Semantics-aware Motion Retargeting with Vision-Language Models

Haodong Zhang[1*] Zhike Chen[1*] Haocheng Xu[1] Lei Hao[2] Xiaofei Wu[2]
Songcen Xu[2] Zhensong Zhang[2] Yue Wang[1] Rong Xiong[1†]
[1]Zhejiang University [2]Huawei Noah's Ark Lab

Abstract

*Capturing and preserving motion semantics is essential to motion retargeting between animation characters. However, most of the previous works neglect the semantic information or rely on human-designed joint-level representations. Here, we present a novel **S**emantics-aware **M**otion re**T**argeting (SMT) method with the advantage of vision-language models to extract and maintain meaningful motion semantics. We utilize a differentiable module to render 3D motions. Then the high-level motion semantics are incorporated into the motion retargeting process by feeding the vision-language model with the rendered images and aligning the extracted semantic embeddings. To ensure the preservation of fine-grained motion details and high-level semantics, we adopt a two-stage pipeline consisting of skeleton-aware pre-training and fine-tuning with semantics and geometry constraints. Experimental results show the effectiveness of the proposed method in producing high-quality motion retargeting results while accurately preserving motion semantics. Project page can be found at* https://sites.google.com/view/smtnet.

1. Introduction

3D animation characters have extensive application in animation production, virtual reality, and various other domains. These characters are animated using motion data, resulting in lifelike and immersive animations. Nevertheless, acquiring motion data for each character can be a costly endeavor. Therefore, the ability to retarget existing motion data for new characters holds immense importance. The goal of motion retargeting is to transfer existing motion data to new characters following motion feature extraction and integration processes, which ensure the preservation of the original motion's characteristics.

Semantics encompasses the meaningful and contextually relevant information conveyed in motion and plays a critical role in ensuring the realism and vividness of the animation characters. Preservation of motion semantics can enhance the efficiency of motion retargeting by reducing the need for time-consuming manual adjustments and refinements. However, previous methods [2, 15, 22] are mainly based on retargeting of joint positions and make less use of the extraction of semantic information. They focus on trajectory-level motion retargeting with few attention to motion semantics. Consequently, this leads to a significant loss of motion semantics and necessitates the labor-intensive intervention of animation artists for manual trajectory adjustments. Recent advancements have introduced self-contacts [23] and joint distance matrices [25] as the representation of motion semantics. Nevertheless, self-contacts are not applicable to non-contact semantics and require intricate vertex correspondence. The human-designed joint distance matrices primarily focus on joint relative relationships and still lack consideration of high-level semantic information.

To address the intricate task of capturing and preserving motion semantics, we introduce a new perspective: the most general and comprehensive form of motion semantics is human-level natural language, reflecting the user's intu-

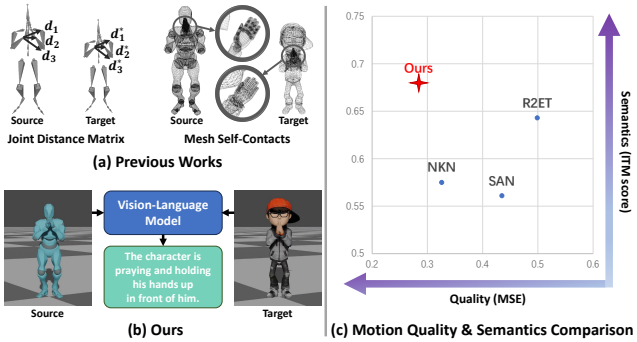

Figure 1. Comparison with previous motion retargeting methods. (a) Previous works rely on human-designed joint distance matrix [25] or self-contacts between mesh vertices [23] to ensure semantics preservation. (b) Ours work enforces human-level motion semantics consistency with the extensive knowledge of vision-language models. (c) Comparison of motion quality and semantics preservation on the Mixamo dataset [1]. Our method achieves the best motion quality and semantics consistency.

*These authors contributed equally to this work
†Corresponding author: rxiong@zju.edu.cn

itive understanding of motion. However, the main challenge of human-level motion semantics representation lies in the scarcity of labelled data. It is difficult and expensive to label sufficient semantic textual descriptions for motion data.

In this paper, we introduce the incorporation of robust, state-of-the-art vision-language models to provide semantic guidance to the motion retargeting network. In the absence of labelled semantic data, we leverage the capabilities of a vision-language model to serve as a semantic supervisor in an unsupervised manner, which can extract motion semantics in a more intuitive way, as illustrated in Fig. 1. This approach offers a solution to the challenge of the limited availability of labelled semantic datasets for motion retargeting. To establish a connection between the vision-language model and motion semantics extraction, we employ the differentiable skinning and rendering modules to translate 3D motions into image sequences. Subsequently, we adopt visual question answering with guiding questions to inquire about the most relevant motion semantics from the vision-language model.

To guarantee the preservation of motion semantics during motion retargeting, we introduce a semantics consistency loss that enforces the semantic embeddings of the retargeted motion to closely align with those of the source motion. For dense semantic supervision and computational efficiency, we utilize latent features extracted by the vision-language model as the semantic embeddings instead of textual descriptions. To alleviate the non-linearity of the semantics consistency loss, we introduce a two-stage training approach. We categorize motion information into two distinct levels: the skeletal level and the semantic level. Our approach involves pre-training the motion retargeting network at the skeletal level, which is then further refined and fine-tuned at the semantic level with the power of vision-language models. To the best of our knowledge, we are the first to leverage the extensive capability of vision-language models for the task of semantics-aware motion retargeting.

