POLITECNICO DI TORINO
Repository ISTITUZIONALE

Improving fidelity of close social interaction animations in social VR with a machine learning-based refinement framework

Original
Improving fidelity of close social interaction animations in social VR with a machine learning-based refinement framework / Visconti, A., Macaluso, R., Di Bartolomei, G., Calandra, D., Lamberti, F.. - ELETTRONICO. - 15915:(2026), pp. 204-219. (38th International Conference on Computer Animation and Social Agents (CASA 2025) Strasbourg (FR) June 2 - 4, 2025) [10.1007/978-981-95-0100-7_13].

Availability:
This version is available at: 11583/3001112 since: 2025-09-15T08:02:32Z

Publisher:
Springer

Published
DOI:10.1007/978-981-95-0100-7_13

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
Springer postprint/Author's Accepted Manuscript

This version of the article has been accepted for publication, after peer review (when applicable) and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/10.1007/978-981-95-0100-7_13

(Article begins on next page)

05 October 2026

Improving Fidelity of Close Social Interaction Animations in Social VR with a Machine Learning-based Refinement Framework

Alessandro Visconti[1][0000−0002−8323−9484], Roberta Macaluso[1][0009−0003−5340−8739], Gabriele Di Bartolomei[2], Davide Calandra[1][0000−0003−0449−5752], and Fabrizio Lamberti[1][0000−0001−7703−1372]

[1] Department of Control and Computer Engineering, Politecnico di Torino, Italy
{alessandro.visconti,roberta.macaluso,
davide.calandra,fabrizio.lamberti}@polito.it
[2] Politecnico di Torino, Italy gabriele.dibartolomei@gmail.com

Abstract. Social Virtual Reality platforms enable users to embody avatars and interact in virtual worlds. While research suggests that full-body avatar representations are generally preferred, high behavioral fidelity in avatar animations can be hard to achieve. Hardware-based tracking can be particularly effective but is costly, whereas Inverse Kinematics (IK) is more affordable but less accurate, leading to less realistic motion. Recent neural network-based approaches have shown promise in improving IK-based animations by predicting natural movements; however, guaranteeing high levels of fidelity in avatar-to-avatar interactions, particularly those involving close contact, remains challenging even with those approaches. With the aim to address such issue, this paper proposes a neural network-based refinement framework to enhance behavioral fidelity in close social interactions. To investigate its effectiveness, hugging has been selected as a use case. The framework, trained on motion capture data, has been evaluated via a user study, showing improved behavioral fidelity in avatar social interactions.

Keywords: Social VR, avatar-based interactions, hugs, machine learning.

1 Introduction

The rise of Social Virtual Reality (VR) platforms, driven by the increasing availability of affordable VR devices, has allowed users to embody avatars and interact in Virtual Environments (VEs) [16]. In these platforms, two primary factors contribute to the enhancement of realism and social presence: sensory fidelity, defined as the degree to which spatial, auditory, and haptic cues replicate real-world sensory experiences, and avatar fidelity, which encompasses both visual and behavioral dimensions. Specifically, visual fidelity refers to the degree to which avatar's appearance and movements resemble those of a human, while

behavioral fidelity refers to the degree to which avatar's actions, gestures, and social interactions authentically emulate human behavior [13].

High sensory fidelity can be achieved, for example, through haptic and tactile interfaces (e.g., vests and suits) that provide a pseudo-physical sense of embodiment [3]. As for visual fidelity, full-body representations are generally preferred in Social VR settings to support more natural and immersive interactions [7].

To achieve behavioral fidelity in full-body avatars, various methods have been proposed. Methods relying on hardware-based tracking deliver superior realism but necessitate complex and expensive setups [19]; Inverse Kinematics (IK), in turn, offers a more cost-effective solution but tends to produce less natural movements [1]. Recent advancements in machine learning offer a promising alternative, generating more realistic and smooth animations by leveraging large motion datasets [9].

Despite advances in full-body animation, achieving realistic interactions with physical elements in VEs remains a significant challenge [12]. In this context, several studies have focused on improving hand-object interactions [11,17], in some cases leveraging neural network-based approaches to resolve overlaps and ensure realistic poses [11]. These methods have improved the accuracy [28] and plausibility [22] of hand-object interactions without significantly increasing application latency. However, research on direct avatar-mediated human-to-human interactions involving physical contact, particularly those utilizing neural networks, remains limited. This complexity arises from the need to manage social interactions involving multiple avatars, where each participant dynamically influences the other's movements rather than merely responding to predefined animations [14]. This issue is especially evident in Social VR, where close-proximity interactions such as handshakes, high-fives, and hugs often appear unnatural, reducing the sense of immersion [10,14].

