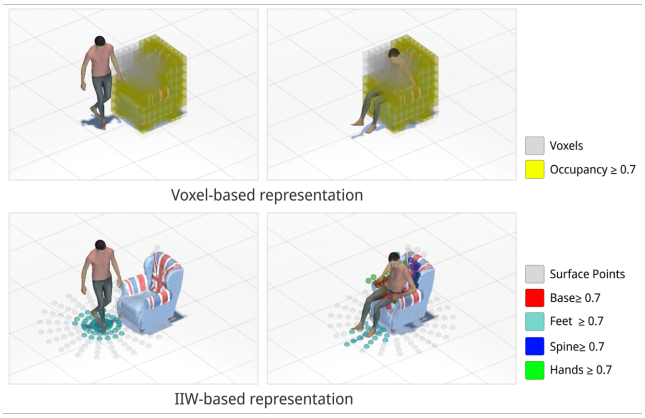

Figure 7: *Comparison of environment representations. Uniform voxels (top) statically fill a bounding box, while our method (bottom) generates dynamic interaction points in real-time.*

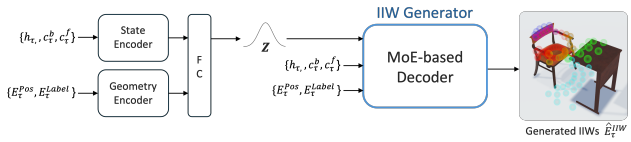

Figure 8: *Overview of the network architecture for training the IIW generator using a conditional VAE.*

but it increases when the person sits down and puts the hand on the armrest.

IIW Generator In the retargeting process, the IIW generator should generate the IIWs for the target furniture to maintain the contact status labels in the input scene. The IIW generator is trained in a supervised manner to infer the IIW matrix $E^{IIW} \in \mathbb{R}^{n\times 6}$ based on the interaction scene feature, defined as:

$$X_\tau = \{E^{pos}_\tau, E^{label}_\tau, c^b_\tau, c^f_\tau, h_\tau\} \tag{4}$$

Here, h_τ refers to a history condition, consisting of the previous key pose $p^{key}_{\tau-1}$ and the root velocities of the past 10 frames, where the root velocity from $R_{\tau-1}$ to R_τ is defined with respect to $R_{\tau-1}$. Adding h_τ to the input of the IIW generator enhances the continuity of the generated IIW. Figure 5 shows the result of generating IIWs in the scene X.

Network Structure We adopted a conditional VAE structure to train E^{IIW}. The network structure is shown in Figure 8. Specifically, we employed a Mixture-of-Experts (MoE)-based decoder, which we named the IIW generator. MoE-based decoders have been recently used in cVAE frameworks to handle multi-condition generation tasks, as shown in previous works [HCV*21, ZBS*22, SZKS19, JYJ*23].

We divide the multi-condition inputs into two distinct types encoded separately. The state encoder S is responsible for encoding controllable conditions, specifically the body contact status label c^b, the furniture contact status label c^f, and the history condition

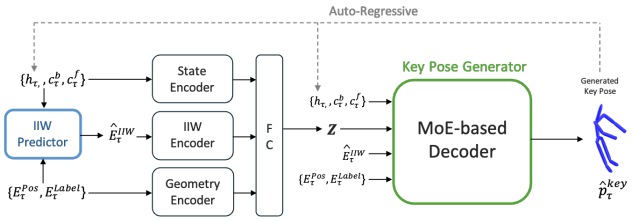

Figure 9: *Structure of our key pose generator.*

h. The geometry encoder G encodes two geometrical conditions: the InterFaceRays positions, E^{IIW}_{pos}, and the furniture label, E^{IIW}_{label}. These outputs represent the positional and categorical information of the InterFaceRays. Finally, the outputs from both the state encoder and the geometry encoder are passed through Fully Connected layers (FC) to compute the parameters of the Gaussian distribution. The reparametrization trick is then applied to these parameters to transform them into the latent value z, which serves as the input to the MoE-based decoder.

