PERSONALITY EXPRESSION IN THREE-DIMENSIONAL HUMAN-OBJECT INTERACTION SEQUENCES

A DISSERTATION SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

By
Yalım Doğan
September 2025

PERSONALITY EXPRESSION IN THREE-DIMENSIONAL HUMAN-OBJECT INTERACTION SEQUENCES

By Yalım Doğan

September 2025

We certify that we have read this dissertation and that in our opinion it is fully adequate, in scope and in quality, as a dissertation for the degree of Doctor of Philosophy.

Uğur Güdükbay(Advisor)

Özgür Ulusoy

Yusuf Sahillioğlu

Ramazan Gökberk Cinbiş

Hamdi Dibeklioğlu

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

PERSONALITY EXPRESSION IN THREE-DIMENSIONAL HUMAN-OBJECT INTERACTION SEQUENCES

Yalım Doğan
Ph.D. in Computer Engineering
Advisor: Uğur Güdükbay
September 2025

Personality expression is an essential component of communication between virtual characters and the viewers. Human motion, encompassing temporal facial expressions, body movements, and posture, conveys personalities while controlled by high-level parameters, such as the OCEAN personality model. Research in personality expression, however, focuses solely on isolated behavior, where the subjects don't interact with their environment in a physical sense. Even though object interaction is not a social context, we investigate whether the aforementioned features regarding human motion can be used to communicate the personality of the subjects. Arbitrary object types, actions, and multiple iterations of such interactions are subject to differences in perceived expression, which we quantitatively and qualitatively analyze and synthesize in this work. We applied crowdsourcing to annotate our custom subset of a three-dimensional object interaction dataset and discuss the perceived personalities in multiple object categories. Then, we developed a personality-aware motion augmentation framework to increase the scale and diversity of our dataset. The dataset is used to train a neural motion field-based network architecture to alter a given motion's expressed personality, controlled via OCEAN factors. We validated our approach with a separate user study to assess the resulting motions' personality, accuracy, and realism. We compared the performance of synthetic motions from our neural network and augmentation-only motions. Results suggest that while augmentation-only motions are better at differentiating positive and negative traits of personality, their realism and semantic accuracy are lower than neural-based motions. Neural-based motions' control over the personality via a single factor is observed to be subtle, but multiple factors display a more noticeable difference.

Keywords: Personality expression, human-object interaction, Five Factor Model, Laban Movement Analysis, neural motion synthesis, user study.

ÖZET

ÜÇ BOYUTLU İNSAN-NESNE ETKİLEŞİM ANİMASYONLARINDA KİŞİLİK İFADESİ

Yalım Doğan
Bilgisayar Mühendisliği, Doktora
Tez Danışmanı: Uğur Güdükbay
Eylül 2025

Kişilik ifadesi, sanal karakterler ve izleyiciler arasındaki iletişimin temel bir bileşenidir. Zamansal yüz ifadeleri, vücut hareketleri ve duruşu kapsayan insan hareketi, OCEAN kişilik modeli gibi üst düzey parametreler tarafından kontrol edilirken kişilikleri yansıtır. Ancak kişilik ifadesi araştırmaları, yalnızca öznelerin çevreleriyle fiziksel anlamda etkileşime girmediği izole davranışlara odaklanır. Nesne etkileşimi sosyal bir bağlam olmasa da, insan hareketine ilişkin yukarıda belirtilen özelliklerin öznelerin kişiliğini iletmek için kullanılıp kullanılamayacağını araştırıyoruz. Rastgele nesne türleri, eylemler ve bu tür etkileşimlerin birden fazla yinelemesi, algılanan ifadedeki farklılıklara tabidir ve bu çalışmada bunları nicel ve nitel olarak analiz edip sentezliyoruz. Üç boyutlu bir nesne etkileşimi veri kümesinin özel alt kümesini açıklamak ve birden fazla nesne kategorisinde algılanan kişilikleri tartışmak için kitle kaynak kullanımını uyguladık. Ardından, veri kümesinin ölçeğini ve çeşitliliğini artırmak için kişiliğe duyarlı bir hareket artırma çerçevesi geliştirdik. Veri kümesi, OCEAN faktörleri aracılığıyla kontrol edilen belirli bir hareketin ifade edilen kişiliğini değiştirmek için sinirsel hareket alanı tabanlı bir ağ mimarisini eğitmek için kullanılır. Yaklaşımımızı, ortaya çıkan hareketlerin kişiliğini, doğruluğunu ve gerçekçiliğini değerlendirmek için ayrı bir kullanıcı çalışmasıyla doğruladık. Sinir ağımızdan elde edilen sentetik hareketlerin ve yalnızca artırma içeren hareketlerin performansını karşılaştırdık. Sonuçlar, yalnızca artırma içeren hareketlerin kişiliğin olumlu ve olumsuz özelliklerini ayırt etmede daha iyi olduğunu, ancak gerçekçiliklerinin ve anlamsal doğruluklarının sinir tabanlı hareketlerden daha düşük olduğunu göstermektedir. Sinir ağ tabanlı hareketlerin tek bir faktör aracılığıyla kişilik üzerindeki kontrolünün hafif olduğu, ancak birden fazla faktör olduğunda daha belirgin bir fark olduğu gözlemlenmiştir.

Anahtar sözcükler: Kişilik ifadesi, insan-nesne etkileşimi, Beş Faktör Modeli, Laban Hareket Analizi, sinirsel hareket sentezi, kullanıcı çalışması.

Acknowledgement

First of all, I want to thank my supervisor sincerely, Prof. Dr. Uğur Güdükbay, for guiding and shaping me for nearly a decade in the academic environment. He introduced me to computer graphics, which expanded to multiple areas as our collaboration progressed. He was always open to my ideas, encouraged and supported me to explore them further. I hope to collaborate with him for even more inspiring projects.

I want to express my sincere gratitude to my thesis monitoring committee, Prof. Dr. Yusuf Sahillioğlu and Asst. Prof. Dr. Hamdi Dibeklioğlu for providing insightful comments that kept my progress on track. I would also like to thank the jury members, Prof. Dr. Özgür Ulusoy and Assoc. Prof. Dr. Ramazan Gökberk Cinbiş, for their insightful comments and valuable suggestions.

Secondly, I would like to thank my research group members for their extensive support, collaboration, and inspiration via fruitful discussions: Serkan Demirci, Sinan Sonlu, Arçin Ülkü Ergüzen, and Musa Ege Ünalan. This thesis and the studies that led to it would not have been possible without their enthusiastic involvement. I thank them individually and look forward to working with them again.

Through my studies, for providing me support, guidance, and freedom, I would like to thank my current and past colleagues and friends at TÜBİTAK BİLGEM İLTAREN. I want to extend my special thanks to Aydın and Tülin, who not only encouraged me to complete my studies but also provided the help I needed most. "*Yes, it finished training, and I finished my thesis.*" I want to thank all my friends from everywhere for sharing great memories throughout this thesis.

I cannot thank my family enough, especially my parents, Nurcan and Oktay, for constantly supporting and being there for me in the most difficult times. I sincerely thank my grandparents and the rest of my family who continually checked on me during my studies, and I hope to make them proud in the upcoming years.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Thanks to its advancements over multiple decades, computer graphics technology enables the visualization of curated environments both realistically and artistically in a scalable fashion. In entertainment, science, and education, visualizing three-dimensional scenes conveyed complex geometric information to viewers by revealing novel perspectives. Additionally, dynamic environments where scene components perform rigid and non-rigid alternations increase immersion. A pivotal point for realism came with animating characters in the scene, where their behavior is consistent with their real-life counterparts. For instance, animating human behavior made it possible to insert imaginary characters to communicate complex interactions with the viewer in the sense of emotions and complex interactions with the environment.

Representing human motion is based on understanding the human skeletal and facial structure, where inaccuracies can limit the character's movement or embody eerie behavior. Any imperfections involving facial expressions, body posture, and limb movement are easily noticeable due to implicit expectations of a human's behavior. The concept of the uncanny valley [9] explains this phenomenon, where problems with the proportions and movement of a humanoid character become increasingly prevalent with attempted realism. In addition to expectations from human movement, scene components imply additional constraints; a character

who interacts with other characters and objects would emphasize its presence. Characters who grasp, hold on to, and carry objects need to abide by the object's properties and capabilities. For instance, a heavy rock is not expected to be lifted with a single finger, or a person is not likely to go through a door without opening it in realistic scenarios.

Affordance [10] regarding the interacted objects, study the design and ways of making such interactions comfortable and possible for humans. For instance, a mug is usually associated with a "hold" action. In contrast, a pair of scissors is associated with "cut." Respecting the constraints and properties of the object, subject, and the scene's context is crucial in generating immersive sequences. However, manual authoring of hands is challenging for such scenarios because of contact accuracy and realistic finger grasps. It becomes even harder to control such contacts if the object is articulated, such as the cap on a bottle.

Recent studies in robotics, computer animation, and computer vision have focused on generating scene interactions involving manipulating the objects in the scene. Such systems focus on the agent's movement during the interaction and the resulting trajectory of the object. Recent studies have shifted focus to articulating the fingers of the subjects to reflect proper grasping behavior, which is especially crucial for robotics. Such articulations are also subject to the affordance properties of the object and the subject. Previously perceived personalities of the characters are also expected to influence these interactions, and reactions from the scene are also expected to incite emotions in the characters themselves. Otherwise, the characters can seem "robotic" where they solely perform their assigned tasks without an internal state which gives them the "character" after all, as seen in Figure 1.1. For instance, the trajectory of a hot plate moved by a person who takes high caution is expected to differ from that of a clumsy one, even though their goals are identical. The result of the interaction would dictate whether the person has succeeded in their intended purpose. Additionally, when unexpected events occur, such as dropping or spilling, the person is expected to respond naturally based on the triggered emotion. In literature, these aspects of object interaction regarding computer animation or robotics are usually overlooked, mainly focusing on achieving the desired goal via the optimal path.

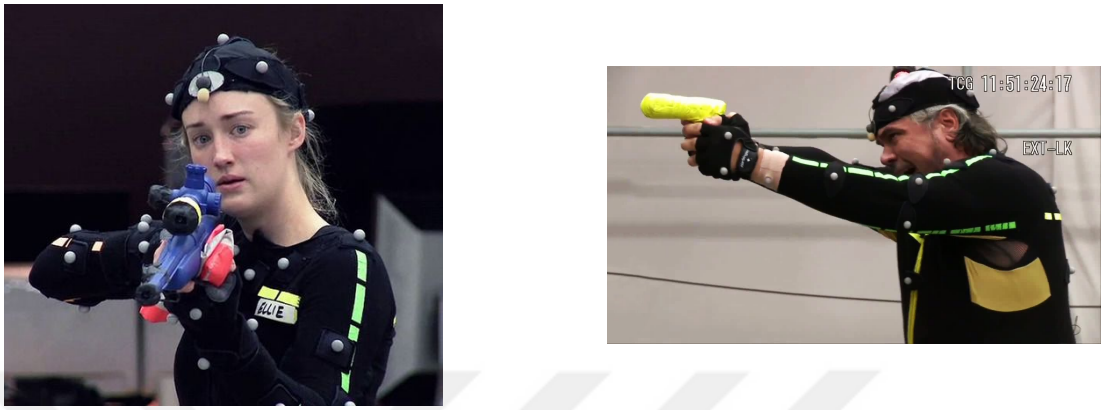

Figure 1.1: Screenshots from performance capture session for *The Last of Us* [1] by Naughty Dog [2]. Both characters, with different personalities, exhibit distinct postures and facial expressions while pointing a weapon. Even without facial expressions, the female character's positioning of the gun while looking forward indicates a "concerned" emotion with an "agreeable" personality. However, the male character grasps and points the gun with high "confidence" which communicates "anger." *The Last of Us* is created and developed by Naughty Dog.

We explore various aspects of a three-dimensional object interaction sequence to explain how they contribute to the expressed personality of the person. Understanding these aspects would benefit the realistic behavior of virtual agents and robots by incorporating natural human behavior and environmental errors. We introduce a neural model for personality manipulation of object interaction motions and use the Five Factor Model for personality representation, which is also known by the acronym of its five orthogonal factors: OCEAN [11]. Our Neural Motion Field (NeMF) [12]-based generator architecture inputs the object properties and target OCEAN factors to transform the input motion to express the desired personality traits. Training of such a generator module is achieved using an adversarial scheme, where we utilize a critic module to assess realism of the synthesized motion and a regressor module to ensure the resulting motion consistently represents the intended personality and object semantics. Since expressive human-object interaction datasets lack personality annotations, we introduce them to a customized 3D object interaction dataset. We utilize our personality-based motion augmentation framework to increase the sample size and diversity, improving our training performance. We also performed user studies to annotate our dataset and validate the feasibility of our approach.

1.1 Contributions

The contributions of the thesis are as follows:

- We examine multiple object interaction datasets and their shortcomings. We then introduce our subset of object interaction sequences based on a well-known dataset.
- Using crowdsourcing, we annotate animations in our dataset using the OCEAN personality model.
- We introduce our object and body-aware motion augmentation, which utilizes Laban Movement Analysis (LMA) to introduce controlled variance to object interaction motion to increase the training sample size. This framework also adjusts the resulting personality after augmentation, making it personality-aware.
- Using the augmented dataset, we train our neural motion-based generative adversarial network to author the personalities in the provided motion. The network also contains modules for inferring the existing personality, object type, and action for the provided motions. We share the code of our augmentation tool and network at: `https://github.com/YalimD/ObjectInteractionPersonalityAuthoring`.
- We conduct additional user studies to compare the realism and expressiveness of augmentation and network-based motions. We alter each personality factor separately to isolate its effects on the motion style.
- We quantitatively analyze the performance of our neural approach by performing ablation studies. We also demonstrate its effectiveness in perceiving personality when using multiple OCEAN factors.

1.2 Publications

The following article is published within the context of this thesis:

- *Exploring Personality in Human-Object Interactions* [13]. This work forms the basis of the thesis. This work is published in *Signal, Image and Video Processing* in 2025.

The following article and conference proceedings papers that I contributed to are indirectly related to this research and were published during this thesis study.

- *Personality Perception in Human Videos Altered by Motion Transfer Networks* [14]: This work was published in journal *Computers & Graphics* in 2024.
- *Human Movement Personality Detection Parameters* [15]: This work was published in *Proceedings of the IEEE Signal Processing and Communications Applications* in 2024.
- *Towards Understanding Personality Expression via Body Motion* [16]: This work was published in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, Workshop on Multi-modal Affective and Social Behavior Analysis and Synthesis in Extended Reality* in 2024.

1.3 Organization of the Thesis

The thesis structure is as follows:

- In **Chapter 2**, we provide background information on motion description, its capture, and object capture to cover reconstruction approaches. We then explore recent learning-based motion generation and control approaches and finalize by describing personality and associated factors.
- **Chapter 3** includes two-dimensional and three-dimensional object interaction datasets and compares them based on context and scale. We conclude with our custom dataset, a subset of one of the listed datasets.
- **Chapter 4** describes our augmentation framework for scaling our limited dataset to improve training performance and introduce subtle diversity. We formulate our approach and provide visual examples.
- During **Chapter 5**, we explain our motion authoring framework based on neural networks. We discuss each component and loss related to our model and share our training details.
- In **Chapter 6**, we demonstrate our user study program and annotation, comparison protocol.
- **Chapter 7** includes quantitative analysis of our annotations and comparison results. Additionally, we perform ablation studies for our neural network model to investigate its robustness. We also visualize the authoring performance of our network using single and multiple factors.
- **Chapter 8** concludes the thesis with possible future research directions.

Chapter 2

Background and Related Work

This chapter focuses on the background and related work of motion representation, its reconstruction, motion generation techniques, and personality representation. The datasets regarding object interaction and their categorization are investigated in Chapter 3.

2.1 Motion Representation and Reconstruction

2.1.1 Human Motion Representation

Human motion in animation is represented as a graph of joints that rotate at time t during the sequence. Each t where a joint has a determined orientation is called a "keyframe". While keyframes can exist at all animation frames, the authors can omit in-between frames and allow the system to "interpolate" between orientation changes. Usually, linear interpolations can take many forms, including polynomial functions, to represent realistic and complex movements. The number of keyframes during the animation sequence determines the level of detail, as more keyframes enable smoother animation, especially with high-frequency motions such as fast finger movements.