To summarize, the contributions of our work include:

- We introduce an innovative framework that leverages the expertise of vision-language models as a semantic supervisor to tackle the challenge of limited labelled semantic data for the task of motion retargeting.
- We propose to use differentiable skinning and rendering to translate from the motion domain to the image domain and perform guiding visual question answering to obtain human-level semantic representation.
- We design a semantics consistency loss to maintain motion semantics and introduce an effective two-stage training pipeline consisting of pre-training at the skeletal level and fine-tuning at the semantic level.
- Our model achieves state-of-the-art performance in the challenging task of semantics-aware motion retargeting, delivering exceptional performance marked by high-quality motion and superior semantics consistency.

2. Related Works

Optimization-based Motion Retargeting. Motion retargeting is a technique to adapt existing motion data from a source character to a target character with different bone proportions, mesh skins, and skeletal structures. Early works formulate motion retargetting as a constrained optimization problem [4, 6, 11, 18]. Gleicher *et al*. [6] introduced a motion retargeting method, which identifies motion features as constraints and computes an adapted motion using a space-time constraint solver to preserve the desirable qualities. Lee *et al*. [11] proposed a method to adapt existing motion of a human-like character to have the desired features with specified constraints and combined a hierarchical curve fitting technique with inverse kinematics. Nonetheless, these methods necessitate the tedious and time-consuming process of formulating human-designed constraints for specific motion sequences.

Learning-based Motion Retargeting. With the rise of deep learning, researchers have been developing learning-based motion retargeting methods in recent years [2, 9, 15, 22, 23, 25]. Villegas *et al*. [22] presented a recurrent neural network architecture, which incorporates a forward kinematics layer and cycle consistency loss for unsupervised motion retargetting. Aberman *et al*. [2] designed a skeleton-aware network with differentiable convolution, pooling, and unpooling operators to transform various homeomorphic skeletons into a primary skeleton for cross-structural motion retargeting. However, these methods tend to concentrate on trajectory-level motion retargeting with limited consideration for motion semantics, which often results in a notable loss of motion semantics and increase the heavy burden of manual adjustments to the trajectories. To address these problems, Zhang *et al*. [25] presented a residual retargeting network that uses a skeleton-aware module to preserve motion semantics and a shape-aware module to reduce interpenetration and contact missing. While this method successfully preserves joint relative relationships, it still falls short in addressing high-level motion semantics.

Vision-Language Models. Vision-language models have empowered various vision-language tasks, including visual question answering and image captioning. Tevet *et al*. [20] introduced a human motion generation model that aligns the latent space with that of the Contrastive Language-Image Pre-training (CLIP) model. Li *et al*. [13] proposed a pre-training strategy from off-the-shelf frozen pre-trained image encoders and frozen large language models for vision-to-language generative learning. Zhu *et al*. [27] presented a vision-language model, which uses one projection layer to align a frozen visual encoder with a frozen advanced large language models (LLM). However, these efforts primarily concentrate on vision-language tasks, leaving the question

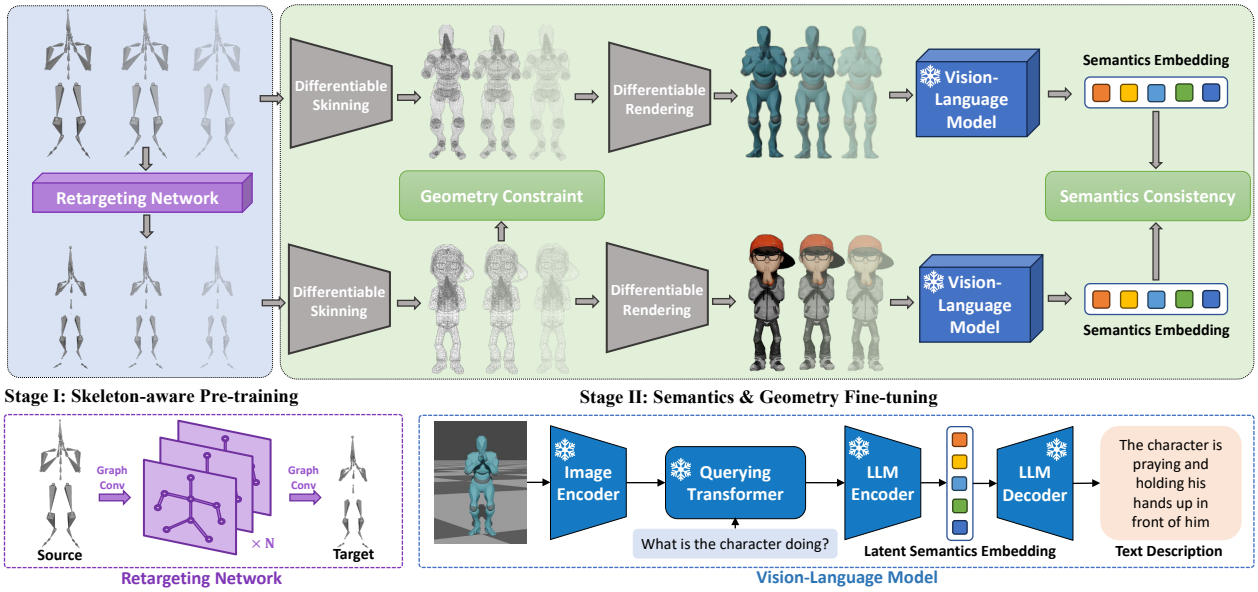

Figure 2. **Model Architecture.** Our semantics-aware motion retargeting framework employs a two-stage pipeline. Initially, the retargeting network consisting of multiple spatial-temporal graph convolution layers is trained at the skeletal level to establish a base model. Subsequently, this model undergoes further refinement and fine-tuning at the semantic level by the alignment of latent semantic embeddings of the source and target, leveraging the extensive knowledge of vision-language models. The latent semantic embedding is extracted by guiding visual question answering. Additionally, the geometry constraints are also enforced during fine-tuning to avoid interpenetration.