The inferred $\hat{E}^{IIW}_\tau$ is produced by the MoE-based decoder D^{IIW} using the latent value z and the interaction scene feature X:

$$\hat{E}^{IIW}_\tau = D(z, X_\tau). \tag{5}$$

The entire networks are trained an supervised manner by minimizing the loss terms below:

$$\mathbb{L}^{IIW} = \mathbb{L}^{IIW}_{rec} + \beta \mathbb{L}^{IIW}_{KL}. \tag{6}$$

The reconstruction term $\mathbb{L}^{IIW}_{rec}$ is defined as the mean squared error between the ground truth interaction intensity weight E^{IIW} and the predicted value $\hat{E}^{IIW}$. The prior loss term $\mathbb{L}^{IIW}_{KL}$ ensures that the Gaussian distribution learned by the encoder remains close to the standard normal distribution, promoting consistency during the latent space representation. The weight β is set to 0.1.

3.3. Pose Retargeting with Key Pose Generator

The key pose generator produces the key pose for the current time step using noise sampled from the prior distribution, the interaction scene feature X_τ, and the generated interaction intensity weight E^{IIW}_τ. The final pose retargeting is achieved by using the generated key pose as the initial solution and performing pose adaptation based on the positional information E^{pos} and the label information E^{label} for both the source and target furniture.

Key Pose Generator We trained our key pose generator using an MoE-based cVAE, similar to the IIW generator. Figure 9 shows the detailed structure of the key pose generator. The main difference between training the cVAE for the key pose generator and the IIW generator is that we incorporate the IIW encoder I to encode the IIW generated by the IIW generator. The key pose $\hat{p}^{key}$ is computed by the MoE-based Decoder D^{key} from the latent value z and the interaction scene feature X:

$$\hat{p}^{key}_\tau = D^{key}(z, X_\tau). \tag{7}$$

The entire networks are trained in an supervised manner with fixed IIW generator by minimizing the loss terms below:

$$\mathbb{L}^{key} = \mathbb{L}^{key}_{rec} + \beta \mathbb{L}^{key}_{KL} \tag{8}$$

The reconstruction term $\mathbb{L}^{key}_{rec}$ is modeled as a mean squared error between p^{key} and $\hat{p}^{key}$. The prior loss term $\mathbb{L}^{key}_{KL}$ ensures consistency between the Gaussian distribution from the encoder and the normal distribution. The trained key pose generator is an auto-regressive model that current key pose is input to the interaction scene feature X_τ for inferring the next key pose $p^{key}_{\tau+1}$.

Whole-body Pose Optimization The optimization process obtains the whole-body pose q of current time step τ by minimizing the cost terms below:

$$q = \arg\min_q C_k + C_h + C_r, \tag{9}$$

where C_k ensures that the pose p follows the generated key pose p^{key}.

$$C_k = \sum_{i=0}^{15} \omega_i \left\| p^{key}_i - p^c_i(q) \right\| \tag{10}$$

The function $p^c_i(q) \in \mathbb{R}^9$ extracts the i-th key joint from a pose. Another cost C_h ensures that the positions of the hands, $p^c_h(q)$, follow the corresponding source hand positions p_h in the target space.

$$C_h = \omega_h \left\| p_{rh} - p^c_{rh} \right\| + \omega_h \left\| p_{lh} - p^c_{lh} \right\| \tag{11}$$

To obtain the target hand positions, we search for corresponding points between source and target InterFaceRays. We identify the furniture type related to the hand interaction during the contact status extraction process. In the retargeting process, we compare E^{label} values and identify the source and target points with matching indices as the corresponding points cp.