The joint's orientation is stored in a selected rotation format at each keyframe. A joint rotation can be represented in terms of Euler angles [17], Quaternions [18], and Axis-Angle form [19]. Since Euler angles can cause gimbal lock, which involves losing a rotation axis due to rotating one axis to 90 degrees, other rotation formats are usually preferred. For simplicity, each joint stores its rotation in its respective axes. As the human skeleton maintains connection during movement, the translation component for the joints is omitted, except for the root joint of the body, which is usually the pelvis. Pelvis joint translation, which is represented as a 2D or 3D vector, moves the body around the scene, giving the notion of traversal. The global position of each joint according to the root joint is solved using the Forward Kinematics (FK) process [20]. FK starts from the root joint and hierarchically combines the rotations at each child joint. Then, if any, the root translation is added to all joints once. Examples of joint hierarchies are demonstrated in Figure 2.1. There are arbitrary motion representation formats available in the industry, which will be discussed further in the following chapters.

SMPL [23] is a skinned, fully rigged human model including articulated face and hands. It is parametric regarding joint orientations and body shape: $M(\beta, \theta, \psi)$. β is for pose, θ for shape and ψ for facial expressions. The model is trained on numerous 3D scans from a diverse range of individuals, varying in proportions. The average model, called the template, is modified using principal components of human body shapes. The model also handles self-collisions and invalid poses using an interpenetration penalty and pose priors.

The hand (MANO [24]) and face (FLAME [3]) models are later incorporated into the model, which is named SMPL-X [22]. Like SMPL [23], the MANO [24] model includes hands captured in high detail, and can handle arbitrary finger articulations. An example model can be seen in Figure 2.1, right. The hand and face are modeled similarly: shape and pose priors are calculated beforehand to obtain plausible skinning and articulations. A skinned model like this is crucial for synthesizing contacts with objects, as humans interact with them using their skin, not joints. This representation can be used to penalize penetrations of the object during contacts and realistically correct the grasp.

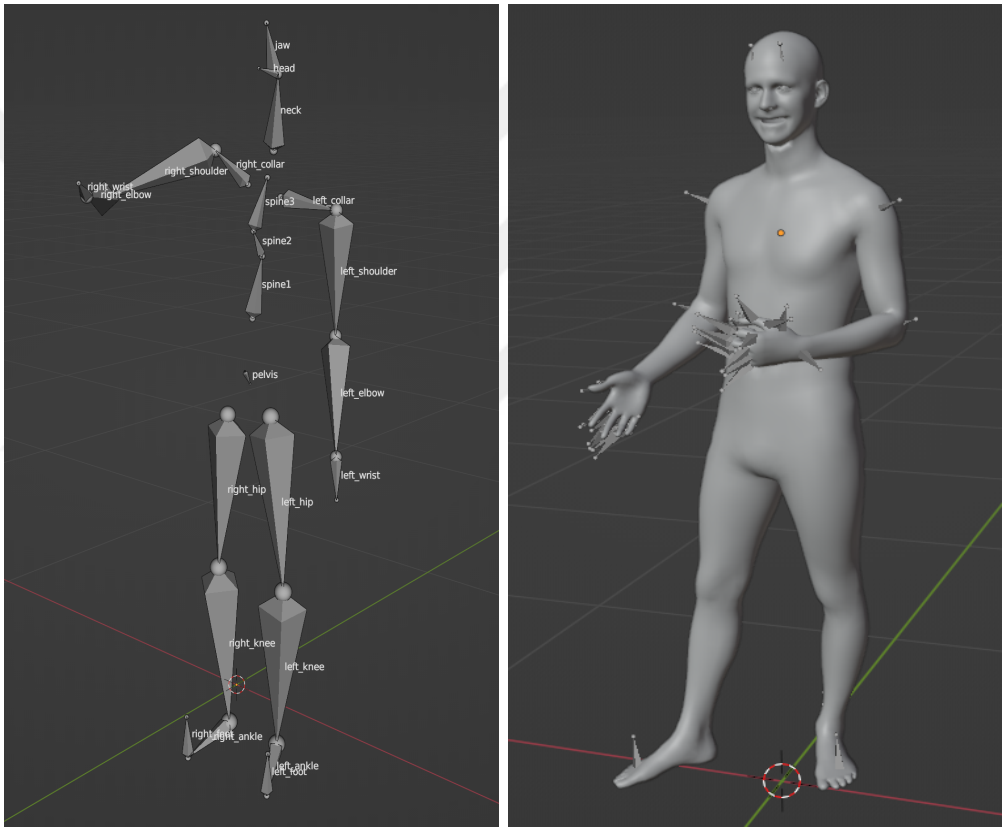

Figure 2.1: Examples of joint hierarchies. On the left, a screenshot from Blender [21] containing the skeleton joint of a human, performing a "pour" action. Each joint is visualized in terms of octahedra. Except *pelvis*, each joint points forward to its respective child in the hierarchy. Hand joints are omitted for visual simplicity. An SMPL-X [22] model for another sequence can be observed on the right.

2.1.2 Motion Capture and Object Reconstruction

Motion capture technology reconstructs the joint and facial movements of the actors to provide real-life references for realistic motions. In general, MoCap artists wear suits equipped with special devices, such as light balls and face capture cameras, and are tracked in studio environments with multiple cameras. Motion capture has been available for decades in the movie and gaming industry, corresponding to a tremendous amount of available data with diverse motions. A prominent example of these datasets is the CMU Graphics Lab Motion Capture Database [25].

Some sensor-based approaches are available for motion capture, such as XSENS [26] and Rokoko [27]. Such affordable motion capture devices allow creators with limited budgets to create motion assets for their projects. However, tools cannot be utilized for offline content, such as videos. Additionally, wearing sophisticated suits might interfere with the performance and lower the mobility of the actors.

Today's recent reconstruction systems, such as OpenPose [28], can capture the whole body motions of humans provided with a single RGB image; however, the reconstruction results in a 3D skeleton without a skinned mesh like SMPL-X [22]. EasyMocap [29], [30], [31], [32], [33], [34] and FrankMocap [35], [36] can infer SMPL-based models from images and videos. ROMP [37], [38], and PARE [39] can reconstruct multiple people in RGB videos in SMPL-X form. However, their solution does not include hands or faces. PIXIE [40] infers an SMPL-X model given a single RGB image. It can infer body, face (FLAME [3]), and hands (MANO [24]) for a single person. Figure 2.2 presents example results on a preview of emotional behavior. Each part of the person, body, face, and hands, is initially reconstructed using separate networks. Each network estimates a camera parameter based on its own cropped image features. Afterward, using a weighted expert system, body features are combined with hands and face separately to form the final parameters for both shape and pose. In Figure 2.2, the blue parts have high confidence, while the purple parts have low confidence. During normal

postures, hands and faces are reconstructed accurately. When a self-contact occurs, which causes occlusion, confidence in the hand lowers and gives realistic yet inaccurate results.

Figure 2.2: Example reconstructions using PIXIE, without using the FLAME [3] model. On the right, the elbow rests on the actor's hand and causes occlusions. Such occlusions lower the confidence of reconstruction.

Object models for our 3D object interaction sequences are in the form of 3D meshes. In earlier work, such objects could be represented in low-poly form compared to fully reconstructed models, which resulted in grasps with no articulations involved [41]. The objects in such scenes were modeled after their real-life counterparts, which requires manual labor that can be intensive and inaccurate depending on the complexity of the model. Datasets such as GRAB [7] use 3D printing to accurately know the model dimensions and level of detail, making it easier to track and represent in virtual scenes.

In case the 3D model of the object is not available via manual design or printing, reconstruction techniques can be utilized. When the object model is available for capture, approaches like NeRF [42] and Gaussian Splatting [43] can be used.

By capturing the object from multiple angles in RGB images, the methods mentioned earlier can convert its structure into a 3D volume or set of ellipsoids, which can then be converted into a mesh. It is also possible to capture large scenes with more images and angles with a detailed authoring process [44], [45].

To capture object mesh information in the wild with a single view, recent deep learning based methods can be separated into two categories: template-based and template-free. Template-based methods rely on providing a template model for the object mesh based on its category. In PHOSA [46], authors fit multiple people and objects in a given RGB image and optimize for contacts for plausible reconstruction. Given an interaction sequence, the objects are first detected and localized using off-the-shelf detectors such as Detectron2 [47]. Since objects can vary significantly in shape and orientation, they apply instance segmentation and fit a matching 3D template using differentiable rendering. The silhouette of the forward rendering result is expected to match the corresponding segmentation. This result is achieved by calculating one-way chamfer distance [48], from the render to the silhouette. Object size is determined as the average dimensions from the internet. Agents fit into their positions and orientations as SMPL [23] models using an occlusion and contact-aware pose-fitting approach. A more recent approach by Xie et al. [49] improves the contact estimation using a better depth-scale estimation to solve depth ambiguity. Hasson et al.[50], tries to solve the joint optimization problem with hands and objects by articulating the fingers to grasp properly. Collisions between hand and object are calculated using sign distance functions (SDF) rather than vertices in PHOSA [46]. It also handles tracking the template-based object to some degree across frames using Kalman filtering. Their method is limited to hands without whole body estimation. Another template-based method, Intercap [51], captures the interaction between an object and the subject using a fully calibrated multi-RGB-D camera setting with object templates. Even though it is accurate, the capture setting is tedious and limited.

In more recent approaches, the focus has shifted to template-free solutions, enabling the representation of diverse objects without a proper template. Ye et al. [52] propose to generate the object on the fly: no templates are used. However,

the object mesh varies greatly between different frames and is not expected to be stable. A more stable, diffusion prior-based approach was later introduced by Ye et al. [53]. Fan et al. [54] improves the quality of the reconstructed mesh by sampling rays in the 3D space projected to the image. InterTrack [55] captures both the subject and object during an interaction sequence by initializing with temporarily inconsistent point clouds from per-frame reconstruction. The noisy data is converted into a smooth reconstruction with autoencoders.

2.2 Learning-based Motion Generation and Authoring

In classical methods, each motion in the motion database is either captured using a MoCap system or manually authored. Depending on the scene context, character state, and interaction prompt, a proper animation can be transitioned smoothly via in-betweening, procedural blending, and blend trees based on the state of the character, including speed and direction. The state machine, or motion tree, regarding the character must be curated manually or procedurally to cover all plausible motions with accurate transition logic. This process becomes even more complicated when the characters are expected to be sensitive to their surroundings and interact with arbitrary objects. Such scenarios rely on predefined object types and contact regions, which are not scalable due to variations in object geometry, type, orientation, and relative body positioning. This approach leads to absurd transitions where the character suddenly "slides" to a predefined position and posture.

Learning-based motion generation approaches enable complex animations with high-level control, fluent composition, and adaptive scene interactions compared to manual and procedural motion generation. The adaptive nature of learning-based methods can provide an edge over procedural approaches with better scalability; however, their performance heavily relies on extensive and diverse datasets.

The learned latent representations for both animations and scene context additionally increase the style variation and combination of generated motions via novel motion synthesis.

Inspired by Neural Radiance Fields (NeRF) [42], He et al. [12] learn two separate embeddings to represent a manifold of motions across temporal sequences, represented by $f : (t, z) \mapsto (x_t^r, r_t^o)$ where t represents time domain, z represents latent space. Resulting motion is defined by joint rotations x_t^r and global orientation r_t^o. Their method learns two embeddings where global orientation and local joint movements are decoupled. The learned latent spaces can be interpolated, enabling motion in-betweening and generating diverse, plausible motions. The Kullback–Leibler divergence keeps the learned latent spaces close to a unit normal distribution. The reconstruction losses ensure the resulting motion is faithful to the original. This work, called NeMF, forms the basis of our approach and will be further discussed in the following chapters.

The basis of object interaction regarding humans generally involves the hands and feet of the subject, whereas sitting and sleeping positions involve other areas of the body as well. Starke et al. [56] propose an environment-aware neural model for object interaction and trajectory control. Their method encodes scene information regarding object proximity and shape using coarse primitives. Past scene context, user control, and phase information are encoded via a "gating network" which outputs blending coefficients for available experts. Each expert maintains a weight configuration for the future-predicting multi-layer perception (MLP). The future frame is generated in real time based on the "weighted" weights of the MLP and encoded scene information. FORCE [57] extends over this by encoding the weight resistance of objects and outputting the human force necessary for performing the desired action. There are other approaches involving full-body scene interactions in literature [58], [59], [60], [61], [62], [63], [64], [65], [66]; however, for the scope of our thesis, we focus on hand-only interactions with the objects.

Object interaction introduced additional challenges to motion synthesis due to the articulations of fingers. Improper articulations can cause penetrations with the object, resulting in unrealistic behavior. SAGA [67] and GOAL [68] solve

the problem of full-body motion generation towards the initial grasp pose, yet their method does not handle motions after grasping. ManipNet by Zhang et al. [69] addresses the challenge of generating dexterous hand-object manipulation across diverse, arbitrary object geometries by leveraging spatial sensors, including ambient sensors for object size, proximity sensors for distance, and signed distance sensors for fine surface details. The model can use these spatial features to predict future frame full finger poses for novel objects. GRIP [70] synthesizes accurate hand grasps for given body motion, object trajectory, and geometry. Their process consists of multiple autoencoders for arm stabilization, rough hand inference based on sensors similar to [69], and a refinement network.

IMoS [8] employs an attention-based, autoregressive VAE approach to synthesize motions involving objects. The scene constants, such as object and action labels, and subject shape β, are transformed into latent space using a separate condition encoder. Textual features are preprocessed using CLIP [71]. The temporal data, including body pose θ_b, arms θ_a, and object transformations $[T, R]$, are fed through two separate conditional autoencoders, which output body poses. The object position is updated using a separate object optimization module, which optimizes the 6-DoF object pose to maintain plausible hand-object contact. The body pose is processed by adding positional encoding and self-attention. Positional encoding adds temporal information to joint motion, where the attention module provides useful factors of each pose in high-dimensional space.

Braun et al. [72] generate physically plausible object interaction sequences using reinforcement learning to overcome the challenges arising from scale and diversity of object interaction data. Their approach uses separate hierarchical priors and policies for coarse body motion and hands with fine finger articulation. Priors are trained adversarially using MoCap data, while policies operate within Isaac Gym as a physics environment. Guided by adaptive policies, they can handle novel object geometries and localizations. More recent approaches like DiffH20 [73] and Text2HIO [74] use diffusion techniques to generate grasp and articulation motions based on provided textual descriptions. However, both methods are hand-only; they do not contain body motions.

Besides motion generation, controlling the "style" of an existing motion is essential for capturing the intended interaction behavior. Aberman et al. [75] apply style transfer to a content motion from a video recording as a style motion, where no content pairing is required. The style motion is encoded as both two-dimensional and three-dimensional features. The content motion is distilled from its existing style using Instance Normalization during the encoding process. Instance normalization normalizes features per channel and sample, thus removing style-related statistics from the motion, while maintaining its content. The style motion's embedding is learned via an MLP and inserted into the content motion's decoding process via Adaptive Instance Normalization (AdaIN) from work by Karras et al. [76] and applied during the decoding process of the generator network. AdaIN controls generated output via learning mapping parameters γ and β.

Unlike the previous method, Li et al. [77] uses a GAN network in multiple motion resolutions. They progressively upsample the original motion via linear interpolation, and the generator fills in the high-frequency details sampled from random noise. This results in coarse-to-fine learning, which enables maintaining the content of the original motion while enriching it with high-frequency movements. Their proposed method also supports providing a style motion and simple constraints, such as a trajectory. The constraints are concatenated to the input of the generator and discriminator and have the same temporal length as the upsampled motion. Style encoding is achieved by replacing the coarsest level of the output with downsampled content motion. However, style encoding produces limited style variation even for simple walking animation, due to the single-sequence training constraint. The authors also indicate that the semantics of both the content and the style motion must match, pairing walking style with walking content. Jang et al. [78] introduce a finer level of control by dividing the body into multiple parts where different motions can influence each.

Despite their performance, style transfer methods require a source motion for style. This requirement adds an extra challenge to object interaction scenarios, as they are spatially constrained, unlike motions like walking, waving, and greeting. Transferring between unrelated object interactions would cause erratic motions

due to different contact patterns, which can contradict the purpose of the object. In semantically identical motions, even though overall style can be successfully influenced, control over the object's trajectory and subject's motion stays coarse compared to quantitative controls.