of how to effectively employ vision-language models to guide motion retargeting as an open and unexplored area.

Human motion synthesis. Human motion synthesis is a domain related to motion retargeting, which aims to synthesize realistic and lifelike human motions from random noise or other inputs with generative networks. Guo *et al*. [7] proposed to generate human motion sequences based on action type. Guo *et al*. [8] presented a temporal variational autoencoder to synthesize human motions from text input. Tevet *et al*. [21] introduced a diffusion-based generative model for human motion generation. As comparison, we focus on the task of motion retargeting, where existing motion data is transferred from a source character to a target character.

3. Method

3.1. Overview

We present a novel semantic-aware motion retargeting method, as illustrated in Fig 2. In contrast to previous methods that neglect motion semantics [2, 15, 22] or rely on human-designed joint-level representations [25], our approach integrates natural language descriptions from vision-language models to offer an explicit and comprehensive semantic representation of character motions, thereby maintaining the preservation of semantic consistency.

Task definition. Given a source motion sequence, consisting of the skeleton motion and its associated skinning geometry, as well as a target character in the reference pose (e.g., T-pose), the objective of motion retargeting is to generate the target motion while preserving crucial motion characteristics, such as joint trajectory similarity and motion semantics, and satisfying geometry constraints.

Graph representation. The skeleton motion sequence can be modelled as a sequence of graphs according to the skeleton hierarchy where each node corresponds to a joint and each edge represents a directed connection between joints. Assume that the motion sequence has T frames in total and the animation characters have N nodes and M edges. In our approach, we consider motion data as node features $\mathbf{Q} \in \mathbb{R}^{T\times N\times 9}$, which encompass the 6D joint rotation representation [26] and 3D joint positions. Additionally, we utilize skeleton hierarchy information as edge features $\mathbf{E} \in \mathbb{R}^{M\times 3}$, which consists of the 3D position offset between each joint and its parent joint.

Two-stage training. The motion of animation characters can be divided into skeletal movements and skinned movements, represented by skeletal joints and skinned vertices respectively. The skinned movements can be derived from the skeletal movements through the linear blend skinning algorithm [12]. Therefore, motion retargeting at the skeletal level can effectively downscale the data and reduce the complexity of the problem. However, this simplification process can lead to the loss of motion semantics and violations of geometry constraints. To address these issues, we employ a two-stage pipeline. Initially, we pre-train a skeleton-aware network to ensure a general initialization for motion retar-

geting without considering motion semantics and geometry constraints. Subsequently, we fine-tune the pre-trained network for each source-target character pair with the vision-language model to maintain semantic consistency and enforce geometry constraints to prevent interpenetrations.

3.2. Skeleton-aware Pre-training

Retargeting network. We propose a retargeting network consisting of a graph motion encoder and a graph motion decoder for motion retargeting. The motion encoder $\mathcal{F}_\theta$ encodes the motion data $\mathbf{Q}_A$ of the source character A into the latent motion embedding $\mathbf{Z}_A$. Then, the motion decoder $\mathcal{F}_\phi$ generates the joint angles $\mathbf{Q}_B$ of the target character B based on the latent features. Both the motion encoder and decoder are composed of multiple graph convolutions. More details are available in the supplementary materials.

$$\begin{aligned} \mathbf{Z}_A &= \mathcal{F}_\theta(\mathbf{Q}_A, \mathbf{E}_A) \\ \mathbf{Q}_B &= \mathcal{F}_\phi(\mathbf{Z}_A, \mathbf{E}_B) \end{aligned} \tag{1}$$

In the first phase, we train the motion encoder and decoder at the skeletal level to establish a robust initialization for motion retargeting. Following the unsupervised learning setting in [22], we train the network with the reconstruction loss, cycle consistency loss, adversarial loss, and joint relationship loss. The overall objective function for skeleton-aware pre-training is defined as follows:

$$\mathcal{L}_{skel} = \lambda_r \mathcal{L}_{rec} + \lambda_c \mathcal{L}_{cyc} + \lambda_a \mathcal{L}_{adv} + \lambda_j \mathcal{L}_{jdm} \tag{2}$$

The reconstruction loss $\mathcal{L}_{rec}$ encourages the retargeted motion to match the source motion when the target character is the same as the source character. Let $\mathbf{Q}_{A,t}$ be the motion data of source character A at frame t, and $\hat{\mathbf{Q}}^{rec}_{A,t}$ be the reconstructed motion. Then $\mathcal{L}_{rec}$ is defined as:

$$\mathcal{L}_{rec} = \sum_t \left|\left| \hat{\mathbf{Q}}^{rec}_{A,t} - \mathbf{Q}_{A,t} \right|\right|_2^2 \tag{3}$$