To address this challenge and considering the positive effects of machine learning reported in previous studies on hand-object interactions, this paper proposes a neural network-based framework designed to simultaneously model and refine avatar motion during social interactions in real-time, ensuring more natural and synchronized movements between multiple avatars. The framework consists of two main components: a module that employs a base IK system to reconstruct the avatar body movements, followed by a neural network module implemented as a multilayer perceptron (MLP); the network processes motion data from user-controlled avatars, refining their synchronized movements during social interactions. The network needs to be trained on a dataset containing avatar social interactions like those mentioned above, so that it can learn and apply motion refinements dynamically.

In this paper, hugs are used as a case study, as they are considered one of the most emotionally charged forms of physical contact between humans and convey feelings of warmth, affection, and social affiliation [3]. In particular, hugging someone in Social VR requires making a hug movement in the real-life world, which strengthens emotional engagement in relationship [6]. The case study scenario aims to replicate a typical situation observed in Social VR, where two

avatars perform a hug while being observed by external viewers within a VE. In this context, none of the users are physically co-located. For the training of the network, a dataset of avatar-mediated hugging interactions was utilized. At training time, motion capture data was used (following a preprocessing step, performed according to [9]), whereas at inference time the input consisted of animations generated using a base IK system. To assess the effectiveness of the proposed refinement framework, besides evaluating improvements in objective terms, a user study was also conducted, asking participants to subjectively assess improvements in behavioral fidelity by observing the animations of two avatars hugging in VR.

2 Background

The need to refine avatar animations primarily arises during physical-like interactions within VEs. When an avatar interacts with elements such as virtual objects [26], virtual agents [24], or avatars controlled by remote users [14], its movements should align with the ongoing physical action. However, the absence of a direct physical counterpart in the real world often causes inconsistencies between the movements of the avatar controlled by the user and the virtual elements in the environment, leading to unnatural and unrealistic animations [10]. Hence, avatar movements may need to be "refined".

2.1 Human-Object Interacton

One of the earliest studies in motion refinement, by Schmidl and Lin [20], introduced a geometry-driven physics approach that computes penetration depth and contact points when an avatar interacts with virtual elements. This information is used to adjust the avatar movements, resolving overlaps and ensuring realistic contact handling. Subsequently, Li et al. [11] introduced a grasp synthesis data-driven method using a shape-matching algorithm to align virtual hand shapes with objects. However, this approach was limited to single-hand interactions. Zhang et al. [28] extended this research by developing a bimanual interaction framework that uses a spatial representation combining object geometry with motion data. Their neural network-based approach demonstrated that incorporating object movement significantly improves the accuracy of hand-object interactions. A further step was taken by Taheri et al. [22] with GRIP, a learning-based model designed to generate realistic full-body grasping motions, synchronizing realistic motion for both hands before, during, and after object interaction based on predefined object interactions. These works highlight the importance of refining avatar motion to enhance the plausibility of interactions with virtual elements.

Besides hand-object interactions, some studies explored full-body motion refinement. Choi et al. [2] introduced a motion optimization technique for Mixed Reality environments, aimed at refining both locomotion and object interactions. Their approach included a penetration depth-based correction to minimize

hand and foot intersection artifacts. However, their method did not incorporate learning-based techniques, which could further enhance realism. Similarly, Vogt et al. [24] proposed an optimization-based framework for virtual agents, leveraging prerecorded motion capture data of human interactions. Their system adapted and refined agent movements in real-time, enabling more natural interactions between human users and virtual agents.

2.2 Human-to-Human Interacton

While the above studies have contributed to improving avatar realism in object interactions, research on direct avatar-to-avatar interactions remains scarce. Oh et al. [14] tackled this issue by developing an avatar animation technique for remote handshakes. Their system refined avatar movements to maintain hand contact between users, improving social telepresence. However, their approach lacked fine-grained finger contact modeling. To address this limitation, Lee et al. [10] designed a framework enabling multifinger contact interactions, using a database of human hand gestures to refine avatar animations. Although their method improved realism for hand interactions, it did not extend to encompass full-body animations, which are essential for complex social interactions.

Hence, despite reported advancements, close-contact interactions in Social VR like handshakes, high-fives, hugs, etc. remain a significant challenge. Existing approaches have yet to fully address the bidirectional influence between interacting elements, such as avatars: most of the current methods refine avatar motion in isolation, without accounting for the dynamic interplay between two bodies during interaction.

To cope with the above limitation, this paper presents a neural network-based framework for real-time motion refinement enabling synchronized and realistic avatar interactions. The proposed refinement framework utilizes a network trained on virtual social interaction data to dynamically refine avatar motion. By jointly modeling the movements of the two interacting avatars it enhances behavioral fidelity, treating each avatar as an interactive entity rather than an independent object. This approach represents a step forward in improving the realism of complex social behaviors in VEs, ultimately providing a way to enrich the sense of presence and user experience in Social VR.

3 Materials and Methods

This section introduces the proposed refinement framework, designed to improve user's social interactions within a VE. Moreover, it outlines the use case that was devised to evaluate it.