The relationship vectors dir from the source corresponding points cp^{src} to the hand positions are computed. The final target hand position p_{lh},p_{rh} are calculated using Eq.12.

$$p_h = \sum_i r_i (dir_i + cp^{tar}_i) \tag{12}$$

Finally, C_r ensures that the pose q remains consistent with the previous pose q_{prev}, serving as a regularization term in the optimization process.

$$C_r = \omega_r \left\| q - q_{prev} \right\| \tag{13}$$

We set the weight ω_r to 0.1. The weight ω_i is assigned a higher value for the trunk joints. We assign different weights to the arm joints depending on the types of furniture interaction. When the character is engaging with the furniture, we assign a lower value to ω_h than ω_i. Once the character finishes engaging and begins to hand movements, such as pointing, we assign a higher value to ω_h than ω_i. Figure 10 shows the different whole-body poses optimized by varying ω values.

4. Implementation Details

We obtained the motion and furniture status data using the Unity game engine. Multi-contact motion data from DAFNet [JL23] was used to train all network structures. We augmented the DAFNet dataset by retargeting poses to accommodate various heights of furniture. This augmentation process expanded the dataset to more than 2 hours of 30 FPS motion data, while the test data was around 23 minutes long. Figure 11 shows the furniture used in the training dataset.

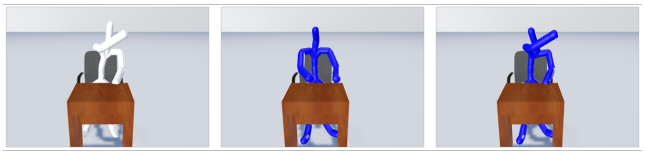

Figure 10: *Comparison of different cost weights for arm joints. Left: source pose. Middle: ω_i is higher than ω_h. Right: ω_i is lower than ω_h.*

Figure 11: *Furniture in the training set.*

All network layers for both the encoder and decoder were modeled with MLP and implemented using Pytorch. Pose optimization was performed with Theseus [PFM*22], an optimization library. We employed the Levenberg-Marquardt algorithm with an error tolerance set of 0.3, a maximum of 15 iterations, and a step size of 0.7. The IIW predictor and key pose generator are autoregressive networks, where the generated key pose is set as the previous key pose $p^{key}_{\tau-1}$ for inferring the next time step. To improve autoregression performance, we applied scheduled sampling, beginning with ground truth previous key poses as inputs in the early epochs and gradually replacing the input data with the network's output.

We standardized the data for both the character's pose and root velocity. Training took approximately 9 hours for the IIW generator and 10 hours for the key pose generator on a GeForce NVIDIA H100. After training, all experiments were run on a Ryzen 5 CPU with 6 cores.

5. Experiment

In this section, we qualitatively and quantitatively evaluate the plausibility and consistency of target motions generated by our retargeting framework. Please refer to the accompanying video for clear qualitative evaluations.

5.1. Qualitative Result

We demonstrate the effectiveness of our framework in adapting the source motion across various shapes of furniture. We use motion sequences not included in the training data and experiment with both trained and unseen furniture.

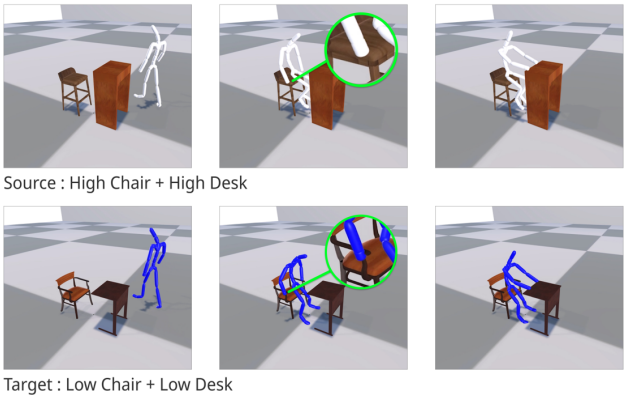

Figure 12: *Retargeting to different object heights and shapes. The intermediate pose (middle) illustrates the retargeted pose, maintaining proper hand contact (green circle) aligned with the target furniture geometry.*