The cycle consistency loss $\mathcal{L}_{cyc}$ promotes the consistency of retargeted motion from the source character A to the target character B and then back to the source character A, ensuring it remains in line with the original motion. Let $\hat{\mathbf{Q}}^{cyc}_{A,t}$ represent the retargeted motion.Then $\mathcal{L}_{cyc}$ is defined as:

$$\mathcal{L}_{cyc} = \sum_t \left|\left| \hat{\mathbf{Q}}^{cyc}_{A,t} - \mathbf{Q}_{A,t} \right|\right|_2^2 \tag{4}$$

The adversarial loss $\mathcal{L}_{adv}$ is calculated by a discriminator network, which utilizes the unpaired data of the target character to learn how to distinguish whether the motions are real or fake. Let $\mathcal{F}_\gamma$ be the discriminator network, and $\mathbf{Q}_{B,t}$ be the retargeted motion at frame t. Then it is defined as:

$$\mathcal{L}_{adv} = \sum_t \log\left(1 - \mathcal{F}_\gamma\left(\mathbf{Q}_{B,t}\right)\right) \tag{5}$$

The joint relationship loss $\mathcal{L}_{jdm}$ is calculated by the joint distance matrix (JDM) $\mathbf{D} \in \mathbb{R}^{N\times N}$, which represents the relative positional relationships of the joints. The element $d_{i,j}$ of $\mathbf{D}$ represents the Euclidean distance between joint i and joint j. We extract the joint distance matrix from the target character and compare it with the source character. Then $\mathcal{L}_{jdm}$ is defined as:

$$\mathcal{L}_{jdm} = \sum_t ||\eta(\mathbf{D}_{A,t}) - \eta(\mathbf{D}_{B,t})||_2^2 \tag{6}$$

where $\eta(.)$ is an $\mathbf{L1}$ normalization performed on each row of the distance matrix. This normalization operation eliminates the difference in bone length to some extent.

3.3. Semantics & Geometry Fine-tuning

In the second phase, we fine-tune the pre-trained retargeting network for each source-target character pair to preserve motion semantics and satisfy geometry constraints. The motion semantics is maintained by the semantics consistency loss, which aligns the semantic embeddings extracted from a vision-language model for both the source and target. Additionally, the geometry constraint is satisfied by minimizing the interpenetration loss. The overall objective function for fine-tuning is outlined as follows:

$$\mathcal{L}_{fine} = \lambda_s \mathcal{L}_{sem} + \lambda_p \mathcal{L}_{pen} \tag{7}$$

Differentiable skinning & rendering. To make the fine-tuning process differentiable for gradient back-propagation, we first use the differentiable linear blend skinning algorithm [12], denoted as $\mathcal{F}_{lbs}$, to transform the target joint angles $\mathbf{Q}_B$ into skinned motions $\mathbf{V}_B$, represented by 3D mesh vertices. Subsequently, we employ the differentiable projection function $\mathcal{F}_{proj}$ as introduced in [16] to convert the skinned motions into 2D images $\mathbf{I}_B$. A limitation for the differentiable rendering process is that when projecting the 3D skinned mesh onto 2D images, the depth information is lost. To obtain a comprehensive semantic representation of the motion, we render the character from multiple perspectives and then combine the extracted features, following the Non-rigid Shape Fitting task in [16].

$$\begin{aligned} \mathbf{I}_A &= \mathcal{F}_{proj}(\mathcal{F}_{lbs}(\mathbf{Q}_A)) \\ \mathbf{I}_B &= \mathcal{F}_{proj}(\mathcal{F}_{lbs}(\mathbf{Q}_B)) \end{aligned} \tag{8}$$

Frozen vision-language model. To obtain an explicit and reliable semantic feature of the motion, we employ a frozen vision-language model as our semantic supervisor. Current 3D vision-language datasets [3, 28] mainly focus on the occupation or the segmentation of the object in a spatial scene like rooms, and thus the state-of-the-art 3D vision-language models [28] lack prior knowledge relevant to animation characters. In contrast, 2D vision-language models achieve better results in semantic tasks, such as image captioning, visual question answering and image-text

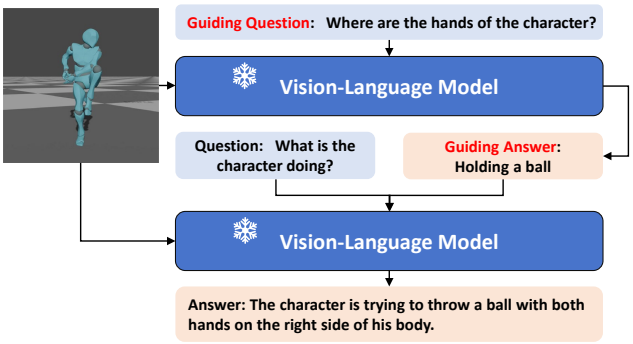

Figure 3. An example of guiding visual question answering.

retrieval, and provides cleaner and richer semantics [24]. Therefore, we utilize a frozen 2D vision-language model to extract latent embeddings of motion semantics. The frozen 2D vision-language model employed in our work is BLIP-2 [14], which incorporates a lightweight querying transformer as a bridge between the off-the-shelf frozen pre-trained image encoder and the frozen large language model.