3.1 Framework

The proposed framework, illustrated in Figure 1, has been designed to be integrated into a Social VR platform accessible to users with consumer-grade VR

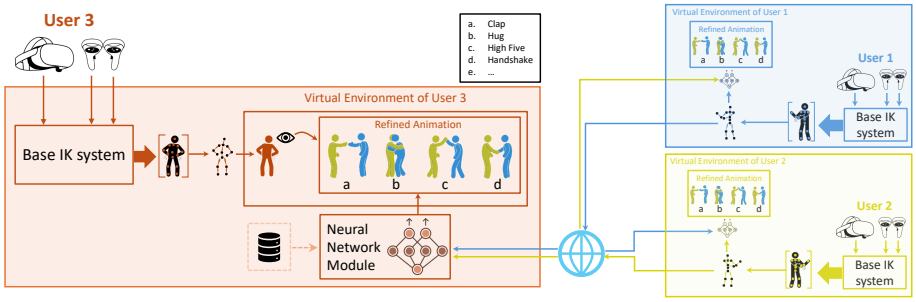


Fig. 1: Overview of the proposed refinement framework. In this case, User 3 acts as an observer, while User 1 and User 2 are engaging in a social interaction. The framework integrates a local neural network module for each user, which processes the structural information of the avatars and refines their animation by leveraging the output of a base IK system.

kits consisting of a headset and a pair of hand controllers. In the figure, User 3 is an external observer (the role played by the participants involved in the user study), while User 1 and User 2 are engaged in a social interaction.

When users wear the VR headset, the position and rotation of their head and hands are tracked. In a typical Social VR platform, this data is processed by an IK system that reconstructs the avatar's pose based on user input, providing a robust foundation for managing human motion. Consequently, a "base" IK system has been integrated into the proposed refinement framework, serving as a functional starting point for the second stage of the framework, which refines the motion. Although the base IK system module may introduce some errors and artifacts, it alleviates the need for the network to learn human motion from scratch, a highly complex task requiring also an extensive and well-prepared dataset. Thus, its integration helps mitigate the computational burden of neural network-based processing, which is performed locally on the user's machine. Furthermore, leveraging an existing IK system ensures adaptability, enabling the framework to integrate with various IK systems, rather than being limited to a specific one.

The second stage of the framework is the neural network module. The use of a neural network for pose refinement builds on the effectiveness of learning-based models and motion reconstruction methods. Neural networks offer high adaptability and the ability to learn complex motion patterns from data that would be impractical to encode manually. Additionally, running the neural network locally on each client rather than on a central server reduces computational load and enhances system scalability.

To limit computational load, the neural network module of a given user is activated only when proximity between two or more avatars is detected. Specifically, when proximity between User 1 and User 2 is detected, i.e. the distance between the users' hips is less than the so-called personal distance, set at 1.2

meters, and avatars fall in the respective field of view, their local neural network modules are activated. Similarly, for User 3 the module is activated when the proximity between User 1 and User 2 is detected and when both User 1 and User 2 are within User 3's field of view. However, the locally activated module does not affect the User 3's own avatar, but instead refines the movements of other users involved in the interaction. Thus, from the perspective of User 3, who acts as an observer, the module refines the motions of both User 1 and User 2, while User 3's avatar remains unaffected. When the local neural network module is active, the IK-based reconstruction is used to extract the position and rotation of each joint for User 1 and User 2, which are then leveraged to compute per-joint velocities and angular velocities following the methodology from a previous study [9]. These four per-joint parameters define a numerical pose representation for each frame.

These pose representations are then processed by each user's local neural network module, which has been trained on a dataset of close social interactions. In this way, the network can refine motion with greater behavioral fidelity. In the current implementation, the network used is a MLP, chosen for its simplicity, flexibility, and effectiveness in learning complex motion patterns even with a single hidden layer, as shown in previous studies [21]. Specifically, in this framework, an MLP is used to infer full-body joint positions and orientations from partial input data, i.e., the positions and orientations of the head and hands. The model is optimized to minimize the difference between its predictions and the reference poses provided in the dataset, with the aim to provide a more realistic and coherent full-body avatar motion. The local neural network module is designed to reduce overlaps between avatars' virtual bodies and to suppress unnatural movements, particularly in the arms and torso, thereby improving the overall plausibility of the social interaction in the VE.

The refined poses generated by the neural network module correct artifacts or unnatural movements introduced by the base IK system reconstruction of User 1 and User 2. Refined poses are applied to the avatars of User 1 and User 2, making the updates visible to both users. Additionally, refinements applied to users engaged in the interaction are also visible to any close observer, like User 3 in this case.

The framework operates independently of avatars' global positions and orientations. To this purpose, a master-slave relationship is established between interacting avatars. The slave avatar's root position and rotation are expressed in the local coordinate system relative to the master's root, whereas all the other joints are represented in local coordinates relative to their respective parent joints. Additionally, during training, the master's global root position is zeroed out to ensure that the network learns motion patterns independently of absolute world positioning. At inference time, the master's root position and rotation are retrieved from the base IK system output, maintaining consistency between training and inference phases.