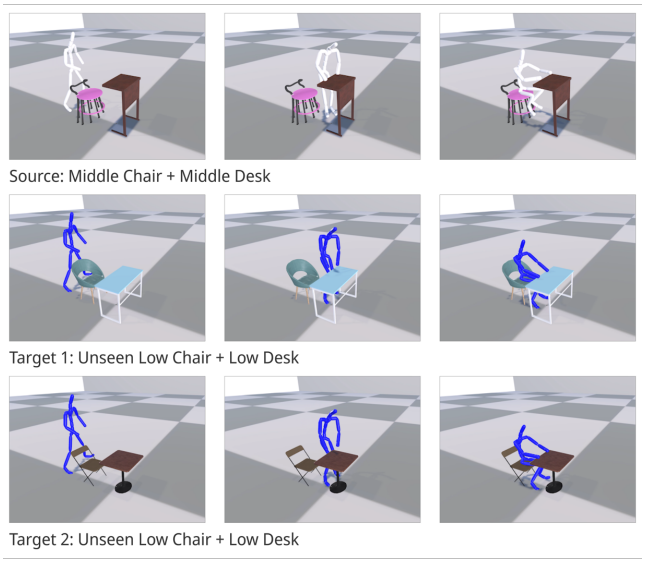

Figure 13: *Retargeting motion to unseen furniture.*

Retargeting for Diverse Shapes Our model is capable of retargeting the source pose for furniture with diverse shapes. The top of Figure 12 demonstrates that the source interaction is preserved when retargeting to a desk and chair combination with different heights or shapes but the same configuration. Figure 13 illustrates how the model generates target interaction motions suitable for previously unseen furniture.

Adaptation to Highly Dissimilar Furniture Our framework addresses the retargeting problem in challenging cases by modifying the corresponding interaction labels. Figure 14 shows that that even when the corresponding target furniture is missing, the source interaction is generated as appropriately as possible for the target furniture. When no target furniture is available for interaction, we

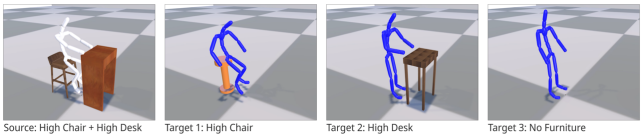

Figure 14: *Retargeting to un-paired source and target furniture state.*

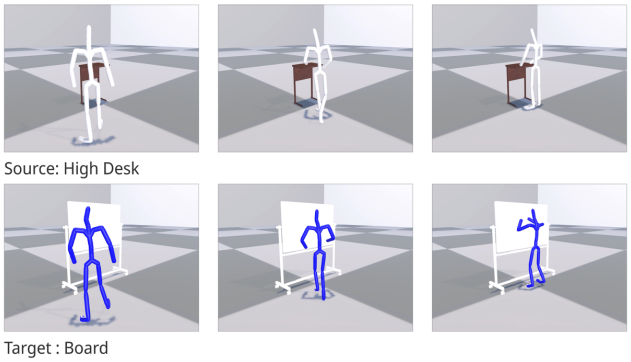

Figure 15: *Retargeting to different surface normal.*

specify the corresponding body contact label c^b and furniture label f^c as null values. Figure 15 illustrates the translation of a pose from a desk to a board, where we change the corresponding environment label from a desk to a board.

5.2. Quantitative Result

We used two types of motion datasets for comparison and ablation studies. We compared the InterFaceRays with the trained IIW generator against three existing types of environmental sensors. For comparison, we constructed a single furniture dataset that includes a low/high chair, a high desk, and a board, totaling 3 minutes of motion sequences. Training our entire framework took approximately 12 hours, while training the other models required about 7 hours. For the ablation study, we trained the ablated models using our training datasets. The test dataset included five types of furniture: low chair, high chair, high desk, chair-desk combinations of low and high heights, and a whiteboard, totaling approximately 8.7 minutes of motion sequences. Each test data set includes motions that approach furniture from different directions, sit or stand, and initiate interaction motions, with furniture shapes that were unseen during training.

We assessed the models' performance using three metrics. For pose consistency relative to the input interaction label, we employed the Fréchet Motion Distance (FMD) metric and the Bad Interaction (BI) metric. FMD calculates the distance between the feature vectors of the ground truth motion and the generated motion. This metric is inspired by the Fréchet Inception Distance (FID) [HRU*17], commonly used in image generation to evaluate the plausibility of generated images, but adapted here for motion evaluation. A lower FMD value indicates that the generated motion exhibits similar pose consistency to the ground truth motion.