Prompt design. Since the vision-language model has the capability to extract rich information from images, it is possible that the extracted features might contain redundant details, such as the appearance of the character. To guide the vision-language model to obtain semantic embedding relevant to character motions, we adopt a guiding visual question answering approach for motion semantics extraction, as depicted in Fig. 3. We believe that there is a strong correlation between motion semantics and hand movements. To acquire a more comprehensive description of the motion, we initially provide a guiding question to BLIP-2: *"Where are the hands of the character?"*. Subsequently, we introduce a new question and combine it with the first answer as the input to BLIP-2: *"[The answers to the first question generated by the vision-language model] What is the character in the image doing?"*. For more details, please refer to the supplementary materials.

Latent semantic embedding. We opt to align the latent semantic embeddings of the source and target generated by the vision-language model rather than relying on textual descriptions, specifically leveraging the encoder output of the large language model. This approach enables us to acquire a more accurate and denser representation, while also mitigating computational costs and the non-linearity of the training objective caused by the large number of parameters of the vision-language model. Let $\mathbf{E}_A$ and $\mathbf{E}_B$ be the latent semantic embeddings of the source and target motions, $\mathcal{F}_\omega$ be the frozen pre-trained image encoder, $\mathcal{F}_\sigma$ be the frozen querying transformer, $\mathcal{F}_\psi$ be the encoder of the frozen large language model, and $context$ be the question.

$$\begin{aligned}\mathbf{E}_A &= \mathcal{F}_\psi(\mathcal{F}_\sigma(\mathcal{F}_\omega(\mathbf{I}_A), context)) \\ \mathbf{E}_B &= \mathcal{F}_\psi(\mathcal{F}_\sigma(\mathcal{F}_\omega(\mathbf{I}_B), context))\end{aligned} \tag{9}$$

Fine-tuning with semantics consistency. As illustrated in Fig. 2, our approach aligns the latent semantic embeddings of both the source and target motions in an unsupervised manner, ensuring a high degree of semantic consistency in the retargeted results. The semantics consistency loss $\mathcal{L}_{sem}$ is calculated using the mean square error and it is defined as follows:

$$\mathcal{L}_{sem} = \sum_t \|\mathbf{E}_{A,t} - \mathbf{E}_{B,t}\|_2^2 \tag{10}$$

Fine-tuning with geometry constraints. From our observations, most interpenetration problems occur between the limbs and the main body. To address this, we incorporate the signed distance field between the limb vertices and the body mesh as the interpenetration loss. First, we convert the skeleton motion output from the network into mesh vertices using the linear blend skinning method [12]. Then, the interpenetration loss is defined as follows:

$$\mathcal{L}_{pen} = \sum_t ReLU(-\Phi_{b,t}(\mathbf{V}_{l,t})) \tag{11}$$

where Φ_b indicates the signed distance field function, $\mathbf{V}_l$ is the vertices of the limbs. If the vertex locates inside the body, the value of the function is less than zero. Therefore, we use the $ReLU$ function to penalize the inner vertices.

4. Experiments

4.1. Settings

Datasets. We train and evaluate our method on the Mixamo dataset [1], an extensive repository of animations performed by various 3D virtual characters with distinct skeletons and geometry shapes. The training set we use to pre-train our skeleton aware module is the same as that used in [2], which contains 1646 motions performed by 7 characters. It's important to note that the Mixamo dataset does not provide clean ground truth data, since many of the motion sequences suffer from interpenetration issues and semantic information loss. To mitigate this, we have carefully selected a subset of motion sequences that are both semantically clean and free of interpenetration issues for fine-tuning and testing. Our fine-tuning process involves retargeting 15 clean motions including 3127 frames, originally performed by 3 source characters, namely "Y Bot", "X Bot", and "Ortiz", onto 3 target characters, including "Aj", "Kaya", and "Mousey". Then we evaluate the performance of our model on the task of retargeting 30 additional motions that are previously unseen in the training set and fine-tuning sets. More details could be found in the supplementary materials.

Implementation details. The hyper-parameters λ_r, λ_c, λ_a, λ_j, λ_p, λ_s for pre-training and fine-tuning loss functions are set to $10.0, 1.0, 0.1, 1.0, 1.0, 0.1$. For semantics fine-tuning, we use BLIP-2 [14] with pre-trained FlanT5-XXL

[5] large language model. To extract the semantic representation of the motion, we render animation from three perspectives, including the front view, left view and right view. The fine-tuning process takes 25 epochs with 5 clean motion sequences of the source character for each target character. During pre-training and fine-tuning, we use an Adam optimizer to optimize the retargeting network. Please refer to the supplementary materials for more details.

Evaluation metrics. We evaluate the performance of our method across three key dimensions: skeleton, geometry, and semantics. At the skeletal level, we measure the Mean Square Error (MSE) between retargeted joint positions and the ground truth provided by Mixamo, analyzing both the global and the local joint positions. At the geometric level, we evaluate the interpenetration percentage (PEN). At the semantic level, we utilize the Image-Text Matching (ITM) score, Fréchet inception distance (FID) and semantics consistency loss (SCL) as metrics. The ITM score quantifies the visual-semantic similarity between the source textual description and the rendered retargeted motion. FID is calculated between the semantic embedding distribution of retargeted motion and source motion. More details are provided in the supplementary materials.