More recently, due to the emergence of Transformers [79] and Diffusion processes [80], [81], text-based motion control became accessible to the research community. MotionFix [82], for instance, uses a Denoiser Transformer [83] approach to iteratively condition the denoising process on both source motion and edit text to generate the target motion. The target motion is noised to timestep t, then denoised conditioned on source motion and edit text, which is random noise $N(0, 1)$ during inference. Upon training, the source motion can be edited solely using an edit text such as "walk faster". Even though text-based motion authoring provides much finer control than style transfer approaches, textual descriptions are not discretely measured. The overall scale of texts is undefined and non-standardized compared to models like the OCEAN personality model [84], [85]. A natural language prompt can be perceived as a "vague" description for the intended behavior.

2.3 Personality Representation

2.3.1 OCEAN Personality Model

A character's personality is related to every feature they represent, from static posture to dynamic movements, and is an essential communication model for social contexts. In object interaction, movement style is expected to be the dominant feature [86] as the object of interest becomes the focus point for both the subject and the observer. Representing personality via proper quantization is crucial for determining the scale of perceived attributes of the motion. OCEAN model by McCrae et al. [84], [85] presents five orthogonal factors collectively representing the character's personality. Via these factors, it becomes possible to

control the overall motion style in a fine-grained manner in a vast space of potential motions. The five factors are: *Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness* and *Neuroticism*. Each factor, in the context of object interaction, can be summarized as follows:

- *Openness* is associated with the tendency for new experiences and creativity. Openness can be associated with curiosity for the object itself and the action performed during object interaction. High openness, for instance, can cause the agent to try novel trajectory patterns for experimentation. They can also perform their actions with more intent and visible gestures. Low openness, however, is expected to display an enclosed pose with less interest in the interaction. They can be hesitant in acting or might complete them in a rigid, unenthusiastic form.

- *Conscientiousness* is one of the most related factors for object interaction as it represents organization and caution of the character. A highly conscientious character would take the necessary care when acting, following expected and appropriate transformations for the object. They would also abide in affordance of the object, using it as intended. A low level of conscientiousness is expected to cause the character to handle the interaction carelessly and is even likely to fail at performing.

- *Extraversion* increases with a character's desire to engage in social activities and outgoing personality. Not directly associated with object interaction, this factor might influence the perceived enthusiasm for interaction with the extent of their body pose.

- *Agreeableness* is also associated with social dynamics of the character with others. A character with high agreeableness is expected to get along and cooperate, while a character with low agreeableness results in confrontational behavior. For object interaction, this can be associated with passing an object. A high agreeableness character would handle the object with care and offer it gracefully. At the same time, an inverse case might display hostility via fast and harsh movements while posing a danger to the object itself.

- *Neuroticism* negatively affects the character's stress. A character with high neuroticism experiences anxiety much more frequently than a character with low neuroticism. This factor can be associated with sudden and frightful movement during interaction with exaggerated expectations from the object. For instance, a highly neurotic character can worry that hot tea from the teapot might spill and unintentionally act clumsily.

2.3.2 Laban Movement Analysis

Laban Movement Analysis (LMA) [87], [88] is a method that emerged from dance to study human movement. LMA was developed with four categories as components: *Body*, *Effort*, *Shape*, and *Space*. For control over the motion of the character, Laban *Effort* values can be used to adjust its dynamics and intention. Laban Movement Analysis is also used to estimate OCEAN personality factors for an input motion by Erkoç [89]. Based on studies by Durupınar et al. [86], we define mappings between OCEAN and LMA effort parameters and utilize them for augmenting our data. More detailed information can be found in Chapter 4.

2.3.3 Motion Authoring with Personality

In PERFORM, Durupınar et al. [86] investigated the relationship between OCEAN and LMA Effort parameters within two action types: pointing a finger and picking up a pillow from the ground. Following the guidance of laban movement experts, they modeled an empirical mapping function for conversion between OCEAN and LMA, which is validated via user studies. This mapping enables manipulating a given motion towards a particular OCEAN factor combination by adjusting LMA effort parameters. Sonlu et al. [90] extends over this approach by incorporating voice and facial animations in scenarios with complex dialogue. Yurtoğlu et al. [14] transfers motions between different videos using deep learning to compare the contribution of motion and appearance to personality expression.

Chapter 3

Human-Object Interaction Datasets

To train our motion-generation framework, we investigated a diverse set of human-object interaction datasets in the literature. The main distinction for datasets is the representation format: two-dimensional or three-dimensional. 2D datasets merely provide bounding boxes and labels for apparent interaction in visual media, such as still images or video sequences. 3D datasets capture human motion, including hands, object mesh, and orientation information that encapsulate richer information regarding the context.

3.1 2D Datasets

2D datasets contain videos with one or more people, where their physical bodies and interactions are annotated as captions and bounding boxes. Due to their simple capture process and wide availability in visual media, 2D datasets include diverse categories of interactions involving multiple and articulated objects. Such scenes can involve reading a book, playing with a ball, or peeling an apple. However, due to missing 3D information and a challenging reconstruction process,

they cannot be directly utilized for 3D motion generation purposes. It is possible to extract 3D geometric information from existing 2D datasets using the aforementioned methods in Chapter 2. However, due to such interactions' noisy and occluded nature, the post-processing process becomes infeasible. Even then, they can be subject to determining the categories of interaction, which may represent, for example, an emotional interaction. This section covers various object interaction-focused 2D datasets to provide possible heuristics for selecting interaction types with potential emotional representation. The summary of contents for covered datasets is included in Table 3.1.

Dataset	Context	Actions	Sequences	Labels
AVA Actions [4]	YouTube Videos	80	230K	BB, labels and caption
EPIC Kitchen [91], [92]	Kitchen Activities	97	90K	BB, labels and caption
20BN-Someting-Something [93], [94]	Daily objects	174	221K	Something action something
HICO-DET [5]	Arbitrary Objects	117	600	Verb-object, BB and associations

Table 3.1: Summary of 2D Datasets, BB: Bounding Boxes

AVA actions dataset [4] includes YouTube videos where each frame is annotated with bounding boxes, their labels, and interaction captions. The captions for the frames only include the action itself, where the involved object is denoted as "an object" and not localized. Even though interaction types are highly diverse, the lack of object annotations and the unconstrained nature of the videos make this dataset a challenge to traverse.

Similarly, EPIC Kitchen Dataset [91], [92] contains unscripted, egocentric kitchen activities. It only contains bounding boxes and scene captions. 20BN-Something-Something Dataset [93], [94] contains humans performing predefined, basic actions with everyday objects. The labels are less descriptive than the others: "something" verb "something." HICO-DET [5], on the other hand, provides descriptive annotations for both human and interacted objects in terms of bounding boxes. Example annotations can be found in Figure 3.1.

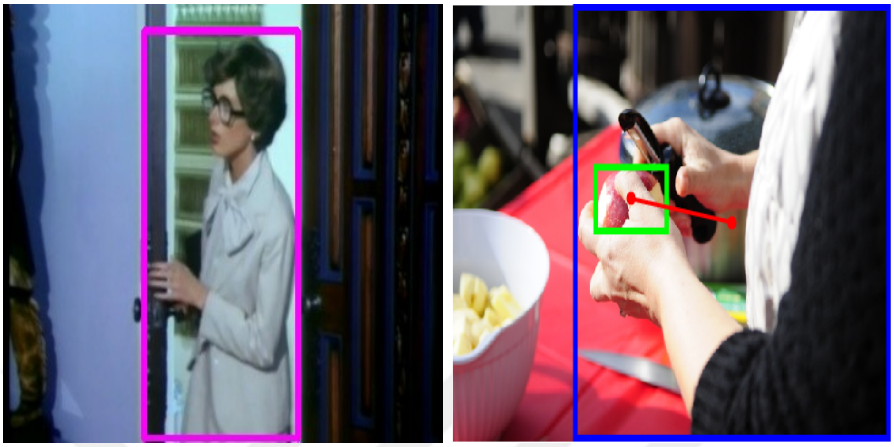

Figure 3.1: Examples from 2D datasets, AVA [4] (left) and HICO-DET [5] (right). In HICO-DET, the object is also annotated and associated with the person interacting with it.

3.2 3D Datasets

Thanks to their frequent usage in computer animation research, there are vast 3D animation datasets regarding human motion. However, only a subset of such motions includes interactions with the surrounding environment, whereas even a smaller subset includes the mesh model. We summarize *whole body* 3D datasets in Table 3.2.

Dataset	Subjects	Actions	Objects	Sequences	Size	SMPL	OM & OT	Segments	Contacts	Text
KIT Bimanual [6]	2	12	21	588	588K	✗	✓	✓	✗	✗
D3D-HOI [95]	5	8	24	256	6286	✓	✓	✗	✓	✗
GRAB [7]	10	4	51	1,334	1.62M	✓	✓	✓	✓	✗
ARCTIC [96]	10	2	11	339	2.1M	✓	✓	✓	✓	✗
BEHAVE [97]	8	5	20	321	15.2K	✓	✓	✓	✓	✗
Intercap [51]	10	1	10	223	67.4K	✓	✓	✓	✓	✗
HIMO [98]	34	3,376	53	3,376	4.08M	✓	✓	✓	✓	✓
Parahome [99]	38	22	22	208	61.2M	✓	✓	✓	✓	✓

Table 3.2: Summary of 3D Datasets. OM: Object Mesh, OT: Object Tracking, Text: Detailed Text Description. The dataset size is given in terms of the number of frames.

3.2.1 Hand Only Datasets

ContactDB [100] and ContactPose [101] datasets contain 3D objects that are grasped by one or both hands of the subject and scanned using RGBD thermal images. The heat on the object's surface induces contact areas while ignoring finger semantics during contact. The intents are limited to "use" and "hand-off." ContactPose [101] is an extended version of ContactDB. It contains finger semantics and fits the parametric MANO [24] model of SMPL-X [22]. Even though both datasets contain grasping data, they are limited to a single frame and lack body posture.

Part-Net [102] and its articulated version SAPIEN [103], [104] include various household objects and ways to interact with them. In addition to the information on the parts, the purpose of each part during various activities is crucial. A more densely annotated dataset, Oak [105], has collected 1800 everyday household objects and annotated affordance information using crowdsourcing. The 3D real-life objects are interacted with and tracked in a multi-camera setting. Objects are used, lifted, held, and handed over to another participant. Participants only used a single hand during the interaction, and their bodies were not captured. Then, to adapt the extracted grasping over similar objects, they learned a latent representation of the object with contacts that can be interpolated while preserving plausible contact annotations. The latent space representation can be helpful when transferring contact knowledge within the same class of objects.

3.2.2 Whole Body Datasets

KIT Bimanual Intersection Dataset [6] is an extensive dataset on the dense reconstruction of human-object interactions. The data is captured on RGB-D cameras, data gloves, a marker-based motion capture system, a head-mounted camera, and inertial measurement units (IMU). Data gloves allow the system to handle cases where hands are close to each other, which would cause problems with marker-only solutions. The subject and the object(s) are tracked throughout daily house

activities. The data is available for the public via sign-up.

The scene's actors interact with various kitchen-related objects, including cups, peelers, and spoons. Each interaction occurs on or behind a table with a predefined height. At the beginning and end of the interaction, the subject is in a hand-down position, where the hands touch the table where the object resides. Depending on the action type, one hand stabilizes the object while the other manipulates it. In some cases, both hands can be seen to manipulate the object. There are variances in the object starting location to cover arbitrary localization cases. The manually annotated interaction segments, including *approach, lift, hold, move, place*, and *retreat*, are also available for both hands. Figure 3.2 shows several object types and an example action sequence from the KIT dataset.

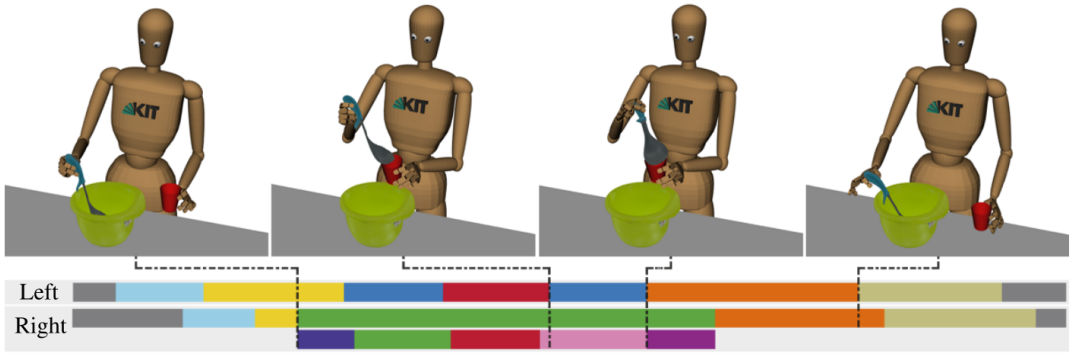

Figure 3.2: The object types and an example action sequence from KIT [6]. During the sequence, the actions of both hands are denoted in different colors. For example: yellow=lift, pink=pour, and orange=place. The bottom line indicates the stages of the action.

Each motion is captured in C3D format and standardized in a reference model Master Motor Map (MMM) [106]. This reference model, resembling a puppet, fits on the tracked landmarks on the real subject, including their hands. The provided objects are tracked using markers and scanned/printed in 3D. Objects are available in WRL format. No heatmaps are available for parts of the objects manipulated by the subjects.

GRAB [7] uses the SMPL-X model where the subject interacts with a single object using their hands. The dataset is captured using MoCap markers for both humans and objects. Objects are 3D printed; therefore, their dimensions

are known beforehand. The contacts between the hands and the object are not detected using sensors but are annotated by finding penetrations between the vertices. Such contacts are collected and represented as heat maps. See Figure 3.3 for an example render. Using real objects, ARCTIC [96] provides bimanual interaction sequences with detailed object articulations and contact. HIMO [98] further enhances the interaction scenarios by introducing a second object, text labels, and temporal segments.

As neural networks tend to take inputs in fixed dimensions, providing arbitrary meshes to model with varying levels of detail becomes a challenge. In GRAB [7], Taheri et al. represented meshes in the form of Basis Point Sets (BPS) [107]. BPS contains a fixed number of points in a spherical manner around the object of interest, where their locations and count are predefined. Each point contains the distance from the closest vertex on the mesh. This way, the entire mesh can be represented in a fixed-length vector based on distance information.

BEHAVE [97] dataset combines multi-camera motion tracking with contact estimation to determine the human and object 3D models. Their approach uses multiple Kinect RGB-D cameras. The unified point cloud of the human and the object is used to calculate a distance field, and 3D models are fit to minimize the surface error. They represent the human in SMPL form. The contacts are found based on the distance between the vertices of each model. Since the hands are modeled using SMPL without MANO, finger articulations are unavailable, and contacts are not detailed. However, specific actions involve body parts other than hands. The available action types are: *Lift, carry, sit, push*, and *pull*. Intercap [51] dataset enhances over BEHAVE [97] by adding finger articulations while using a similar setup containing multiple RGB-D cameras.

D3D-HOI [95] includes articulated kitchen appliances such as refrigerators, washing machines, and dishwashers. The actors in the scene are reconstructed as SMPL-X [22] models, and object joint articulations are annotated in real-life measures. Kim et al. [99] capture and validate complex object interactions in a home environment using an extensive motion capture system containing multiple markers, lights, and cameras.

As objects of interest in the recent studies are not uniform, each part of the mesh has a different purpose relative to the usage of the object. This association is referred to as the object's affordance, which guides the subjects towards proper interaction. Datasets such as PartNet [102], [104] investigate part information, while 3D AffordanceNet [108] tries to segment a given object geometry based on visual affordance reasoning. A more recent work, ComA [109], predicts contact regions for both object geometry and SMPL-X [22] human body models.

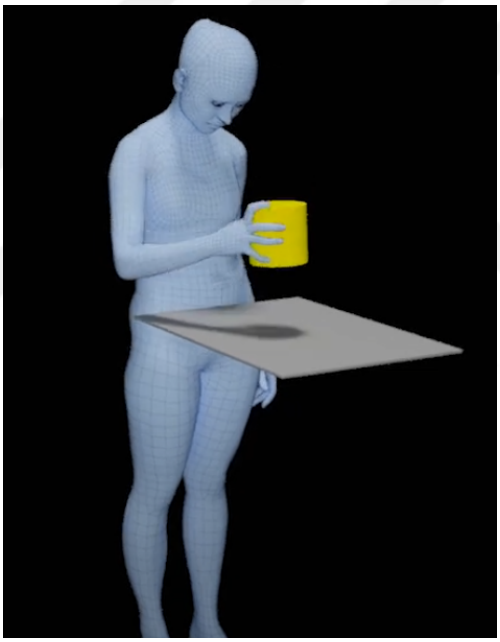

Figure 3.3: A woman grabbing a cup in GRAB [7] dataset.