In this setup with User 3 acting as an observer, its neural network module dynamically applies the master-slave distinction to refine the social interaction

between User 1 and User 2. Specifically, the avatar closest to the observer is identified as the master, while the other as the slave. Correspondingly, for the module running on each of the users directly engaged in the social interaction (User 1 and User 2 in this case), the master-slave distinction is applied such that the local user assumes the role of the master, while the other user is handled as the slave.

3.2 Use Case

As said, to evaluate the proposed refinement framework, hugs were considered as a use case. In the following, the use case is described by also making reference to all the technologies and settings used to implement, operate and assess the framework.

Hugging Among social interactions, hugging involve very close interplay of the users limbs and bodies, making it a particularly complex case. The two most common types of hugs identified in real-life, the "criss-cross" and "neck-waist" [4,5] hugs shown in Figure 2, were considered when collecting the dataset. Additional variability was introduced during acquisition by modifying the crossing side of the heads and the positioning of the arms, whether over or under.

(a) (b)

Fig. 2: The two types of hugs considered: (a) criss-cross, and (b) neck-waist.

Base IK System In this work, the Final IK Unity asset by RootMotion[3] was selected as base IK system. Final IK is an IK solution for full-body character animation, featuring a VR-oriented IK solver for natural movement and real-time adaptation to user input given by head and hand tracking. It uses IK for the upper body and animation blending to manage walking. It is widely used in studies involving virtual avatars [23,25].

[3] Final IK: https://tinyurl.com/rMotionFinalIK

Neural Network Module To train the MLP, stochastic gradient descent with a MSE loss function was employed. Various configurations in terms of hidden layer sizes, along with other hyperparameters such as learning rate, batch size, and momentum were evaluated to determine the optimal settings for the network.

Dataset The dataset was created using motion capture with the OptiTrack system[4], which has been used in previous studies [19]. The built-in re-targeting function from the OptiTrack streaming package for Unity was used to adapt the recorded movements to full-body avatars. In particular, full-body avatars downloaded from Mixamo[5] were used.

The dataset consists of the two most common types of hugs, i.e. "criss-cross" and "neck-waist". Specifically, each hug type was performed 29 times, with variations in hand, arm, and head positions between the two actors. Specifically, for each hug, the positions and rotations of both actors were recorded and then replicated in a mirrored way by switching the actors' movements in subsequent takes, ensuring complementary body positioning across all variations. This resulted in a total of 58 usable recordings, capturing different arm, hand, and head position combinations. The duration of each recording ranges from 5 to 7 seconds. To improve the robustness of the neural network, the dataset was further augmented by reversing the master-slave relationship for each hug sample, yielding 90 viable recordings. This augmentation ensured that the network can reconstruct the interaction regardless of which avatar assumes the master role.

Two simplifications were applied to the avatars for which motion data was collected. First, finger tracking was omitted, as finger movements are generally less relevant to the overall motion of a hug. Second, only avatars of the same height were considered; introducing height variations would have significantly increased the number of required recordings for each interaction type, with strong impacts on the data collection process.

For the data collection, the same two actors, both approximately 1.80 meters tall and represented by virtual avatars of matching height, were involved in all usable recordings. Rotations and angular velocities from the collected data were converted to a 6D representation which has been shown to be effective for training neural networks [9,15,29].

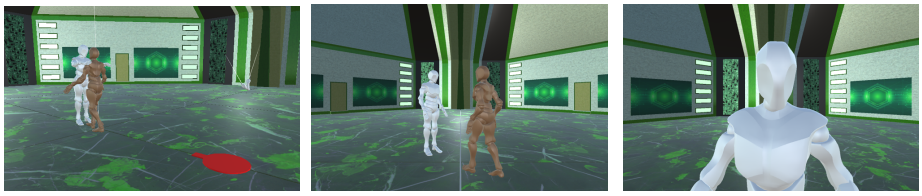
The resulting dataset was split into three subsets for training (60%), validation (20%) and testing (20%).

To train the model, the input consists of the position and orientation of a set of bones (e.g., head, neck, etc.) obtained using Final IK, which receives head and hand positions from the dataset recorded with the OptiTrack system. The ground truth, used for supervision during training, consists of the positions and orientations of the same set of bones directly recorded via OptiTrack. The model outputs the predicted positions and rotations of these bones, which are then applied to the 3D avatar model.

[4] OptiTrack: https://www.optitrack.com/
[5] Mixamo: https://www.mixamo.com/#/

4 Evaluation

This section presents the evaluations conducted to assess the proposed refinement framework in both objective and subjective terms. Specifically, the rotation error with respect to the ground truth for both the selected base IK system and the proposed refinement framework was calculated. Additionally, a user study was conducted to determine whether the framework effectively enhances behavioral fidelity compared to the base IK system.