4.2. Comparison with State of the Arts

Quantitative. In this section, we conduct a comparative analysis of our method against the state-of-the-art approaches as illustrated in Tab. 1. The baseline methods include R2ET [25], SAN [2], NKN [22] and the Copy strategy. The Copy strategy achieves the lowest local MSE because the ground truth data in the Mixamo dataset are not entirely clean, and many of them are generated by copying rotations. As a result, this strategy comes at the cost of semantic loss and interpenetration issues. SAN [2] and NKN [22] focus on skeleton-level motion features, which results in a high interpenetration rate and relatively low semantics preservation. R2ET [25] treats motion semantics as the joint distance matrix and mesh distance field, which helps it obtain better motion semantics than SAN and Copy. Nevertheless, there is still a gap between the human-designed distance matrix and the human-level semantics. Notably, our model exhibits the best interpenetration rate and semantics preservation among all methods, showcasing the capability of the proposed method in producing high-quality retargeted motions with semantics consistency.

Qualitative. In Fig. 4, we visualize the text descriptions of the motions and the qualitative comparison between the state-of-the-arts and our method. SAN [2] and Copy neglect the preservation of semantics and have severe interpenetration. R2ET [25] utilizes joint distance matrix as semantics representation and fails to capture high-level semantic information. For example, the salute motion retargeted by R2ET [25] appears more like a hand-up motion. As a comparison, our method is able to successfully preserve high-level motion semantics leveraging the vision-language model. We observe that our approach reaches the best results among all methods, achieving more reliable semantics preservation and lower interpenetration rates. It suggests that with semantics and geometry fine-tuning, our method could effectively solve interpenetration issues together with semantics preservation.

Method	MSE ↓	MSE^{lc} ↓	Pen.% ↓	ITM ↑	FID ↓	SCL ↓
Source	-	-	4.43	0.796	-	-
GT	-	-	9.06	0.582	26.99	1.331
Copy	-	**0.005**	9.03	0.581	26.58	1.327
NKN [22]	0.326	0.231	8.71	0.575	27.79	1.414
SAN [2]	0.435	0.255	9.74	0.561	28.33	1.448
R2ET [25]	0.499	0.496	7.62	0.643	5.469	0.405
Ours	**0.284**	0.229	**3.50**	**0.680**	**0.436**	**0.143**

Table 1. Quantitative comparison with the state-of-the-arts. MSE^{lc} denotes the local MSE. ITM indicates the image-text matching score. FID is Fréchet inception distance of motion semantics. SCL is the semantics consistency loss.

Method	MSE ↓	MSE^{lc} ↓	Pen.% ↓	ITM ↑	FID ↓	SCL ↓
SMT_{tws}	**0.248**	**0.129**	8.37	0.586	7.727	0.769
SMT_{twf}	7.798	7.083	**0.44**	0.432	56.53	13.29
SMT_{twa}	0.335	0.288	5.36	0.658	2.826	0.266
SMT_{fwp}	0.439	0.368	1.22	0.597	7.241	0.583
SMT_{fwi}	5.418	4.576	4.41	0.552	78.46	18.96
SMT_{fwq}	0.739	0.517	4.56	0.668	2.497	0.191
SMT_{Ours}	0.284	0.229	3.50	**0.680**	**0.436**	**0.143**

Table 2. Ablation study. SMT_{tws} is the network trained with only skeleton-aware pre-training. SMT_{twf} is the network trained with only semantics and geometry fine-tuning. SMT_{twa} is the network trained in one stage. SMT_{fwp} is the network fine-tuned with only the interpenetration loss. SMT_{fwi} is the network fine-tuned with image features. SMT_{fwq} is the network fine-tuned with the features of the querying transformer.

4.3. Ablation Studies

Skeleton-aware pre-training. The proposed method can be divided into two stage: pre-training and fine-tuning. To illustrate the importance of skeleton-aware pre-training, we evaluate the network trained with only the semantics consistency loss and the interpenetration loss in Tab. 2, denoted as SMT_{twf}. The network trained without skeleton-aware pre-training performs worst in MSE and semantics preservation. A reasonable explanation is that the semantics consistency loss is highly non-linear, so it is important to pre-train the network at the skeletal level to provide better initial values. We also visualize qualitative results in Fig. 5.

Semantics & geometry fine-tuning. We also conduct ablation study to illustrate the importance of semantics and geometry fine-tuning in Tab. 2. We first evaluate the performance of the skeleton-aware model without fine-tuning, denoted as SMT_{tws}. Though it reaches the best global posi-