3.2.2.1 Custom Dataset

We opted for the GRAB dataset, which includes non-articulated object interactions, to minimize reconstruction errors. Personality expression is not the primary goal of GRAB. The subjects do not aim to convey personality traits or emotions; such expressive features, if any, occur naturally. Since object interaction happens in a nonsocial context, detecting an expressive style in arbitrary samples is challenging. Consequently, we manually filtered the GRAB dataset to 10 basic household items with identifiable contexts. We also removed two hand interaction samples for simplicity, leaving us with 10 object categories and 102 sequences.

The statistics of the selected subset of the GRAB [7] dataset are shown in Table 3.3. Note that a set of subjects interacts with more than one object category, and some action definitions are shared across objects while causing distinct behaviors.

Object	**Action**	**Subject**	**Sequence**	**Frame**
Bowl	1	4	4	305
Cup	1	8	8	690
Flashlight	1	7	14	1,930
Frying pan	1	4	4	460
Hammer	1	10	22	2,510
Knife	2	4	6	805
Mug	1	10	20	2,030
Scissors	1	5	6	735
Teapot	1	6	10	1,115
Wineglass	1	7	8	635
TOTAL	8	10	102	11,215

Table 3.3: Statistics of the customized GRAB [7] dataset.

Chapter 4

Motion Augmentation Framework

In the previous chapter, we discussed various object interaction datasets involving diverse actions and object types. Due to the vast space of possible trajectories, hand articulations, and body poses, certain motions are repeated to demonstrate variations. Even though such "reshoots" increase the motion count, due to a lack of focus on personality, current datasets do not provide sufficient data regarding variance in terms of personality. To address the issue of low sample size and limited expressive variance, we developed a motion augmentation framework using Blender [21].

The framework alters input motion regarding various joint rotations and temporal resolution, following adjustments based on previous work by Ergüzen et al. [110]. Eight additional joints are utilized via Inverse Kinematics (IK) to control the hands, feet, knees, and elbows. The framework also updates the expressed personality of the edited motion, based on empirical studies by Durupınar et al [86]. During authoring, we used LMA Effort parameters: *Space*, *Time*, *Flow*, and *Weight*, with each parameter being in the normalized $[-1, 1]$ range.

For the rest of the chapter, each factor and its influence on the motion is described and demonstrated. Then we explain our mapping function to update the authored motion's personality, which we finalize by describing additional adjustments to the motion to ensure its integrity.

4.1 LMA Space

Space factor controls the distance between *hand* and *feet* joints and declared as factor f_s. Positive f_s values increase the distance between each hand and feet along the frontal axis, while negative f_s values bring the joints mentioned before closer. This behavior results in "indirect," thus, "spreading" motion for positive values and "direct," thus, "enclosing" for negative values. Each behavior can be associated with the object's weight or the subject's conscientiousness.

The target bones are paired with the second parent of their attached bone, namely the *"Limit"* bone. Limit bones are the upper arm and leg bones for hand and foot targets, respectively. The delta vector V_L between the limit and target bones is calculated for each frame and decomposed into the frontal (V_L^f) and longitudinal (V_L^l) components to be used for rotation. The rotation amount R is proportional to f_s and has a maximum of 15 degrees, determined empirically. When f_s is negative, and the distance between symmetric target joints d_{sym} is below the max limit distance, f_{sym} is applied as a correction term. f_{sym} is non-zero in case the target limbs are *not* crossing each other. f_{sym} is calculated using Equation 4.1:

$$f_{\text{sym}} = \begin{cases} \left|1 - \frac{|d_{\text{sym}} - \|\mathbf{V}_L\||}{\|\mathbf{V}_L\|}\right|, & \text{not crossing} \\ 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

The final R value is multiplied by the f_{sym} and is used to find the frontal ($shift_f$) and longitudinal ($shift_l$) shifts, as in Equation 4.2. Figure 4.1 provides an example visualization of a "cup" object in "pour" action, showing the effect

of the *space* parameter.

$$\begin{aligned} shift_f &= (V_L^f \cdot \cos(R) - V_L^l \cdot \sin(R)) - V_L^f \\ shift_l &= (V_L^f \cdot \sin(R) + V_L^l \cdot \cos(R)) - V_L^l \end{aligned} \tag{4.2}$$

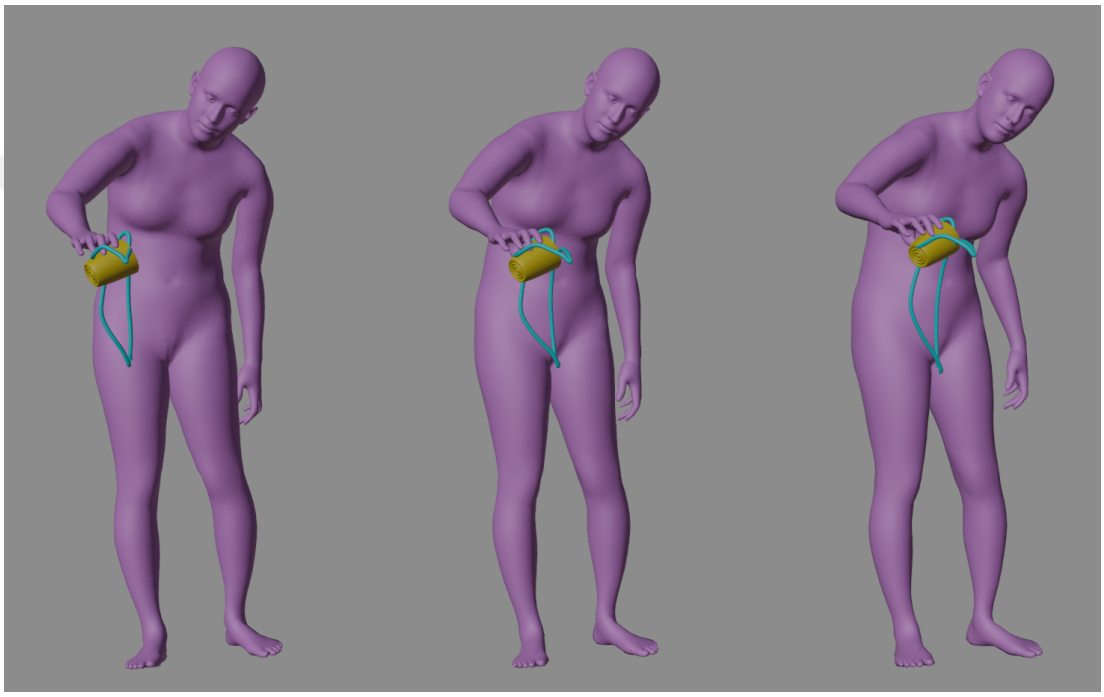

Figure 4.1: Example augmentations with *space* from −0.75 to 0.75; left to right. Notice the change is subtle, as seen in object height, due to the enclosed posture of the subject.

4.2 LMA Weight

Weight parameter modifies the vertical posture of the agent while keeping feet grounded. Increasing f_w imposes "heavy" weight on the subject, causing descent, while decreasing it causes the subject to stand taller with a "lighter" weight. This parameter is essential for conveying the weight of the interacted object, for instance, carrying a heavy box.

The weight factor primarily affects the "pelvis" bone of the subject; the effect is scaled according to the midpoint M between the two foot targets. This midpoint is used to calculate shifts on sagittal and frontal axes $shift_{sf}$ (Equation 4.3).

In contrast, the distance between the pelvis and the floor d_{lim} determines the longitudinal shift $shift_{ln}$ (Equation 4.4). The distance between M and the pelvis is indicated as d_p. When the shift factor f_w is negative, the distance between the pelvis and either of the upper leg bones is calculated as d_n, and the distance between the lower foot bone and its upper leg bone is d_l. The shifts are combined into a 3D vector and applied to the pelvis and the hand joints. For neck and spine bones, positive and negative f_w rotate around the frontal axis in 15 and 5 degrees, respectively. Example visualization in Figure 4.2 indicates that increased *weight* influences the subject as if the "cup" is heavy.

$$shift_{sf} = \begin{cases} d_p \cdot f_w, & \text{if } f_w > 0, \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

$$shift_{ln} = \begin{cases} \frac{d_{lim}}{20} \cdot f_w, & \text{if } f_w \geq 0, \\ \max(\frac{d_n - d_l}{10}, 0) \cdot f_w, & \text{otherwise} \end{cases} \tag{4.4}$$

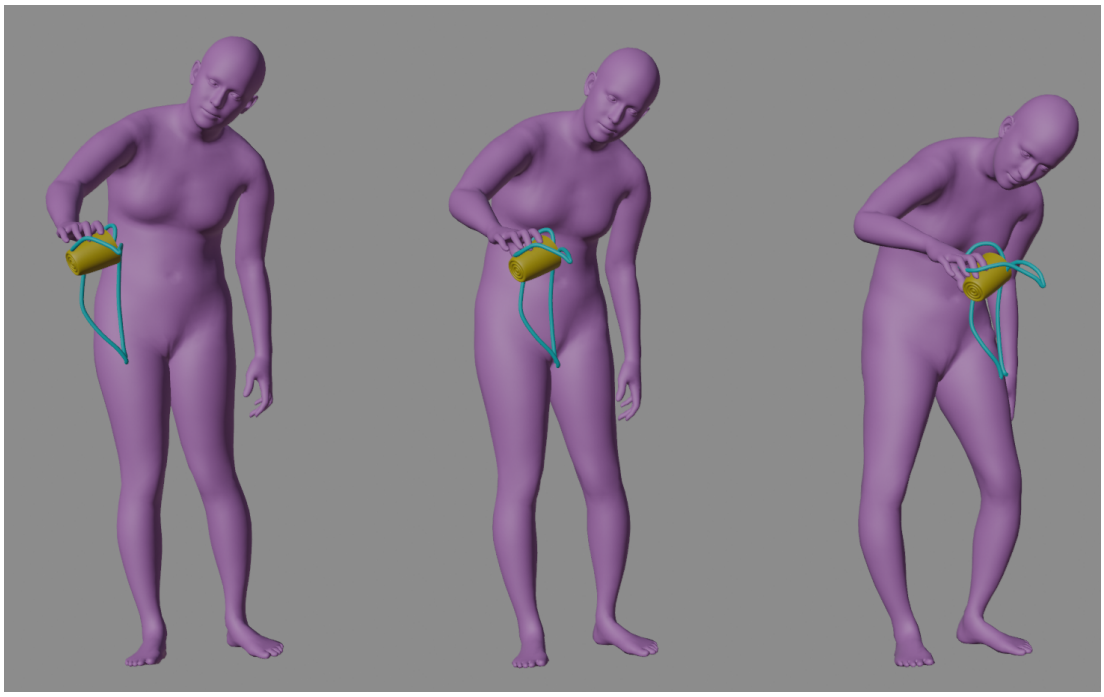

Figure 4.2: Example augmentations for *weight* between −0.75 and 0.75; left to right. The subject bends forward like the "cup" is a heavy object.

4.3 LMA Time

Time changes the playback speed of the animation, depending on the delta changes between consecutive frames of hands and feet. The lower the delta changes, the slower the movement, and the more aggressive the speed adjustment.

The shift amount s_k is calculated according to the time factor f_t (Equation 4.5). The new shift $S(i)$ for keyframe i is calculated in the incremental fashion where N is the total number of keyframes, except the first frame (Equation 4.6). The shift can be noticed in different subject postures in Figure 4.3.

$$s_k = \begin{cases} f_t + 1, & \text{if } f_t \geq 0, \\ \frac{1}{|f_t|+1}, & \text{otherwise} \end{cases} \tag{4.5}$$

$$S(i) = s_k + i \cdot |1 - s_k|/N \tag{4.6}$$

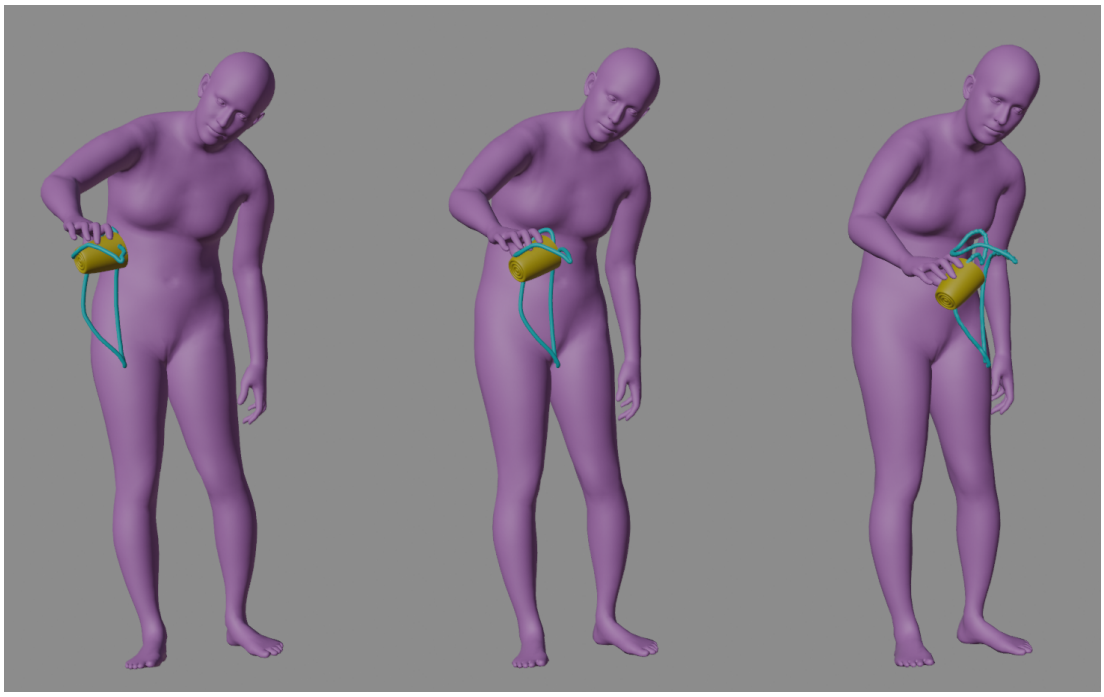

Figure 4.3: Example augmentations for *time* between -0.75 and 0.75; left to right. Due to the resolution change, the subject can be seen in a different pose at the same timestep as others.

4.4 LMA Flow

Flow reduces the number of keyframes without changing the animation duration if the provided factor value f_f is negative. We achieve this using Blender's internal *Decimate Keyframes* operation. Further reducing f_f causes the animation to become more "bounded" as more keyframes are reduced to interpolated animation curves.

The decimation amount is determined as $f_{dec} = 0.9 \cdot |f_f|$. For positive values of f_f, random keyframes are selected to add noise to their rotations. The noise is sampled from two normal distributions with means $\mu_1 = -2f_f$ and $\mu_2 = 2f_f$ and common standard derivation of 0.75. If the sampled noise is above 2 in magnitude, it is resampled, as higher f_f values result in less variation. The range of frames is also calculated using random distribution, varying from 5 to 30, which is determined empirically (Equation 4.7). Altering the flow changes the trajectory of the object due to changes in the subject's pose, as seen in Figure 4.4.

$$\begin{aligned} range_{min} &= 5 + (1 - f_f) \cdot 15 \\ range_{max} &= 5 + (1 - f_f) \cdot 35 \end{aligned} \tag{4.7}$$

4.5 Personality Augmentation

Augmenting the original motion is expected to alter its apparent personality regarding OCEAN factors. For this purpose, we utilized the mapping in [86]. The mapping originally applies reversely, from OCEAN personality factors to Laban Effort. To obtain the LE2OCEAN mapping, we transposed the normalization axis with the significant portions kept intact. In [86], a value is considered "significant" if they are statistically significant ($\rho < 0.05$). We multiplied the resulting matrix in Table 4.1 with the input Laban Effort parameters to obtain the OCEAN trait adjustments. A factor of 1 for *Space* would cause a 0.86 increase on *Openness*, -0.86 in Conscientiousness, and 0.896 in Neuroticism. The resulting OCEAN values are directly added to the original OCEAN values and clamped if necessary.