(a) (b) (c)

Fig. 3: Position of the users and viewpoints considered in the user study: (a) position of the participant as an external observer (red marker), (b) his or her viewpoint as an external observer, and (c) viewpoint as one of the users engaged in the hug.

4.1 User Study

The user study was designed as a $2 \times 4 \times 10$ within-subjects experiment, with animation modality, type of hug, and point of view as independent variables, and involved 16 participants (12 males, 4 females) aged between 20 and 31 years. Participants were recruited from the student and staff population of the authors' university, volunteering as potential future users of Social VR. The VR experience for the user study was developed in Unity 2022.3, and the selected VR kit was a Meta Quest Pro.

The objective of the user study is to preliminarily investigate the potential effects on the visual perception of an external observer regarding the behavioral fidelity of hugging avatars in a VE, controlled by users who are not physically co-located. To this aim, two animation modalities were compared: a base IK system, specifically Final IK, and the proposed refinement framework, building onto Final IK.

At the beginning of the experiment, the participants filled in a demographic questionnaire covering age, gender, general experience with VR, and familiarity with multi-user Social VR. Most of the participants reported limited experience with VR technology, with approximately 50% rating their experience as 3 or lower on a 7-point scale (where 1 indicated almost no experience and 7 indicated

daily use). Experience with social VEs was even lower, with about 94% of the participants stating they had little to no prior engagement with Social VR.

Afterwards, the participants were invited to enter a VE resembling a typical environment of Social VR platforms[6]. Within the VE, they watched four types of pre-recorded hug animations (each lasting a maximum of ten seconds): criss-cross right, criss-cross left, neck-waist high, and neck-waist low. The use of pre-recorded animations ensured reproducibility and consistency across the participants. Avatars with abstract representations were intentionally used to focus the attention on behavioral fidelity rather than visual fidelity [18].

Each animation was presented ten times, eight times from the perspective of an external observer with a 45-degree incremental rotation to allow for multiple viewpoints, and two times from a front-facing perspective of the hugger (one from the master and one from the slave viewpoint). The position of the users and the various viewpoints are illustrated in Figure 3.

To balance exposure, a Latin square design was used to counterbalance both the order of hug animations and the sequence in which the two modalities were presented. After viewing an animation in a given modality, the participants answered questions aimed to assess the naturalness and realism of hug. They were then shown the same animation in the second modality and asked to respond to the same set of questions, with the option to revise their previous answers. This approach enabled a direct comparison between the two modalities. A video illustrating the various hug types in the two modalities as seen in the VE is available for download [7].

Finally, the participants were invited to provide open-ended feedback.

4.2 Metrics

Evaluation metrics are reported in the following.

Objective Measures As done in previous studies [9], the ability of the proposed framework to refine hugs compared to Final IK was measured in terms of Mean Per Joint Rotation Error (MPJRE), describing the mean rotation error of the joints in degrees.

Subjective Measures As said, subjective evaluation was conducted using questionnaires aimed to assess the behavioral fidelity of the animations generated by the proposed refinement framework and to determine potential improvements over the base IK system. Custom questions were specifically designed to ask the participants to judge arm movements in relation to the use case, focusing on the following aspects: whether the beginning and end of the hug were well-coordinated, the hug appeared as visually harmonious, the postures and movements of the avatars were well-aligned, and the movement of the arms and

[6] Sci-Fi Room: https://tinyurl.com/shifiRoom
[7] Video of hug animations: https://tinyurl.com/2ja4zp7z

hands appropriately adapted to the shape of the other body. These items were evaluated on a 7-point scale. Additionally, a question from [8] regarding perceived collisions (in terms of avatars' mesh interpenetration) was included. Finally, to evaluate realism, the anthropomorphism section of the Godspeed Questionnaire Series [27] was used. The full questionnaire is available for download[8].

4.3 Results and Discussion

Various hyperparameter configurations were explored to optimize the neural network module. The best results came with a 1024-size hidden layer, momentum of 0.9, learning rate of 0.1, batch size of 32, 2000 epochs, and a downsampling factor (used to select a subset of frames from a recorded hug to increase variability and reduce overfitting) of 10.

For what it concerns objective results obtained on the trained network, the proposed refinement framework applied to Final IK yielded a MPJRE of 4.86 degrees, whereas for Final IK alone MPJRE was equal to 17.83 degrees. This outcome demonstrates that the proposed framework was able to produce an improvement of about 13 degrees compared to Final IK in reproducing the ground truth poses.

For subjective results, the Shapiro-Wilk test was used to analyze normality of data. Since data was found to be normally distributed, T-test with a 5% threshold was used to identify significance differences (p-value $< .050$). For what it concerns custom questions (Figure 4), scores regarding coordination of the beginning and end of the hug, visual harmony of the interaction, and alignment of the avatars postures and movements did not show significant differences between the two modalities.