To evaluate BI, we attached sphere colliders with a radius of 10 cm to every key pose joint and performed collision detection between the character and the furniture. For each frame, if the collision of three source body parts—two hands and the pelvis—did not occur in the target scene, it was counted as a Bad Interaction (BI). This count is then divided by the total number of frames.

For pose plausibility, we measured the frame ratio of foot sliding (FS). A foot was considered to be sliding if the toe-base joint collided with the ground and had a velocity greater than 1.0 cm/s. If both feet were sliding in a frame, FS was increased by two. The FS value is then divided by the total number of frames.

Comparison with Environmental Sensors We compare the environment representations of existing environment-conditioned pose generation models with our InterFaceRays. For comparison, our key pose generator is modeled by removing E^{label}, and all environmental conditions of test models are encoded to the same dimension.

We modeled the Voxel, where the environmental condition is represented as a voxel-based geometry sensor, similar to SAMP [HCV*21]. The environmental condition of the Cylinder model is created using surrounding sphere-shaped colliders to obtain occupancy values, following the same setup as DAFNet [JL23]. The environment condition of Voxel+Cylinder combines a voxel-based sensor with sphere sensors, following the same environmental conditions used in the pose regression model NSM [SZKS19].

Table 1 presents the comparison results. Our model outperformed the other models in most criteria regarding the pose consistency of the input interaction. The Voxel shows poor performance across most criteria, suggesting that a simple static geometry condition is not effective for generating the proper key pose with the input interaction labels. While Voxel+Cylinder exhibits lower values in both FMD and FS, the differences are negligible. In FS, the difference is only 0.42 percentage points, meaning Ours produces one additional foot sliding frame every 7.94 seconds at 30 FPS, which has minimal practical impact. Similarly, in FMD, Ours underperforms Voxel+Cylinder by just 0.013. In contrast, Ours outperforms Voxel+Cylinder in BI by 11.38 percentage points.This is a significant difference, as Voxel+Cylinder generates one more Bad Interaction frame every 0.29 seconds. These results show the superiority of the IIWs of InterFaceRays in reflecting user's body part interactions.

Table 1: *Comparison Results and Ablation Study.*

Comparison Results	FMD ↓	FS ↓	BI ↓
Voxel	2.9860	0.0184	0.2902
Cylinder	2.5900	0.0179	0.2073
Voxel+Cylinder	**2.0970**	**0.0175**	0.2797
Ours	2.1097	0.0217	**0.1659**
Ablation Study	FMD ↓	FS ↓	BI ↓
Ours-IL	1.7312	0.0276	0.0826
Ours-L	1.5539	0.0216	0.0880
Ours	**1.4195**	**0.0202**	**0.0761**

Ablation study Ours-IL is an ablated version of our model, where both the IIW generator and the IIW encoder are removed from the key pose generator. Ours-L is another ablation version of Ours, modeled without E^{label} of the InterFaceRays. Table 1 presents the result of ablation study. Our model outperforms the compared models in all criteria, suggesting that the IIW generator plays a crucial role in improving the consistency of the generated key poses with the input interaction labels.

6. Limitation and Future Works

Our framework has several limitations that need to be addressed by future research. We address retargeting for various furniture configurations. However, our method may produce unnatural interactions in cases of significant height differences between the target furniture, such as sitting on a low chair while reaching for a high desk. Addressing this limitation requires developing an algorithm that adjusts source interaction labels to better align with target furniture configurations, paired with an enhanced pose generation approach. Future work could also investigate dynamic furniture manipulation to enable realistic human-furniture interactions, such as repositioning a chair for sitting.

We focused on quantifiable metrics to assess motion quality, but a perceptual evaluation could examine a more comprehensive aspects on the motion quality. Our key pose generator handles multiple contacts in a single frame. However, retargeting for complex scenes, such as varying room sizes and object assembly, may require an advanced reference motion generator to enhance adaptability to the scenes.