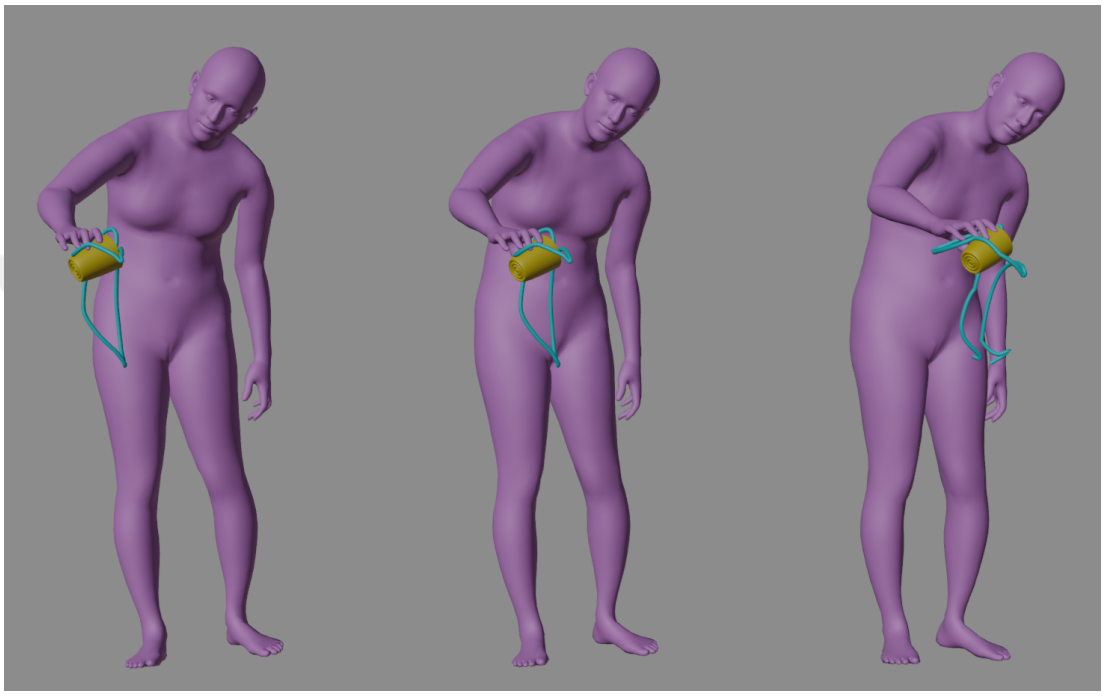

Figure 4.4: Example augmentations for *flow* between −0.75 and 0.75; left to right. The increased flow alters the object's trajectory, as seen in differences in blue lines.

Effort	Space	Weight	Time	Flow
O	0.86	0.0	0.0	1.0
C	-0.86	0.0	0.91	-1.0
E	0.784	0.0	-0.997	1.0
A	0.0	0.685	1.0	0.0
N	0.89	0.0	0.0	0.776

Table 4.1: Matrix to calculate OCEAN delta trait values from LMA effort parameters.

4.6 Motion Integrity

To properly augment the provided animation, obeying the integrity of the motion, we follow the steps below:

- Our Blender-based augmentation works on BVH (BioVision Hierarchical Data) format [111]. To use AMASS [112] motions, we import our SMPL-X motion into the Blender [21] scene and export it in BVH format.
- Using the aforementioned LMA parameters, we augment the motion present in the scene, without considering the interacted object.
- To reduce gender bias in our user studies, we used a gender-neutral SMPL-X model. We also removed facial expressions and set the beta parameters associated with the body shape to zero to remove any potential biases. However, this alters body proportions, such as height and arm length. This change does not cause problems in isolated motions, but extra steps are required to preserve the object's motion.
- We first fix the subject's fingers to their initial articulations and stick the object to the corresponding hand based on its original orientation relative to the wrist. Sticking is achieved via utilizing Blender [21] internal constraint logic. As the initial hand location would be altered due to altered body proportions, the distance between "neutralized" and the original hand is used as an offset for the object, as seen in Figure 4.5.
- The object follows the hand motion, which overwrites the original trajectory. However, we assume the finger's articulation is maintained across the motion, which can cause unrealistic grasps during the interaction. This approach also requires the user to provide motions where the subject maintains contact with the object.
- To convert from BVH to AMASS [112] format, we apply retargeting. The retargeting process maps a provided motion in one skeletal animation to a different skeletal hierarchy, while maintaining proper rotations. We used the add-on from [113] to apply retargeting from BVH to SMPL-X format.

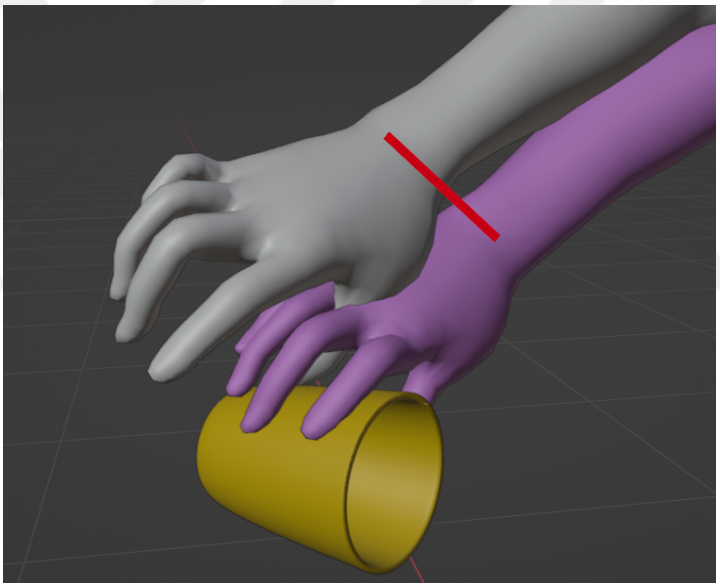

Figure 4.5: Updating the hand location due to altered body proportions. The gray agent is the original movement, where the purple agent is the neutralized one. The drawn offset is considered for the object's initial movement and thus sticking.

Chapter 5

Human-Object Interaction Motion Control

5.1 Generator Model Structure

We base our neural motion control on a Variational Auto Encoder (VAE) based approach, NeMF by He et al. [12]. The summary of our model can be found in Figure 5.1. Our network controls the motion's personality via high-level OCEAN factors while maintaining semantic consistency. The original NeMF architecture considers local and global motion components for separate latent encodings. The local component is responsible for representing the joint rotations, positions, and velocities, while the global component contains the orientation of the body and its translation. As motions in our dataset lack any noticeable translations, we discard the global component of the network and copy such features directly to the output motion. Our approach accepts motions in 128 frames, similar to the original architecture.

We split our NEMF-based generator into two parts: encoders for each aspect of the motion and a decoder module to emit different features regarding the synthesized motions. For encoders, we split features of the motion as: *Local*

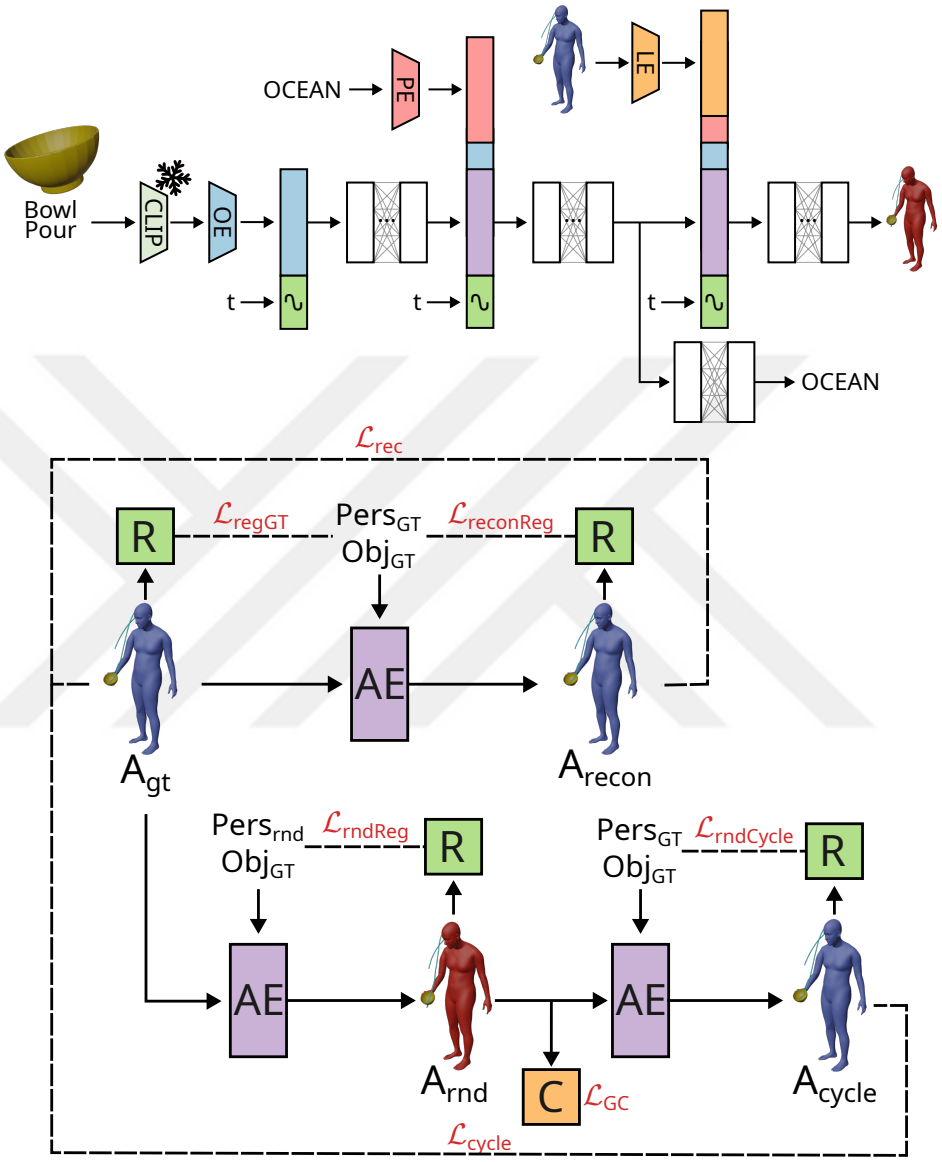

Figure 5.1: The proposed neural model for personality manipulation of human-object interaction motions. Top: The NeMF-based, neural personality manipulation model. The model takes scene features in the following order (determined empirically): object name and purpose, target OCEAN factors, and motion parameters of the source motion. For temporal resolution, the latent vectors are accompanied by a positional encoded time t. The output motion's (red) personality matches the input personality and exported OCEAN factors. Bottom: During training, *critic* (C) model assesses the overall motion quality while *regressor* (R) model assesses motion's semantic accuracy, based on provided object features. We also apply cycle loss to ensure the neural model can transition between original and edited personalities. Both *critic* and *regressor* are trained together with the generator (AE) in an end-to-end fashion.

Encoder for local features, *Object Encoder* for object information, and *Personality Encoder* for target personality. Each encoded feature is independently forwarded into its separate linear layer to calculate distribution parameters μ and σ to fulfill the variational properties of the network.

Local encoder takes features for each joint, namely position and velocity, angular velocity, and global rotations, according to the subject's pelvis. Each position and velocity vector is represented in 3D vectors, while global rotations are represented in 6D rotations [114]. 6D rotation representation is differentiable, making it suitable for learning. It is based on representing rotation as two vectors that can be converted to rotation matrices for forward kinematics using a Gram-Schmidt-like process. NeMF's local encoder was based on the SMPL model [23], where finger joints are ignored. Similarly, we ignore the fingers in our network model, while using SMPL-X [22], as the subject's fingers are locked. To obtain latent representation from the graph of joints, we concatenate them in the time dimension t and feed them to a predefined graph-based skeleton convolution by Fussell et al. [115].

The *object encoder* takes the object information in the form of object intent and name. Object intent and name are initially in text format; therefore, we preprocessed them using frozen CLIP [71] into fixed-size (512) latent vectors, as in [8]. As fingers are locked and the object follows the hand, we ignore the object mesh and its original trajectory information. Similar to local motion, all features are concatenated into a single vector in the time domain t and fed to layers of residual blocks. Each residual block contains *Conv*,*BatchNorm*,*Activation* layers, accompanied by residual connections. As NeMF expects features to be provided across all motion frames, we repeat the obtained latent representations for intent and object name for the expected number of frames, which is 128 in our work.

The *personality encoder* takes OCEAN personality factors to control the personality of the motion. Each OCEAN factor with the original scale of $[-3,\ 3]$ is normalized to $[-1,\ 1]$ before proceeding. The factors are constant across frames, like object intent; it is repeated across frames. Similar to *object encoder*, *personality encoder* has multiple residual blocks with different input dimensionality.

The base NeMF decoder has multiple residual blocks, where each encoder is introduced to the network as input at different steps. The latent vector fed to each layer is initialized with positional encoding of time t to provide a temporal relation between frames. The first encoder's latent is concatenated to the current latent vector. For a predetermined number of upcoming layers, this latent vector is concatenated with the immediate output of each layer using skip connections. After processing a particular encoder's layers, the latest feature vector is fed to a fully connected network for the related output. The next encoder's latent is concatenated to the latest latent vector, and the process is repeated for the next set of layers. In the original work, the order of latents was *Local* ($\boldsymbol{z}_l$) and *Global* ($\boldsymbol{z}_g$), which implies a dependency. In our work, we determined order as *Object* ($\boldsymbol{z}_o$), *Personality* ($\boldsymbol{z}_p$), and lastly *Local* ($\boldsymbol{z}_l$). This way, object-related context and target personality affect the subject's motion, not vice versa.

5.2 Training Details

Our training process involves applying multiple losses to all our core components, which are trained end-to-end: *generator*, *critic*, and *regressor*.

To ensure the *generator* can reconstruct a provided motion, it produces the original motion according to its original personality. The resulting motion is subject to reconstruction loss, $\mathcal{L}_{rec}$, relative to the original motion. Complementarily, the generator is then fed with a random personality to produce a random motion, which is reverted to the original motion using the original personality through the generator. The "cycled back" motion contributes to the generator's overall loss via cycle loss $\mathcal{L}_{cycle}$.

The original encoders of NeMF outputs Gaussian distribution parameters in terms of mean μ and standard deviation σ, and samples using the reparametrization trick, to represent a motion prior from input. Updated with our additional inputs, the variational lower bounds are given in Equation 5.1:

$$\begin{aligned}\log p_\theta(\mathbf{x}^r, \mathbf{r}^o, \mathbf{o}^p, \mathbf{x}^p) \geq \mathbb{E}_{q_{\phi_1}, q_{\phi_2}, q_{\phi_3}} \left[\log p_\theta(\mathbf{x}^r, \mathbf{r}^o, \mathbf{o}^p, \mathbf{x}^p | \boldsymbol{z}_o, \boldsymbol{z}_p, \boldsymbol{z}_l)\right] \\ -\mathcal{D}_{KL}(q_{\phi_1}(\boldsymbol{z}_o|\mathbf{o}^p) \parallel p(\boldsymbol{z}_o)) \\ -\mathcal{D}_{KL}(q_{\phi_2}(\boldsymbol{z}_p|\mathbf{x}^p) \parallel p(\boldsymbol{z}_p)) \\ -\mathcal{D}_{KL}(q_{\phi_3}(\boldsymbol{z}_l|\mathbf{x}^r, \mathbf{r}^o) \parallel p(\boldsymbol{z}_l))\end{aligned} \tag{5.1}$$

where $\mathbf{x}^r, \mathbf{r}^o, \mathbf{o}^p, \mathbf{x}^p$ represent joint rotations, root orientation, object properties, and personality, respectively. θ stands for decoder weights and each ϕ stands for encoder weights. ϕ_1 stands for object, ϕ_2 for personality and ϕ_3 for local encoder. Each Kullback-Leibler Divergence term $\mathcal{D}_{KL}$ acts as a regularization to ensure the resulting Gaussian distributions are similar to the standard normal distribution Z, thus providing a continuous motion space with movements, objects, and personalities. By giving time t and latent components, the variational property of the network enables smooth traversal of the motion space.

We use Equation 5.2 for the rotation and orientation losses.

$$\mathcal{L}_{rot/ori} = \sum_{t=1}^{T} \arccos \frac{\mathrm{Tr}(R_t^{GT}(R_t^{Synt})^{-1}) - 1}{2} \tag{5.2}$$

where R^{GT} represents ground truth rotation features, while R^{Synt} represents synthetic rotation features. We calculate the reconstruction ($\mathcal{L}_{rec}$) and cycle ($\mathcal{L}_{cycle}$) losses using Equation 5.3. We use the following terms for these losses:

Position loss ($\mathcal{L}_{pos}$): The $L2$ norm of the difference between predicted and ground-truth joint positions. λ_{pos} is set as 20 during training.