This outcome suggests that the proposed refinement framework does not degrade the harmony and fluidity of the animations, which are already well-refined in Final IK. A significant difference in favor of the proposed refinement framework, instead, was observed in the evaluation of how well the movement of the arms and hands adapted to the shape of the other body (3.84 vs 4.22, $p = .036$). This improvement is likely due to the neural network, which, having been trained on a dataset of high-fidelity hugging interactions, refines the motion to better conform to the shape of the other avatar's body. Another statistically significant difference was found in the perception of perceived collisions, with the participants reporting fewer instances of unnatural visual collisions when watching the animations generated by the proposed refinement framework (3.27 vs 3.56, $p = .028$). This finding further supports an improvement in the behavioral fidelity of the hug, likely attributable to the neural network's ability to better adapt the movements of both the avatars, avoiding collisions and interpenetration.

In contrast, when analyzing the answers to the Godspeed questionnaire (Figure 5) no significant differences were observed for most items, suggesting that

[8] Questionnaire: https://tinyurl.com/54nbsaxu

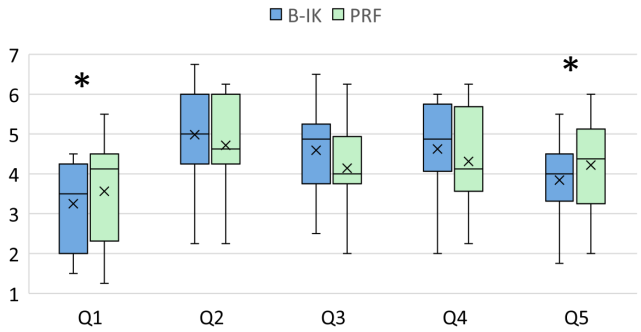


Fig. 4: Box plot showing the distribution of scores for the base IK system (B-IK) and the proposed refinement framework (PRF) in relation to custom questions assessing hug animations. Statistically significant results are marked with a * symbol. (Q1: I have never seen overlaps between the avatars; Q2: The beginning and end of the hug were well coordinated; Q3: The hug was visually harmonious; Q4: The postures and movements of the avatars were well aligned; Q5: The movement of the arms and hands adapted correctly to the shape of the other body).

the proposed refinement framework does not negatively impact the motion generated by Final IK. However, a notable exception was found in the assessment of the extent to which the movements appeared as human-like; in this case, animations generated by Final IK were perceived as more natural (4.52 vs 4.03, $p = .032$). This outcome could be attributed to the fact that, while the proposed refinement framework improves natural adaptation, the resulting movements may have lacked smoothness, appearing slightly rigid and contributing to a more mechanical perception.

5 Conclusions and Future Work

In this work, a neural network-based framework has been designed to refine avatar animations during social interactions in real-time. The framework consists of two main components: a base IK system to reconstruct the avatar's body movements, followed by a neural network module implemented as a MLP. The neural network processes motion data from avatars controlled by users, refining their synchronized movements during social interactions. To achieve this goal, the network can be trained on a dataset containing motion capture recordings of avatar-to-avatar interactions, allowing it to learn and apply realistic motion refinements.

To evaluate framework's effectiveness in enhancing behavioral fidelity, hugs were chosen as a use case, and a user study was conducted to preliminary investigate the potential effects on the behavioral fidelity of the hugging avatars based on the visual perception of an external observer. In the study, both objective

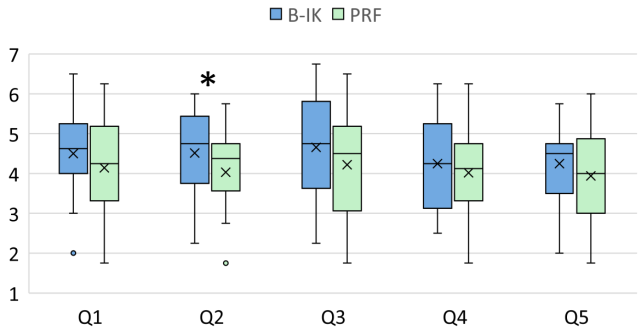


Fig. 5: Box plot showing the distribution of scores for the base IK system (B-IK) and the proposed refinement framework (PRF) in relation to items of the anthropomorphism section of the Godspeed questionnaire. Statistically significant results are marked with a * symbol. (Q1: Fake/Natural, Q2: Machinelike/HumanLike, Q3: Unconscious/ Conscious, Q4: Artificial/Lifelike, Q5: Moving rigidly/Moving elegantly.)

and subjective measures were collected. Objective results showed improvements in terms of MPJRE compared to a base IK system commonly used in VEs. Subjective results were gathered through the user study, in which participants evaluated two animations: one using the base IK system and the other using the proposed refinement framework. Subjective results indicated a reduction in perceived collisions and an improvement in animation harmony. However, participants noted a slight decrease in the humanlikeness of movements, likely due to the lack of smoothness in the refined movements despite the improvement in natural adaptation.