We use body part interactions to capture the semantics of both source and target scenes. Recently, Zhang et al. [ZCX*24] fine-tuned a pre-trained skeleton retargeting network to align the semantics of two characters by inferring each textual token from the same description and rasterizing the characters' images into a pre-trained Vision-Language Model (VLM). By incorporating the VLM, we could represent the more complex body part interactions, such as "reading a book while drinking coffee," from the source scene.

7. Conclusion

In this paper, we proposed a novel framework for retargeting human-furniture interaction motions to furniture with significantly different shapes, structures, or types. Our approach operates in two stages: first, motion generative models are used to create key joint poses that preserve the interaction state of the source motion; second, optimization-based pose adaptation refines the full-body pose. We introduce InterFaceRays and the Interaction Intensity Weight (IIW) generator to capture the interaction between body parts and the surface, ensuring key pose consistency. Extensive experiments validated the framework's ability to effectively retarget motions across a wide range of diverse furniture configurations, offering a robust solution for adapting character motions to environments with dissimilar furniture layouts.

Acknowledgement

This work was supported by IITP, MSIT, Korea (2022-0-00566) and NRF, Korea (2022R1A4A5033689).

References

[ALL*20] ABERMAN K., LI P., LISCHINSKI D., SORKINE-HORNUNG O., COHEN-OR D., CHEN B.: Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG) 39*, 4 (2020), 62–1. 2

[AMN24] ANNABI L., MA Z., NGUYEN S. M.: Unsupervised motion retargeting for human-robot imitation. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction* (2024), pp. 204–208. 2

[CHC*23] CHOI S., HONG S., CHO K., KIM C., NOH J.: Online avatar motion adaptation to morphologically-similar spaces. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, pp. 13–24. 2, 3

[DD24] DILLER C., DAI A.: Cg-hoi: Contact-guided 3d human-object interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 19888–19901. 3

[DM24] DOU Y., MUKAI T.: Facial animation retargeting by unsupervised learning of graph convolutional networks. In *2024 Nicograph International (NicoInt)* (2024), IEEE, pp. 69–75. 2

[GFK*23] GRANDIA R., FARSHIDIAN F., KNOOP E., SCHUMACHER C., HUTTER M., BÄCHER M.: Doc: Differentiable optimal control for retargeting motions onto legged robots. *ACM Transactions on Graphics (TOG) 42*, 4 (2023), 1–14. 3

[Gle98] GLEICHER M.: Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques* (1998), pp. 33–42. 3

[HCV*21] HASSAN M., CEYLAN D., VILLEGAS R., SAITO J., YANG J., ZHOU Y., BLACK M. J.: Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 11374–11384. 2, 3, 6, 9

[HRU*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems 30* (2017). 8

[HXD*23] HUANG Z., XU J., DAI S., XU K., ZHANG H., HUANG H., HU R.: Nift: Neural interaction field and template for object manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)* (2023), IEEE, pp. 1875–1881. 3

[HXH*24] HUANG Z., XU H., HUANG H., MA C., HUANG H., HU R.: Spatial and surface correspondence field for interaction transfer. *arXiv preprint arXiv:2405.03221* (2024). 3

[JKY*18] JANG H., KWON B., YU M., KIM S. U., KIM J.: A variational u-net for motion retargeting. In *SIGGRAPH Asia 2018 Posters*. 2018, pp. 1–2. 2

[JL23] JIN T., LEE S.-H.: Dafnet: Generating diverse actions for furniture interaction by learning conditional pose distribution. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, p. e14962. 3, 7, 9

[JYJ*23] JANG D.-K., YANG D., JANG D.-Y., CHOI B., JIN T., LEE S.-H.: Movin: Real-time motion capture using a single lidar. In *Computer Graphics Forum* (2023), vol. 42, Wiley Online Library, p. e14961. 6