Rotation loss ($\mathcal{L}_{rot}$): The geodesic distance between predicted and ground-truth joint rotation matrices. λ_{rot} is set as 7 during training.

Orientation loss ($\mathcal{L}_{ori}$): The geodesic distance between predicted and ground-truth root orientation. λ_{ori} is set as 2 during training.

Joint velocity loss ($\mathcal{L}_{vel}$): The $L2$ norm of the difference between predicted and ground-truth joint velocities. λ_{vel} is set as 1 during training.

Angular velocity loss ($\mathcal{L}_{avel}$): The $L2$ norm of the difference between predicted and ground-truth angular velocities. λ_{avel} is set as 1 during training.

KL divergence loss ($\mathcal{L}_{KL}$): This loss measures the difference between the learned latent distribution and the prior distribution. All latent embeddings' losses are summed. λ_{KL} is set as 10^{-4} during training.

OCEAN loss ($\mathcal{L}_{OCEAN}$): The $L2$ norm of the difference between the predicted and ground truth OCEAN personality factors. λ_{OCEAN} is set as 2 during training.

$$\begin{aligned}\mathcal{L}_{rec/cycle} = \lambda_{rot}\mathcal{L}_{rot} + \lambda_{pos}\mathcal{L}_{pos} + \lambda_{ori}\mathcal{L}_{ori} \\ + \lambda_{vel}\mathcal{L}_{vel} + \lambda_{avel}\mathcal{L}_{avel} + \lambda_{KL}\mathcal{L}_{KL} \\ + \lambda_{OCEAN}\mathcal{L}_{OCEAN}\end{aligned} \tag{5.3}$$

We employ an adversarial methodology via a separate *critic* module to ensure the local motion is correctly influenced by the provided target personality. We utilize Wasserstein GAN [116] with a gradient penalty (GP) approach to ensure the *generator* can synthesize motion consistent with the target personality even when fed with random, unseen personalities. In addition to our *generator*, we train a separate critic network, which distinguishes real and synthesized motion distributions. During our adversarial training strategy, the *generator* tries to synthesize motions with random personalities similar to real motions. On the contrary, the *critic* attempts to maximize the difference between the aforementioned motion's distributions, calculated as the Earth Mover's distance. Real motions receive a much higher score than synthesized motions, unless distributions become similar. The loss for the *critic* module, $\mathcal{L}_C$, is given in Equation 5.4.

Gradient updates for the *critic* are also subject to its losses to ensure stable training. As part of the gradient penalty, synthesized motions are randomly interpolated with real motions and fed to the critic. The calculated gradients'

norm is penalized via its associated loss function, which enforces Lipschitz continuity (Equation 5.5). This way, the *critic* is deterred from destabilizing the training process due to large fluctuations in gradient updates. Additionally, the *critic* provides feedback on synthesized motions by returning a negative loss to the *generator* as $\mathcal{L}_{GC}$ (Equation 5.6), where the generator is expected to increase the synthetic motion's score. The *critic*'s implementation is similar to the *generator*'s; it consists of identical encoders and fewer fully connected blocks.

$$\mathcal{L}_C = \frac{1}{n}\sum_{i=1}^{n} C(\hat{\mathbf{x}}_i) - \frac{1}{n}\sum_{i=1}^{n} C(\mathbf{x}_i) + \lambda_{gp}\mathcal{L}_{gp} \tag{5.4}$$

$$\mathcal{L}_{gp} = \frac{1}{n}\sum_{i=1}^{n} \left(\|\nabla_{\tilde{\mathbf{x}}_i} C(\tilde{\mathbf{x}}_i)\|_2 - 1\right)^2 \tag{5.5}$$

$$\mathcal{L}_{GC} = -\frac{1}{n}\sum_{i=1}^{n} C(\hat{\mathbf{x}}_i) \tag{5.6}$$

where C is the critic, $\mathbf{x}_i$ denotes the true sample, $\hat{\mathbf{x}}_i$ denotes generated (fake) sample, and $\tilde{\mathbf{x}}_i$ is an interpolated sample used for gradient penalty. λ_{gp} was set to 10 during training.

In addition to the *critic*, we trained a *regressor* module to provide feedback for apparent personality, object, and action types to *generator*. This module works solely on local features of the motion, with a similar architecture to *generator*, and is utilized in multiple stages during training. Before the *generator*'s losses are calculated, the *regressor* is fed with real motions and their ground truth personality, object type, and action labels. Object type and action labels are converted into one-hot encoding during preprocessing. The regressed values are part of the *regressor*'s loss $\mathcal{L}_{reg}$. Later, each motion synthesized by the aforementioned *generator* is processed by the *regressor*. Each motion's loss is compared against its expected value. Motions synthesized with ground truth, random labels, and cycled motions are subject to separate loss terms:

OCEAN Regression loss ($\mathcal{L}_{OCEANReg}$): $L2$ norm of the difference between the predicted and ground truth OCEAN personality factors. $\lambda_{OCEANReg}$ is set as 10 during training.

Object name loss ($\mathcal{L}_{name}$): Cross-entropy loss between the predicted and ground truth object name. λ_{name} is set as 5 during training.

Object intent loss ($\mathcal{L}_{intent}$): Cross-entropy loss between the predicted and ground truth object intent. λ_{intent} is set as 5 during training.

$$\begin{aligned}\mathcal{L}_{reg} = \lambda_{OCEANReg}\mathcal{L}_{OCEANReg} \\ + \lambda_{name}\mathcal{L}_{name} + \lambda_{intent}\mathcal{L}_{intent}\end{aligned} \tag{5.7}$$

$$\begin{aligned}\mathcal{L}_{GR} = \lambda_{cycleReg}\mathcal{L}_{cycleReg} \\ + \lambda_{reconReg}\mathcal{L}_{reconReg} \\ + \lambda_{randomReg}\mathcal{L}_{randomReg}\end{aligned} \tag{5.8}$$

$$\begin{aligned}\mathcal{L}_{G} = \mathcal{L}_{rec} + \lambda_{cycle}\mathcal{L}_{cycle} \\ + \lambda_{critic}\mathcal{L}_{GC} + \mathcal{L}_{GR}\end{aligned} \tag{5.9}$$

where $\mathcal{L}_{reg}$ (Equation 5.7) is applied to the *regressor* while $\mathcal{L}_{GR}$ (Equation 5.8) is added to the overall loss of the *generator* ($\mathcal{L}_{G}$) (Equation 5.9). $\mathcal{L}_{cycleReg}$, $\mathcal{L}_{reconReg}$, and $\mathcal{L}_{randomReg}$ all share the same formulation with $\mathcal{L}_{reg}$, where they only differ in input motion. We set $\lambda_{cycleReg}$, $\lambda_{reconReg}$, and $\lambda_{randomReg}$ as 5, λ_{cycle} as 0.1, and λ_{critic} as 1 for training.

During training, each latent size for features is: 512 for *object*, 256 for *personality*, and 1024 for *local*. Number of latent blocks in NeMF after latent is added: four for *object*, two for *personality*, and 11 for *local*. We applied Xavier initialization [117] for training and used Adamax optimizer [118] for all modules. We used a learning rate of 10^{-4} for the *Generator* and *Regressor* and 2×10^{-5} for the *Critic* with one iteration per generator iteration. We trained all modules for 200 epochs with a batch size of 64 using an A100 Nvidia GPU. We only applied weight decay with 10^{-7} to the generator module. We used cyclical KL annealing as a KL scheduler [119].

During our preliminary training, we focused on the reconstruction performance of the network, based on initial hyperparameters from the original NeMF. After

receiving satisfying results, we incorporated the critic into our framework and tuned its iteration count and learning rate to ensure competitive yet ultimately successful training for the generator. Our regression-related parameters are set according to our further experiments, where we focused on performance related to expressiveness. After training our model, we apply latent optimization to novel motions during synthesis. The first step of the optimization extracts the latent vectors for all provided inputs. Then, while fixing *personality* and *object* latent variables, the *local* latent is optimized to maximize reconstruction performance according to the original motion. Then, to ensure proper personality alteration, the process is repeated to minimize the loss from the *regressor* module according to the target personality. In our experiments, we optimized the reconstruction and target personality for 500 and 100 steps, respectively.

Chapter 6

User Study

6.1 Dataset Annotation

We developed a 3D annotation program using Unity Engine [120] to label non-augmented motion sequences in our custom dataset. The program supports two languages, Turkish and English, and allows for the transition at any time during the user study. The program is developed for the web and can stream many highly detailed motions and meshes efficiently. Each animation is loaded in FBX format, including all of its components. We performed a user study where each participant was tasked to annotate seventeen random animations. If the participant logs out, they can continue annotating for the remaining tasks later [1].

For better interpretability, we added a visualization for the resulting trajectory of the object as a post-processing step in Blender [21]. The participant is shown a random animation after logging in with their email. The participants are provided with five independent sliders representing each trait of the OCEAN personality model. Each trait can be annotated in scale $[-3, 3]$, which is the normalized form of the Ten-Item Personality Inventory (TIPI) [121]. The participants could

[1]Bilkent University Ethical Committee for Human Research approved the user study with the decision number İAEK_2024_07_26_01.

rotate around and zoom into the object and control the time of the animation for a better view (see Figures 6.1 and 6.2).

We asked additional questions to rate the motion realism, action label accuracy, object attributes (temperature, weight, softness), and the animation elements that influence the participant's decision the most (body pose, gaze, object trajectory, or finger articulations). We record the participant's annotation duration for each motion.

6.2 Comparison User Study

We extended our annotation user study for simultaneous comparison between five motions. Like the first study, participants are asked to annotate each motion according to the OCEAN factor in a normalized form of the Ten-Item Personality Inventory (TIPI) scale [-3, 3]. Additionally, realism and accuracy relative to the shown action label are evaluated. Under each motion, a slider is provided for the relevant question with its appropriate scale. The participant can control the camera and animation time of the motions simultaneously. At first log in and during the study, participants are provided with a visual help screen for information purposes. Figure 6.3 shows a screenshot from our comparison user study.

To be used in comparison, we selected a subset of motions from our dataset. We determine a single motion, mostly "natural" for each object type, based on its OCEAN annotations. We calculated the overall magnitude of the OCEAN traits and selected motions with a magnitude closest to 0. We altered their isolated factors to assess the authoring performance of our augmentation module and NeMF [12]-based neural network. Compared to complex, multi-factor alteration, we can easily identify differences between approaches via isolated alteration.

We altered each OCEAN trait separately for selected motions and combined them in separate tasks for isolated evaluation. We assigned each participant

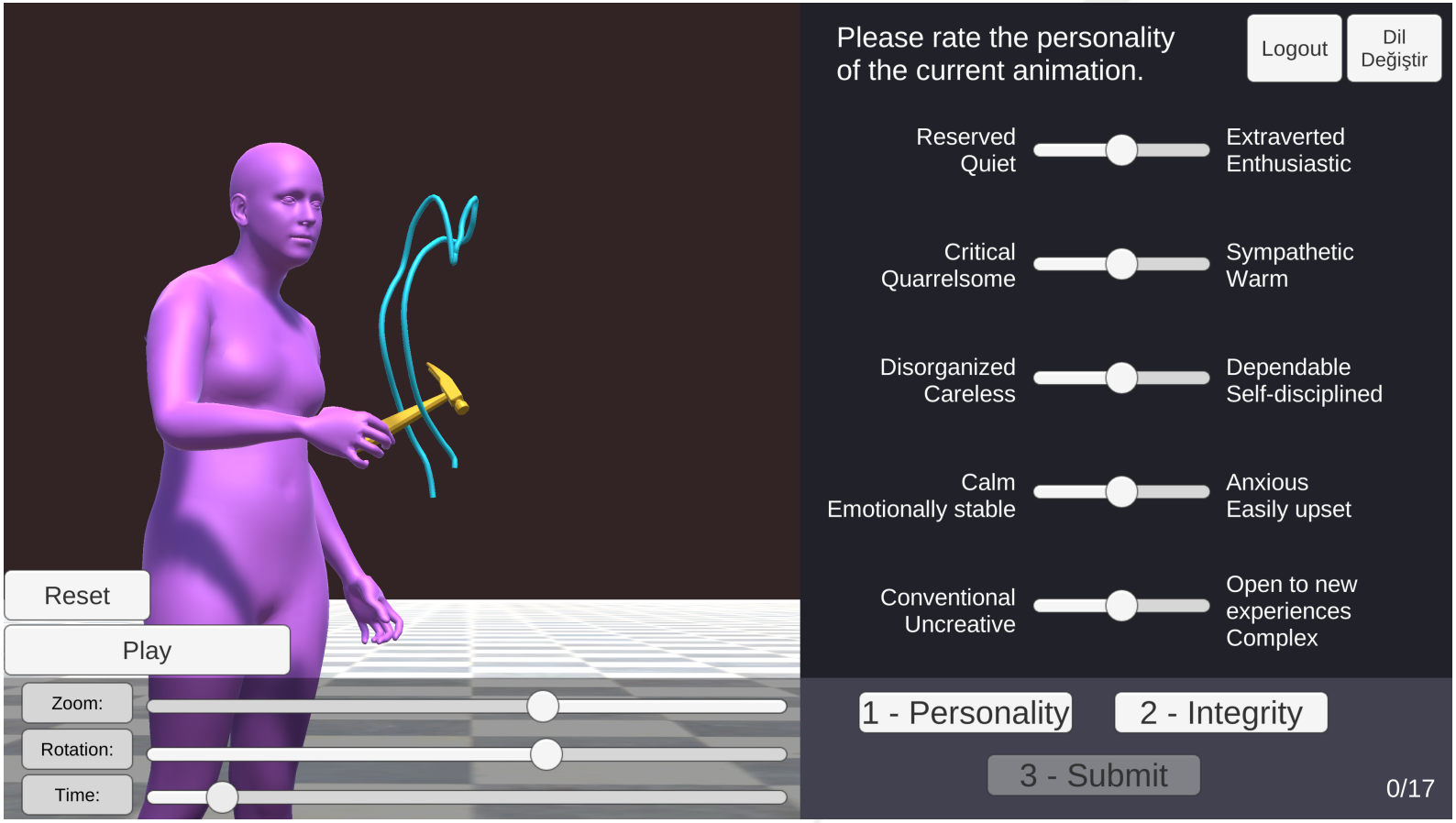

Figure 6.1: The motion annotation tool. The blue lines represent the trajectory of the object's center. On the right, users are given sliders associated with an OCEAN factor.