The study is affected by some limitations, including the limited sample, and suggests potential directions for future research. For instance, more advanced neural network models could be explored, balancing prediction accuracy and computational complexity. It would be interesting to investigate the latency introduced by the neural network, as it could potentially affect the perception of close-quarters interactions. The detection of social interactions used to trigger the neural network could be improved by incorporating classifiers capable of recognizing body gestures associated with specific interactions; this would enhance the scalability of the framework when handling multiple types of social interactions simultaneously, thereby reducing the risk of misinterpreting the users' intended behavior. In terms of scalability, it would be valuable to investigate the framework's ability to manage a large number of users connected at the same time. Future work could additionally focus on enhancing animation smoothness with motion filters to improve behavioral fidelity and evaluating the framework with a wider range of social interactions; this should also include investigating whether the refinement process enhances the sense of immersion. Building on that, new studies may investigate how the visual perception of an external ob-

server is influenced by the nature of the human-human interactions occurring in the Social VE; specifically, such studies could explore the differences between typically more polished but inherently repetitive pre-recorded interactions, and real-time interactions, which, although potentially less refined in appearance, offer uniqueness and spontaneity that may enhance the sense of authenticity.

A further direction for future development would be to investigate user perception from a first-person perspective by integrating the framework in interactive applications and allowing users to actively engage in social interactions; this would enable an assessment of the effects of such interactions from within the experience itself. Finally, the integration of haptic feedback systems or spatial audio technologies could be explored to further enhance sensory fidelity and enrich the overall sense of physical presence.

Acknowledgments. This work has been carried out in the frame of the VR@POLITO initiative.

References

1. Caserman, P., Achenbach, P., Göbel, S.: Analysis of inverse kinematics solutions for full-body reconstruction in virtual reality. In: Proc. 7th IEEE Int. Conference on Serious Games and Applications for Health. pp. 1–8 (2019)
2. Choi, S., Hong, S., Cho, K., Kim, C., Noh, J.: Online avatar motion adaptation to morphologically-similar spaces. Computer Graphics Forum **42**(2), 13–24 (2023). https://doi.org/https://doi.org/10.1111/cgf.14740
3. Cui, D., Kao, D., Mousas, C.: Toward understanding embodied human-virtual character interaction through virtual and tactile hugging. Computer Animation and Virtual Worlds **32**(3-4), e2009 (2021). https://doi.org/https://doi.org/10.1002/cav.2009
4. Dueren, A.L., Vafeiadou, A., Edgar, C., Banissy, M.J.: The influence of duration, arm crossing style, gender, and emotional closeness on hugging behaviour. Acta Psychologica **221**, 103441 (2021). https://doi.org/https://doi.org/10.1016/j.actpsy.2021.103441, https://www.sciencedirect.com/science/article/pii/S0001691821001918
5. Floyd, K.: All touches are not created equal: Effects of form and duration on observers' interpretations of an embrace. Journal of Nonverbal Behavior **23**(4), 283–299 (1999). https://doi.org/10.1023/A:1021602926270, https://doi.org/10.1023/A:1021602926270
6. Freeman, G., Acena, D.: Hugging from a distance: Building interpersonal relationships in social virtual reality. In: Proc. ACM Int. Conference on Interactive Media Experiences. p. 84–95 (2021). https://doi.org/10.1145/3452918.3458805
7. Freeman, G., Zamanifard, S., Maloney, D., Adkins, A.: My body, my avatar: How people perceive their avatars in social virtual reality. In: Proc. CHI Conference on Human Factors in Computing Systems – Extended Abstracts. pp. 1–8 (2020)
8. Hoyet, L., Olivier, A.H., Kulpa, R., Pettré, J.: Perceptual effect of shoulder motions on crowd animations. ACM Transactions on Graphics **35**(4) (2016). https://doi.org/10.1145/2897824.2925931, https://doi.org/10.1145/2897824.2925931
9. Jiang, J., Streli, P., Qiu, H., Fender, A., Laich, L., Snape, P., Holz, C.: Avatarposer: Articulated full-body pose tracking from sparse motion sensing. In: Proc. European Conference on Computer Vision. pp. 443–460 (2022)