[KPBL16] KIM Y., PARK H., BANG S., LEE S.-H.: Retargeting human-object interaction to virtual avatars. *IEEE transactions on visualization and computer graphics 22*, 11 (2016), 2405–2412. 2, 3

[LSY*23] LEE S., STARKE S., YE Y., WON J., WINKLER A.: Questenvsim: Environment-aware simulated motion tracking from sparse sensors. In *ACM SIGGRAPH 2023 Conference Proceedings* (2023), pp. 1–9. 3

[LWC*23] LI T., WON J., CLEGG A., KIM J., RAI A., HA S.: Ace: Adversarial correspondence embedding for cross morphology motion retargeting from human to nonhuman characters. In *SIGGRAPH Asia 2023 Conference Papers* (2023), pp. 1–11. 2

[PFM*22] PINEDA L., FAN T., MONGE M., VENKATARAMAN S., SODHI P., CHEN R. T., ORTIZ J., DETONE D., WANG A., ANDERSON S., DONG J., AMOS B., MUKADAM M.: Theseus: A Library for Differentiable Nonlinear Optimization. *Advances in Neural Information Processing Systems* (2022). 7

[PXW*23] PENG X., XIE Y., WU Z., JAMPANI V., SUN D., JIANG H.: Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553* (2023). 3

[SZKS19] STARKE S., ZHANG H., KOMURA T., SAITO J.: Neural state machine for character-scene interactions. *ACM Transactions on Graphics 38*, 6 (2019), 178. 2, 3, 6, 9

[TAAP*16] TONNEAU S., AL-ASHQAR R. A., PETTRÉ J., KOMURA T., MANSARD N.: Character contact re-positioning under large environment deformation. In *Computer Graphics Forum* (2016), vol. 35, Wiley Online Library, pp. 127–138. 3

[VCH*21] VILLEGAS R., CEYLAN D., HERTZMANN A., YANG J., SAITO J.: Contact-aware retargeting of skinned motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 9720–9729. 2

[VYCL18] VILLEGAS R., YANG J., CEYLAN D., LEE H.: Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 8639–8648. 2

[XBLPM24] XIE X., BHATNAGAR B. L., LENSSEN J. E., PONS-MOLL G.: Template free reconstruction of human-object interaction with procedural interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 10003–10015. 3

[XBPM22] XIE X., BHATNAGAR B. L., PONS-MOLL G.: Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision* (2022), Springer, pp. 125–145. 3

[YML23] YAN Y., MASCARO E. V., LEE D.: Imitationnet: Unsupervised human-to-robot motion retargeting via shared latent space. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)* (2023), IEEE, pp. 1–8. 2

[ZBS*22] ZHANG X., BHATNAGAR B. L., STARKE S., GUZOV V., PONS-MOLL G.: Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision* (2022), Springer, pp. 518–535. 2, 3, 6

[ZCX*24] ZHANG H., CHEN Z., XU H., HAO L., WU X., XU S., ZHANG Z., WANG Y., XIONG R.: Semantics-aware motion retargeting with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 2155–2164. 9

[ZLH*24] ZHENG B., LIANG D., HUANG Q., LIU Y., ZHANG P., WAN M., SONG W., WANG B.: Frame-by-frame motion retargeting with self-collision avoidance from diverse human demonstrations. *IEEE Robotics and Automation Letters* (2024). 2

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2223–2232. 2

[ZWK14] ZHAO X., WANG H., KOMURA T.: Indexing 3d scenes using the interaction bisector surface. *ACM Transactions on Graphics (TOG) 33*, 3 (2014), 1–14. 3

[ZWK*23] ZHANG J., WENG J., KANG D., ZHAO F., HUANG S., ZHE X., BAO L., SHAN Y., WANG J., TU Z.: Skinned motion retargeting with residual perception of motion semantics & geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 13864–13872. 2

[ZWZZ24] ZHANG J.-Q., WANG M., ZHANG F.-C., ZHANG F.-L.: Skinned motion retargeting with preservation of body part relationships. *IEEE Transactions on Visualization and Computer Graphics* (2024). 2