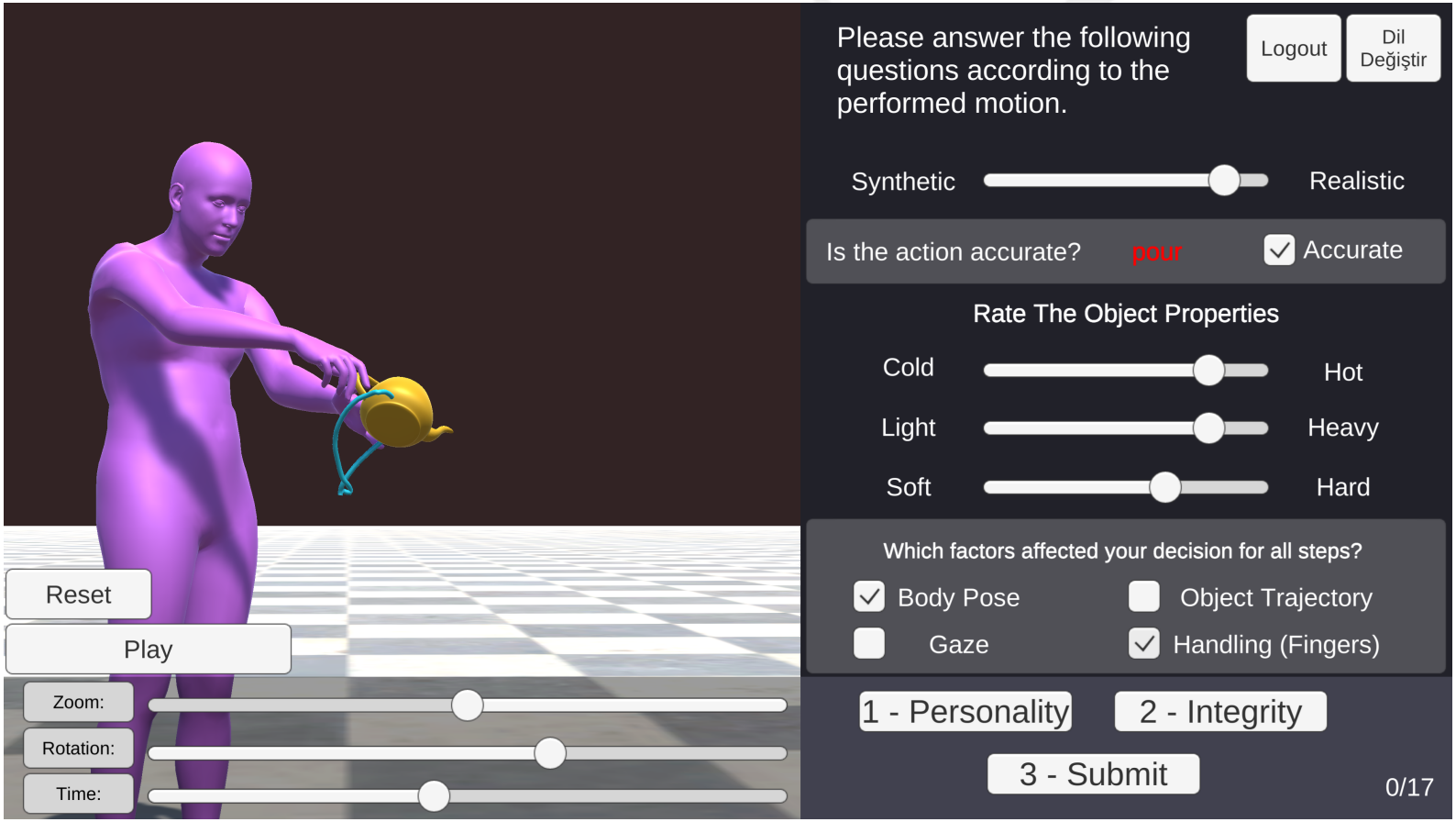

Figure 6.2: Questions regarding the integrity of the motion in the annotation tool.

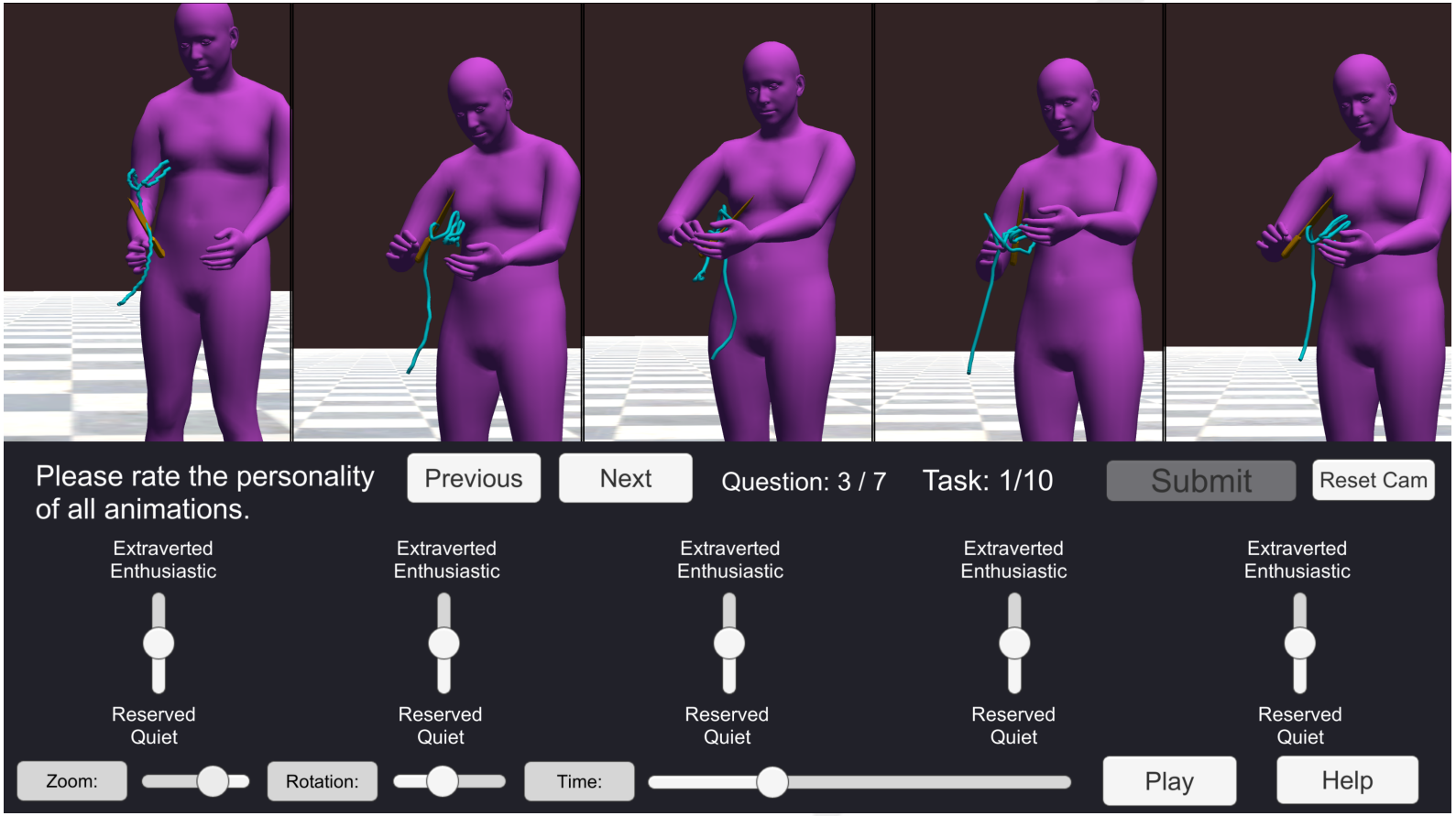

Figure 6.3: The comparison user study visualizes five motions simultaneously and includes camera and animation time controls. Each motion is annotated separately for OCEAN traits and realism. In case the user requires help, a separate help window is shown.

ten tasks, with each task containing five motions. We altered their OCEAN trait via LMA Effort parameters for two motions. To calculate the required parameters based on the altered trait, we used Table 6.1, which is based on the study by Durupınar et al. [86]. We first determine the difference between the annotated trait value and both $-1, 1$. For example, an *Openness* with -0.2 will get -0.8 and 1.2 delta values. We negate all the signs in the original OCEAN2LE matrix [86] and multiply it by our OCEAN delta to obtain the required LMA Effort parameters. For instance, an *Openness* increase with a magnitude of 1.5 results in 1.38 in *space* and 1.39 in *flow*. We synthesized two other motions using our neural network. We set the changed trait to -1 or 1 and kept the remaining traits intact.

Effort	**O**	**C**	**E**	**A**	**N**
Space	0.920	-0.9265	0.8929	0.0	1.0
Weight	0.0	0.0	0.0	1.0	0.0
Time	0.0	0.856	-0.99	-1.0	0.97
Flow	0.931	-0.938	1.0	0.0	0.762

Table 6.1: Matrix to calculate LMA effort parameters from OCEAN delta trait values.

6.3 User Study Demographics

We recruited our participants from the Prolific platform and our local community. We provided participants a link to our online website for both user studies. We collected optional demographic data for participants in Prolific using the platform itself, while the local community is provided with an option to do so. A subset of participants in our local community completed both studies, while participants from Prolific only completed either. As the motions were selected and displayed randomly, no participant who participated in both studies showed identical motions between studies. The participants' demographics are as follows:

For the first user study:

- Gender: 50% male, 36% female and 14% not provided.
- Age: Between 18 and 65.
- Nationality: 36% Türkiye, 20% South African, 16% United States, 14% Others, and 14% not provided.

For the second user study:

- Gender: 45% male, 44% female, 11% not provided.
- Age: Between 18 and 82.
- Nationality: 36% United States, 20% United Kingdom, 8% South African, 22% Others, and 14% not provided.

Chapter 7

Results

7.1 Dataset Annotation and Training

Fifty online participants rated the personality of the chosen samples in the first study, resulting in an average of eight annotations per sample. Each participant annotated 17 randomly selected samples in an average of 96 seconds per sample. We examined the resulting personality distribution of the different object categories in Figure 7.1. The top row depicts the raw data, and the bottom row shows the results after standardization. Our standardization utilizes an optimization process to fit a normal distribution for each participant, which results in a normal distribution with $\sigma = 1$ for object interaction sequences per *OCEAN* factor, using $L1$ loss. The σ of each standardized motion's annotations and σ for the mean of all sequences are subject to the associated loss.

While specific object categories like *Hammer* and *Mug* deviate from the neutral personality, especially for *conscientiousness* and *agreeableness*, we observe less variance for *openness* and *extraversion*. This behavior is expected because object interaction involves less social and intellectual context, characteristic of these traits. In contrast, the perceived effect of *conscientiousness* relates strongly to the attention towards the object of interest, thus diverging from 0.

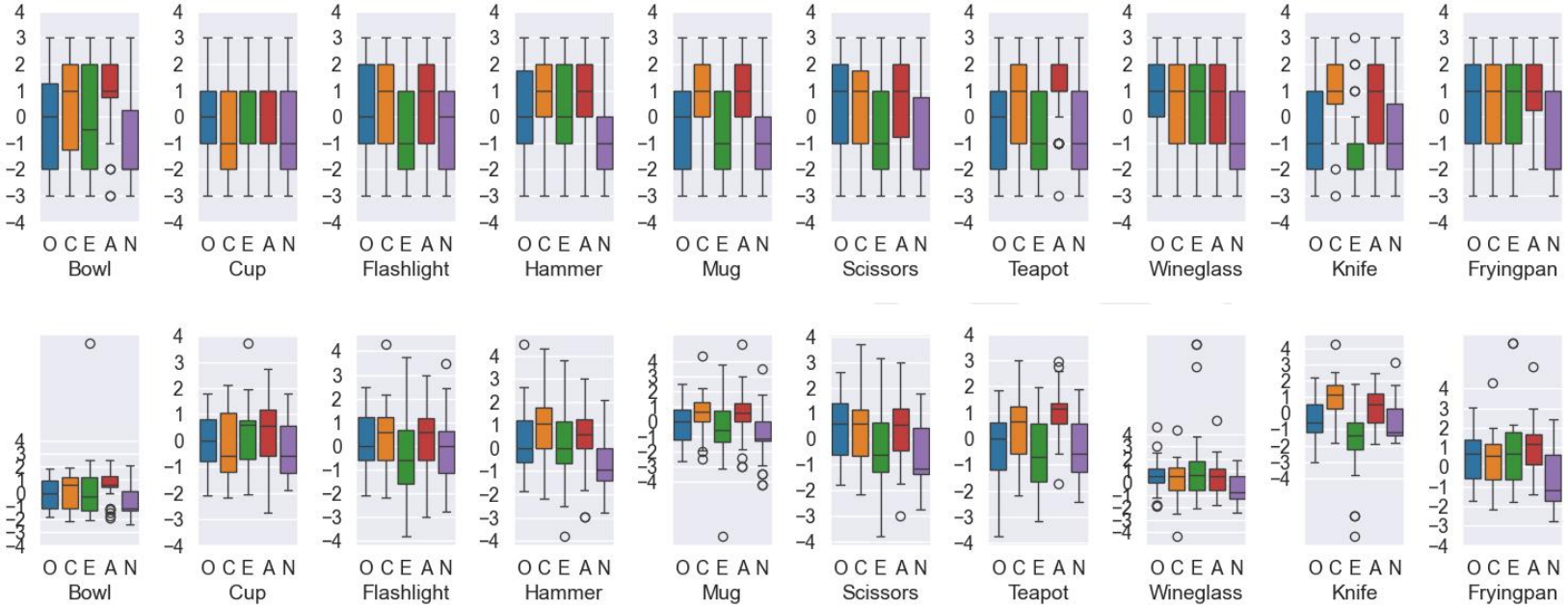

Figure 7.1: Raw (top) and standardized (bottom) annotations for each object in our annotation study. The variance for each OCEAN factor decreases after standardization, thus certain annotations become outliers while maintaining the overall diversity.

We use standardized annotations to train our generator network. To increase the scale of our limited dataset, we also applied augmentation across all LMA Efforts independently with values {-0.75, 0, 0.75}, resulting in eighty-one combinations. Upon augmentation, the resulting OCEAN is also subject to change due to changes in motion, which we described in Section 4.5. This approach dramatically increases our motion count by eighty, while introducing subtle diversity. We followed the approach in IMoS [8] to split our dataset for training. We used Subject 1's motions for the validation and Subject 10's for the test. The rest of the subjects are used for the train set. For all subjects, the augmentations of their motions are included in their relevant splits.

As our network accepts motions with a fixed length of 128 frames, we applied a sliding-window approach to each motion, as some motion lengths are not uniform. We repeat the last frame if the motion is shorter than the fixed length. After training, we apply latent optimization to novel motions during synthesis. The first step of the optimization extracts the latent vectors for all inputs. Then, while fixing *personality* and *object* latent variables, the *local* latent is optimized to minimize reconstruction errors relative to the original motion. Then, to ensure proper personality alteration, the process is repeated to minimize the loss from *regressor* module relative to the target personality. In our experiments, we optimized the reconstruction target personality for 500 and 100 steps, respectively.

We depict the effect of using different OCEAN values for the network input on the generated animation in Figures 7.2 to 7.4. We observe that *Openness* impacts the general posture of the body; High *Openness* appears as a figure standing tall. *Conscientiousness* affects the trajectory of the object's motion; High *Conscientiousness* results in a steady trajectory that helps the figure appear more confident and determined. *Extraversion* affects the general speed of the animation, which also impacts the object's trajectory; High *Extraversion* results in quick movements. *Agreeableness* impacts the posture in the vertical axis; High *Agreeableness* is portrayed as a bowing figure to signal friendliness and humbleness. *Neuroticism* affects the range of movement and, hence, the steadiness of the object's trajectory; High *Neuroticism* is portrayed with an unsteady object trajectory with anxious movements.