10. Lee, Y., Lee, S., Lee, S.H.: Multifinger interaction between remote users in avatar-mediated telepresence. Computer Animation and Virtual Worlds **28**(3-4), e1778 (2017). https://doi.org/https://doi.org/10.1002/cav.1778
11. Li, Y., Fu, J.L., Pollard, N.S.: Data-driven grasp synthesis using shape matching and task-based pruning. IEEE Transactions on Visualization and Computer Graphics **13**(4), 732–747 (2007)
12. Mangalam, M., Oruganti, S., Buckingham, G., Borst, C.W.: Enhancing hand-object interactions in virtual reality for precision manual tasks. Virtual Reality **28**(4), 166 (2024)
13. Narayanan, S., Polys, N., Bukvic, I.I.: Cinemacraft: exploring fidelity cues in collaborative virtual world interactions. Virtual Reality **24**(1), 53–73 (2020)
14. Oh, J., Lee, Y., Kim, Y., Jin, T., Lee, S., Lee, S.H.: Hand contact between remote users through virtual avatars. In: Proc. 29th Int. Conference on Computer Animation and Social Agents. p. 97–100 (2016). https://doi.org/10.1145/2915926.2915947
15. Oreshkin, B.N., Bocquelet, F., Harvey, F.G., Raitt, B., Laflamme, D.: ProtoRes: Proto-residual network for pose authoring via learned IK. arXiv:2106.01981 (2022), https://arxiv.org/abs/2106.01981
16. Pratticò, F.G., Checo, I., Visconti, A., Simeone, A., Lamberti, F.: Designing hand-held controller-based handshake interaction in Social VR and Metaverse. In: Proc. 16th ACM SIGGRAPH Conference on Motion, Interaction and Games (2023). https://doi.org/10.1145/3623264.3624464, https://doi.org/10.1145/3623264.3624464
17. Rezzonico, S., Boulic, R., Huang, Z., Thalmann, N.M., Thalmann, D.: Consistent grasping in virtual environments based on the interactive grasping automata. In: Virtual Environments' 95: Selected papers of the Eurographics Workshops in Barcelona, Spain, 1993, and Monte Carlo, Monaco, 1995. pp. 107–118. Springer (1995)
18. Roth, D., Lugrin, J.L., Galakhov, D., Hofmann, A., Bente, G., Latoschik, M.E., Fuhrmann, A.: Avatar realism and social interaction quality in virtual reality. In: Proc. IEEE Virtual Reality. pp. 277–278 (2016). https://doi.org/10.1109/VR.2016.7504761
19. Roth, D., Waldow, K., Latoschik, M.E., Fuhrmann, A., Bente, G.: Socially immersive avatar-based communication. In: Proc. IEEE Virtual Reality. pp. 259–260 (2017). https://doi.org/10.1109/VR.2017.7892275
20. Schmidl, H., Lin, M.C.: Geometry-driven physical interaction between avatars and virtual environments. Computer Animation and Virtual Worlds **15**(3-4), 229–236 (2004)
21. Sharma, G., Chandra, S., Venkatraman, S., Mittal, A., Singh, V.: Artificial neural network in virtual reality: A survey. Int. Journal of Virtual Reality **15**, 44–52 (2016). https://doi.org/10.20870/IJVR.2016.15.2.2873
22. Taheri, O., Zhou, Y., Tzionas, D., Zhou, Y., Ceylan, D., Pirk, S., Black, M.J.: Grip: Generating interaction poses using spatial cues and latent consistency. In: Proc. Int. Conference on 3D Vision. pp. 933–943 (2024)
23. Visconti, A., Calandra, D., Lamberti, F.: Comparing technologies for conveying emotions through realistic avatars in virtual reality-based metaverse experiences. Computer Animation and Virtual Worlds **34**(3-4), 11 (2023). https://doi.org/https://doi.org/10.1002/cav.2188
24. Vogt, D., Grehl, S., Berger, E., Ben Amor, H., Jung, B.: A data-driven method for real-time character animation in human-agent interaction. In: Proc. 14th Int. Conference on Intelligent Virtual Agents. pp. 463–476 (2014)

25. Wagnerberger, L., Runde, D., Lafci, M.T., Przewozny, D., Bosse, S., Chojecki, P.: Inverse kinematics for full-body self representation in VR-based cognitive rehabilitation. In: Proc. IEEE Int. Symposium on Multimedia. pp. 123–129 (2021). https://doi.org/10.1109/ISM52913.2021.00029
26. Wan, W., Yang, L., Liu, L., Zhang, Z., Jia, R., Choi, Y.K., Pan, J., Theobalt, C., Komura, T., Wang, W.: Learn to predict how humans manipulate large-sized objects from interactive motions. IEEE Robotics and Automation Letters **7**(2), 4702–4709 (2022). https://doi.org/10.1109/LRA.2022.3151614
27. Weiss, A., Bartneck, C.: Meta analysis of the usage of the godspeed questionnaire series. In: Proc. 24th IEEE Int. Symposium on Robot and Human Interactive Communication. pp. 381–388 (2015). https://doi.org/10.1109/ROMAN.2015.7333568
28. Zhang, H., Ye, Y., Shiratori, T., Komura, T.: ManipNet: Neural manipulation synthesis with a hand-object spatial representation. ACM Transactions on Graphics **40**(4) (2021). https://doi.org/10.1145/3450626.3459830, https://doi.org/10.1145/3450626.3459830
29. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)