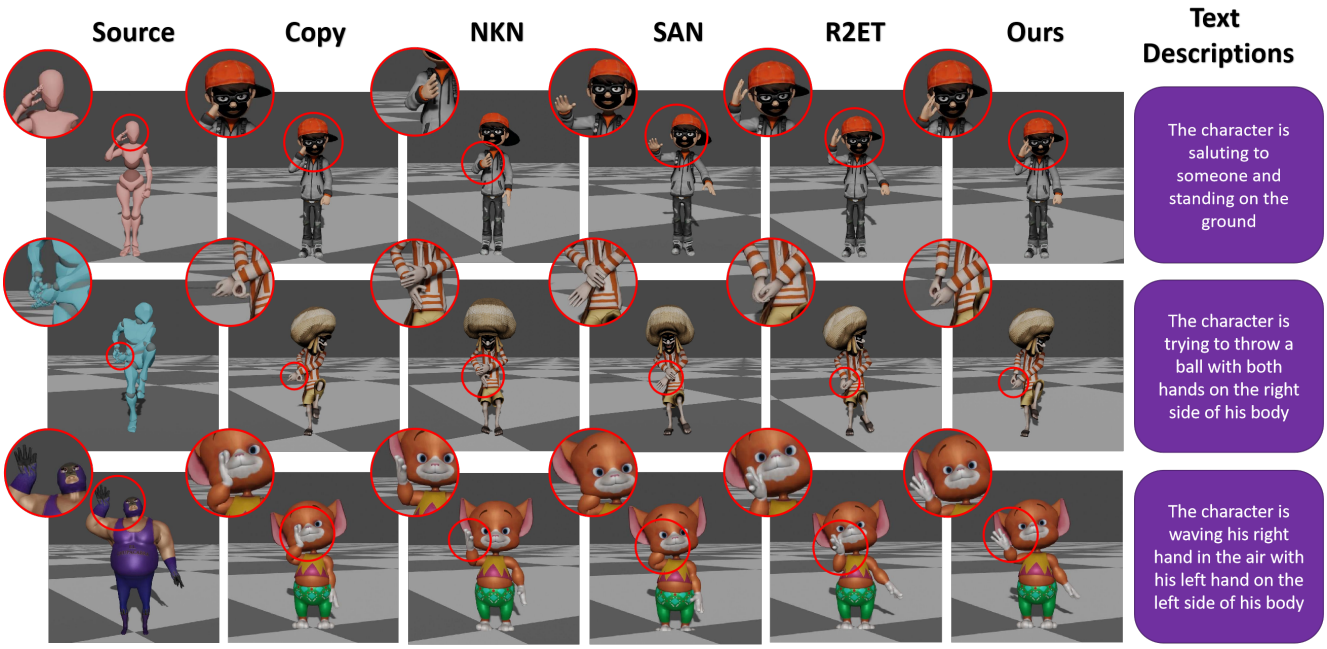

Figure 4. Qualitative comparision. The results demonstrate that our method can effectively preserve semantics while the baseline methods suffer from interpenetration or semantic information loss. From the first column to the last column are the source motion, the Copy strategy, NKN [22], SAN [2], R2ET [25], our method and text descriptions, respectively.

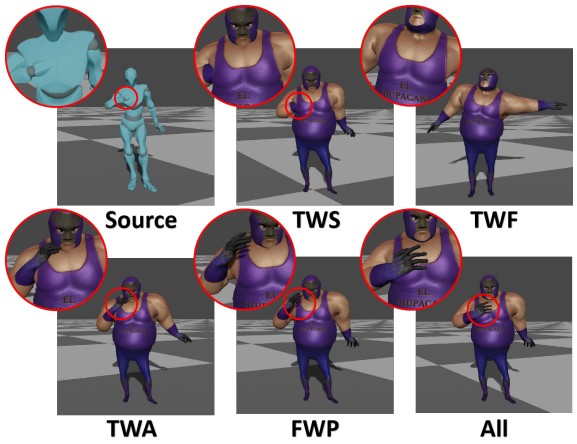

Figure 5. The qualitative comparison of ablation study between the network without fine-tuning (TWS), the network trained with only semantics and geometry fine-tuning (TWF), the network trained with all loss functions (TWA), the network fine-tuned with only the interpenetration loss (FWP) and our full model (All).

tion MSE, it suffers from interpenetration and semantic information loss because of the low-quality motion data provided by Mixamo. We next evaluate the network fine-tuned with only the interpenetration loss, denoted as SMT$_{fwp}$. This version results in a significant boost in terms of penetration rate. However, the gradient of interpenetration loss is only relevant with the face normals of the geometry mesh without considering the semantic information conveyed in the motion. It indicates the importance of the semantic consistency loss that makes the network reach a better balance between interpenetration and semantics. We also try to train the network with all loss functions in one stage, denoted as SMT$_{twa}$. However, it is challenging for the model to acquire general knowledge of interpenetration and semantics that is suitable for every character with limited data. Therefore, training the model with skeleton-aware pre-training and fine-tuning it with semantics consistency and geometry constraints for each target character remains a more reasonable and data-efficient strategy.

Latent semantic embedding. The vision-language model used for semantic extraction can be divided into three parts: the image encoder from CLIP [19], the querying transformer and the large language model. In Tab. 2, we compare the feature outputted by the image encoder, the querying transformer and the encoder of the large language model, denoted as SMT$_{fwi}$, SMT$_{fwq}$, and SMT$_{Ours}$, respectively. The results show that the image feature performs worse since it is greatly affected by the appearance of the character. It indicates that with the help of the large language model, the semantic representation better focuses on the semantic meaning of the motion instead of the character's visual appearance. Therefore, the encoder output of the large language model is more suitable for semantic embedding. More details can be found in the supplementary materials.

Prompt design. To validate the importance of guiding visual question answering, we compare the textual descriptions generated by visual question answering with and without guiding questions as well as image captioning. The re-

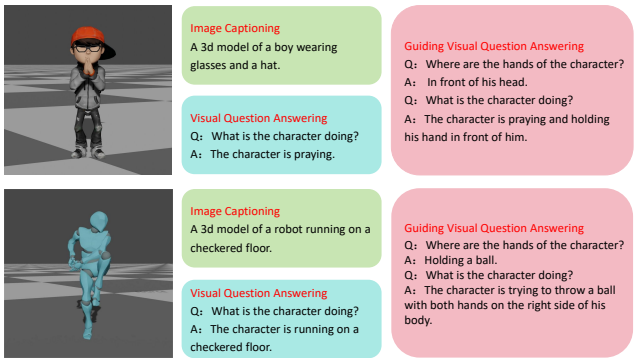

Figure 6. Text descriptions generated by different ways. The guiding visual question answering yields more comprehensive results.

Method	Quality ↑	Smoothness ↑	Semantics ↑
Copy	0.72	**0.86**	0.71
NKN [22]	0.65	0.80	0.66
SAN [2]	0.69	0.82	0.67
R2ET [25]	0.80	0.61	0.85
Ours	**0.89**	0.80	**0.92**

Table 3. User study results. We collect 100 comparisons in three aspects. Our method gets highest scores in the overall quality as well as semantics preservation.

sults in Fig. 6 indicate that using guiding questions for visual question answering yields the most comprehensive and reasonable text descriptions for motion semantics. Compared with image captioning that uses the vision-language model to generate text description directly from images, the answers from visual question answering task can be guided by the designed question to focus on motion semantics.

4.4. User Study

We conduct a user study to evaluate the performance of our method against the baseline methods. Human subjects are given 12 videos. Each video includes one source skinned motion and five anonymous skinned results. The retargeted results are randomly placed. We ask subjects to rate the results out of 1.0 in three aspects: overall quality, motion smoothness and semantics preservation. We collect a total of 100 comparisons. During the evaluation, users are required to extract semantic meaning from the source motion themselves and then evaluate the preservation of retargeted motions. In general, more than 92% of subjects prefer the retargeting results of our method.

4.5. Retargeting Motion from Human Videos

In this section, we evaluate our motion retargeting approach from human videos in the human3.6M [10] dataset. Video retargeting involves two stages: human pose estimation from video and motion retargeting. However, inaccuracies in estimating body postures may result in semantic information loss and thus accumulation of errors in the entire retargeting process. Therefore, we first get the estimated human pose from [17]. Then we utilize the vision-language model to extract the semantic embedding of the original video and calculate the semantic consistency loss to optimize the joint angles acquired from the retargeting process directly. In Fig. 7, we show our results of motion retargeting from human videos to Mixamo characters.

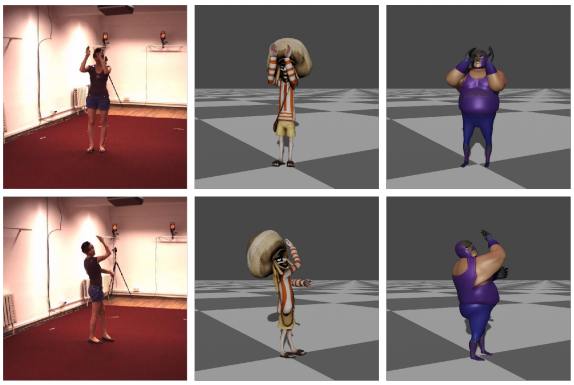

Figure 7. We retarget from human motion clips in the human3.6M [10] dataset. The retargeted motions are free from interpenetration and preserve semantics well.

5. Conclusions

In this paper, we present a novel semantics-aware motion retargeting method that leverages the capabilities of vision-language models to extract semantic embeddings and facilitate the preservation of motion semantics. This approach offers a promising solution to the challenge of lacking labelled semantic data for motion. Our proposed method involves a two-stage process that integrates skeleton-level motion characteristics and semantics-level consistency along with geometry constraints. Experimental results demonstrate that our approach excels in generating high-quality retargeted motions with semantics consistency.

Limitations. The main limitation is the performance of the vision-language model in extracting motion semantics. Without the support of motion semantic datasets of sufficient data size and quality, we rely on the model pre-trained on large image-text datasets. Although the model achieves some remarkable results in motion semantics extraction, there is still room for improvement. In addition, the projection of 3D motion into 2D images loses spatial information and affects the performance.

Future work. Compared with 2D vision-language models, 3D vision-language models have the advantage of capturing spatial relationships directly. Therefore, fine-tuning 3D vision-language models to make them more suitable for the task of motion semantics extraction is worth exploring in our future work.

Acknowledgements. This work was supported by the National Nature Science Foundation of China under Grant 62173293.

References

[1] Adobe's mixamo. https://www.mixamo.com/. Accessed: 2023-02-08.

[2] 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.

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[6] Michael Gleicher. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 33–42, 1998.

[7] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020.

[8] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5152–5161, 2022.

[9] 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*, pages 1–17, 2023.

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2014.

[11] Jehee Lee and Sung Yong Shin. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 39–48, 1999.

[12] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 165–172, 2000.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[15] Jongin Lim, Hyung Jin Chang, and Jin Young Choi. Pmnet: Learning of disentangled pose and movement for unsupervised motion retargeting. In *BMVC*, page 7, 2019.

[16] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7708–7717, 2019.

[17] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2022.

[18] Zoran Popović and Andrew Witkin. Physically based motion transformation. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 11–20, 1999.

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[20] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022.

[21] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[22] 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*, pages 8639–8648, 2018.

[23] 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*, pages 9720–9729, 2021.

[24] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *ICCV*, 2021.

[25] 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*, pages 13864–13872, 2023.

[26] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.

[27] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[28] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. *ICCV*, 2023.