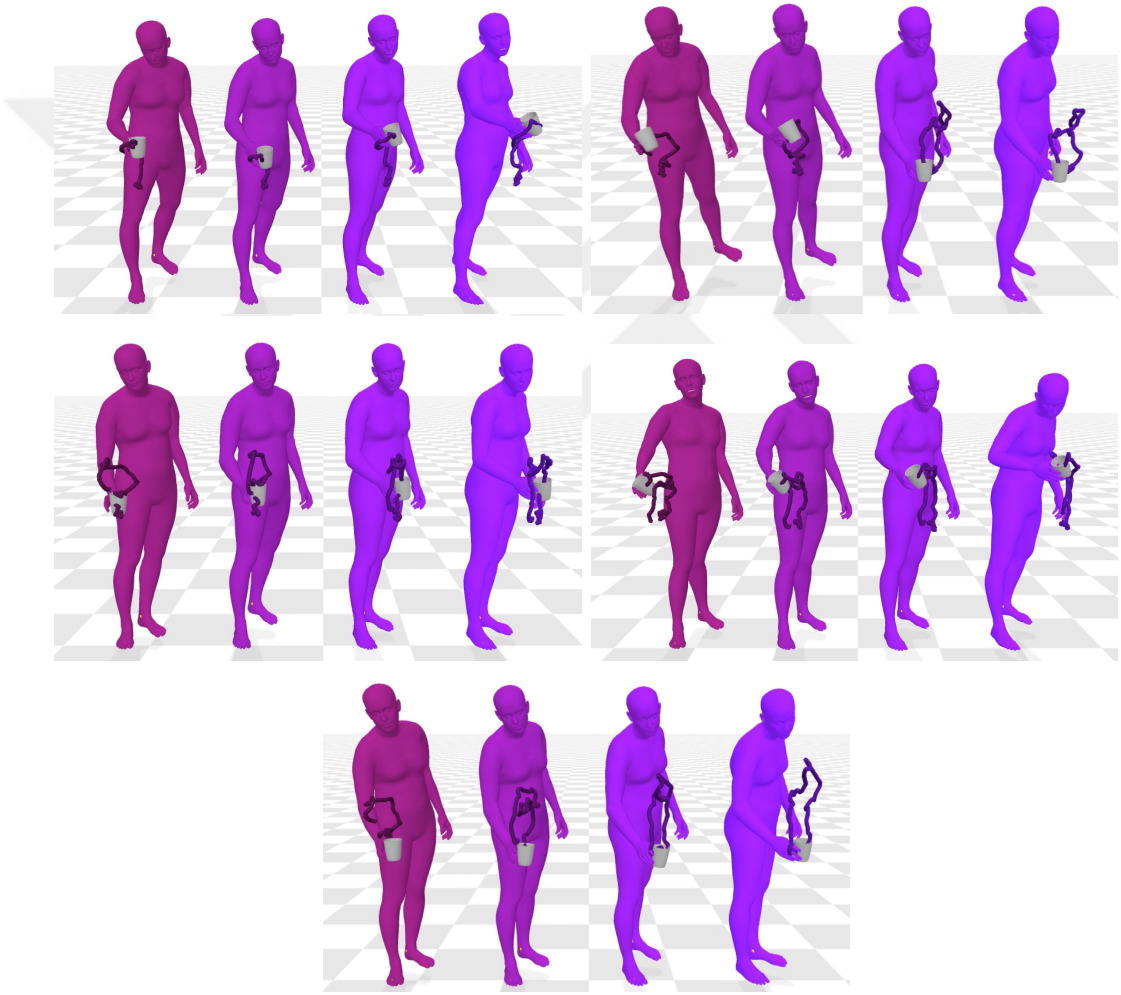

Figure 7.2: The effect of OCEAN values on the network output for “pour” action for “cup” object: From left to right, the factor ranges from -1 to 1. Each image contains the range for a specific trait. From top to bottom, left to right: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

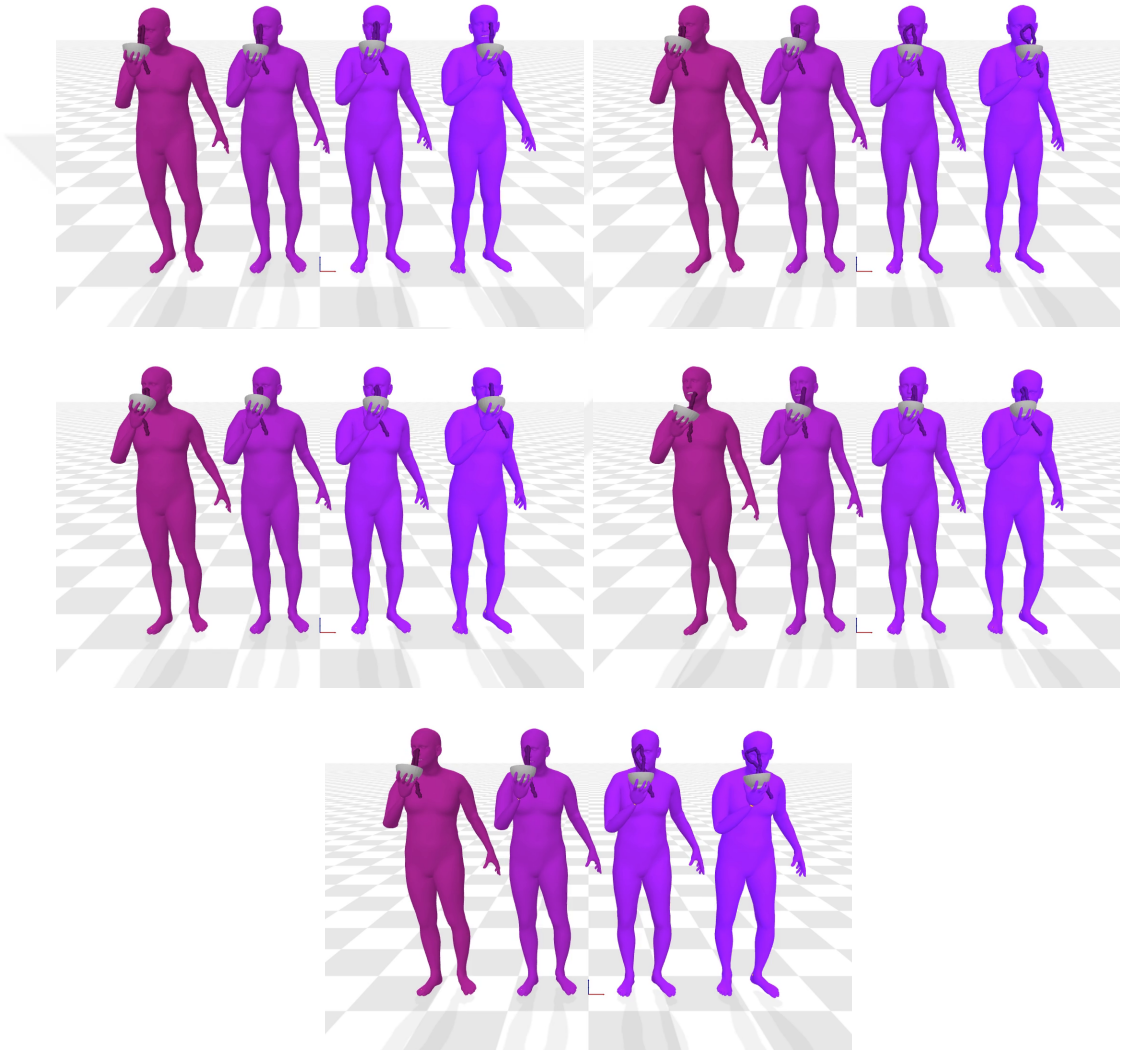

Figure 7.3: The effect of OCEAN values on the network output for "drink" action for "bowl" object: From left to right, the factor ranges from -1 to 1. Each image contains the range for a specific trait. From top to bottom, left to right: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

Figure 7.4: The effect of OCEAN values on the network output for “use” action for “hammer” object: From left to right, the factor ranges from -1 to 1. Each image contains the range for a specific trait. From top to bottom, left to right: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism.

7.2 Motion Control User Study

We generated our samples according to the process described in Section 6.2. Eighty-three online participants rated the generated samples in our second user study, corresponding to 16 samples for each comparison. We report Tukey's Honestly Significant Difference (HSD) [122] adjusted pairwise mean differences for the significant effects in Table 7.1. We compare the mean scores of the models in contrast to the base model and among each other. We expect the successful combinations to deviate significantly from the base samples and their opposite counterparts. For *Agreeableness*, for instance, neither the augmentation nor the neural network model could synthesize distinguishable differences while altering the personality of the original motion. However, all object and action types are subject to perceivable differences in at least one of the traits.

We generally observe that augmentations yield more apparent differences. The factor that is most influenced by the adjustment is *conscientiousness*, followed by *extraversion* and *neuroticism*. We observe that the effect on perceived agreeableness is minimal. Object categories such as *Bowl* and *Scissors* produce more expressive results, likely because these motions give more freedom to the actor. In contrast, categories such as *Hammer* have connotations that discourage expressiveness. For example, the first user study revealed that motions in the *Hammer* category had high *conscientiousness*. Thus, manipulation to alter this effect had less opportunity than an initially more neutral category.

We observe that augmentations differentiate negative and positive samples by exaggerating the negative case for *conscientiousness* and the positive case for *extraversion*. However, this effect adversely affects their realism; particularly, augmentations for positive traits significantly decrease both realism and action semantic accuracy. In contrast, the network's output, especially for the positive samples, has substantially higher realism and accuracy. For the negative manipulation, both approaches have similar results regarding realism and accuracy; however, augmentations yield more inconsistent results between positive and negative cases.

	Category	Base–ANeg		Base–NNeg		Base–APos		Base–NPos		ANeg–NNeg		ANeg–APos		NNeg–NPos		APos–NPos	
		ρ	Δ	ρ	Δ	ρ	Δ	ρ	Δ	ρ	Δ	ρ	Δ	ρ	Δ	ρ	Δ
O	*Bowl*	.832	-.933	.998	.267	.962	.600	.996	.333	.663	1.200	.428	1.533	.999	.067	.998	-.267
	Cup	.996	.250	.991	-.312	.968	-.438	.968	.438	.922	-.562	.851	-.688	.806	.750	.702	.875
	Flashlight	.999	-.077	.999	.077	.999	.077	.999	.000	.999	.154	.999	.154	.999	-.077	.999	-.077
	Fryingpan	.936	-.600	.694	1.000	.694	1.000	.909	-.667	.242	1.600	.242	1.600	.206	-1.667	.206	-1.667
	Hammer	.940	-.417	.999	-.083	.821	-.583	.998	-.167	.973	.333	.998	-.167	.999	-.083	.940	.417
	Knife	.999	.188	.856	.750	.999	.188	.856	.750	.944	.562	.999	.000	.999	.000	.944	.562
	Mug	.906	.786	.801	1.000	.326	1.714	.415	1.571	.999	.214	.840	.929	.969	.571	.999	-.143
	Scissors	.999	-.188	.745	-.875	.993	-.312	.999	.000	.876	-.688	.999	-.125	.745	.875	.993	.312
	Teapot	.510	1.286	.074	2.143	.019	2.571	.510	1.286	.827	.857	.510	1.286	.827	-.857	.510	-1.286
	Wineglass	.998	.235	.991	.353	.521	1.235	.991	-.353	.999	.118	.710	1.000	.897	-.706	.267	-1.588
	Drinking	.832	-.933	.998	.267	.962	.600	.996	.333	.663	1.200	.428	1.533	.999	.067	.998	-.267
	Pouring	.996	.250	.991	-.312	.968	-.438	.968	.438	.922	-.562	.851	-.688	.806	.750	.702	.875
C	*Bowl*	.000	-3.062	.627	-.938	.686	-.875	.916	-.562	.018	2.125	.013	2.188	.980	.375	.990	.312
	Cup	.551	-1.067	.729	.867	.911	.600	.671	.933	.055	1.933	.133	1.667	.999	.067	.989	.333
	Flashlight	.179	-1.867	.968	-.533	.989	-.400	.873	-.800	.505	1.333	.407	1.467	.998	-.267	.989	-.400
	Fryingpan	.563	-1.200	.360	-1.467	.999	.067	.720	-1.000	.997	-.267	.510	1.267	.977	.467	.669	-1.067
	Hammer	.054	-2.214	.388	-1.429	.992	-.357	.819	-.857	.861	.786	.149	1.857	.952	.571	.970	-.500
	Knife	.010	-2.067	.548	-.933	.999	.133	.940	.467	.351	1.133	.005	2.200	.160	1.400	.982	.333
	Mug	.060	-2.214	.624	-1.143	.868	-.786	.999	-.143	.679	1.071	.404	1.429	.732	1.000	.932	.643
	Scissors	.010	-2.533	.395	-1.333	.984	-.400	.900	-.667	.502	1.200	.045	2.133	.900	.667	.997	-.267
	Teapot	.633	-1.200	.885	-.800	.885	-.800	.913	-.733	.990	.400	.990	.400	.999	.067	.999	.067
	Wineglass	.048	-2.214	.165	-1.786	.809	-.857	.137	-1.857	.982	.429	.422	1.357	.999	-.071	.707	-1.000
	Drinking	.000	-3.062	.627	-.938	.686	-.875	.916	-.562	.018	2.125	.013	2.188	.980	.375	.990	.312
	Pouring	.551	-1.067	.729	.867	.911	.600	.671	.933	.055	1.933	.133	1.667	.999	.067	.989	.333
E	*Bowl*	.999	-.067	.999	-.133	.088	1.933	.779	.867	.999	-.067	.072	2.000	.675	1.000	.620	-1.067
	Cup	.905	-.667	.905	.667	.094	1.933	.933	.600	.410	1.333	.009	2.600	.999	-.067	.410	-1.333
	Flashlight	.848	-.750	.999	.000	.305	1.438	.986	.375	.848	.750	.033	2.188	.986	.375	.607	-1.062
	Fryingpan	.998	.188	.998	-.188	.761	.750	.999	.062	.976	-.375	.901	.562	.995	.250	.814	-.688
	Hammer	.989	-.333	.938	-.533	.989	.333	.998	.200	.998	-.200	.869	.667	.825	.733	.999	-.133
	Knife	.989	-.333	.999	.000	.006	2.467	.962	.467	.989	.333	.001	2.800	.962	.467	.041	-2.000
	Mug	.996	.333	.864	.867	.894	.800	.984	.467	.974	.533	.984	.467	.991	-.400	.996	-.333
	Scissors	.971	-.500	.883	-.750	.912	.688	.631	-1.125	.998	-.250	.581	1.188	.990	-.375	.172	-1.812
	Teapot	.995	.333	.999	.000	.082	2.200	.982	.467	.995	-.333	.191	1.867	.982	.467	.256	-1.733
	Wineglass	.969	.467	.924	.600	.373	1.333	.999	.200	.999	.133	.762	.867	.982	-.400	.538	-1.133
	Drinking	.999	-.067	.999	-.133	.088	1.933	.779	.867	.999	-.067	.072	2.000	.675	1.000	.620	-1.067
	Pouring	.905	-.667	.905	.667	.094	1.933	.933	.600	.410	1.333	.009	2.600	.999	-.067	.410	-1.333
A	*Bowl*	.999	-.067	.995	-.267	.112	-1.667	.582	-1.000	.998	-.200	.139	-1.600	.815	-.733	.862	.667
	Cup	.770	-1.000	.912	.733	.971	.533	.812	.933	.267	1.733	.389	1.533	.999	.200	.990	.400
	Flashlight	.348	-1.250	.844	-.688	.746	-.812	.514	-1.062	.918	.562	.966	.438	.981	-.375	.996	-.250
	Fryingpan	.798	-.857	.999	.071	.691	-1.000	.844	-.786	.747	.929	.999	-.143	.798	-.857	.999	.214
	Hammer	.976	.400	.976	.400	.933	.533	.998	.200	.999	.000	.999	.133	.998	-.200	.988	-.333
	Knife	.564	-1.000	.993	-.286	.632	-.929	.967	-.429	.819	.714	.999	.071	.999	-.143	.943	.500
	Mug	.997	-.267	.632	-1.067	.577	-1.133	.906	-.667	.833	-.800	.788	-.867	.985	.400	.973	.467
	Scissors	.802	-.857	.581	-1.143	.357	-1.429	.409	-1.357	.996	-.286	.947	-.571	.999	-.214	.999	.071
	Teapot	.999	.214	.984	-.429	.954	-.571	.618	-1.143	.930	-.643	.865	-.786	.901	-.714	.954	-.571
	Wineglass	.979	-.375	.996	.250	.979	-.375	.999	-.125	.877	.625	.999	.000	.979	-.375	.996	.250
	Drinking	.999	-.067	.995	-.267	.112	-1.667	.582	-1.000	.998	-.200	.139	-1.600	.815	-.733	.862	.667
	Pouring	.770	-1.000	.912	.733	.971	.533	.812	.933	.267	1.733	.389	1.533	.999	.200	.990	.400
N	*Bowl*	.999	.000	.838	.733	.000	3.667	.791	.800	.838	.733	.000	3.667	.999	.067	.001	-2.867
	Cup	.999	-.200	.999	-.200	.868	.800	.900	-.733	.999	.000	.745	1.000	.967	-.533	.349	-1.533
	Flashlight	.872	-.800	.904	-.733	.705	1.067	.904	.733	.999	.067	.178	1.867	.406	1.467	.995	-.333
	Fryingpan	.998	.286	.828	.929	.926	.714	.988	.429	.948	.643	.988	.429	.979	-.500	.998	-.286
	Hammer	.975	-.467	.999	.200	.707	1.000	.997	.267	.914	.667	.343	1.467	.999	.067	.882	-.733
	Knife	.875	.733	.205	1.667	.985	.400	.836	.800	.744	.933	.993	-.333	.792	-.867	.985	.400
	Mug	.999	.125	.138	2.062	.519	1.375	.310	1.688	.185	1.938	.610	1.250	.993	-.375	.996	.312
	Scissors	.999	-.125	.976	.438	.040	2.125	.708	.938	.940	.562	.025	2.250	.960	.500	.494	-1.188
	Teapot	.993	.357	.262	1.714	.136	2.000	.114	2.071	.497	1.357	.303	1.643	.993	.357	.999	.071
	Wineglass	.999	.133	.983	.467	.956	.600	.999	.200	.995	.333	.983	.467	.998	-.267	.990	-.400
	Drinking	.999	.000	.838	.733	.000	3.667	.791	.800	.838	.733	.000	3.667	.999	.067	.001	-2.867
	Pouring	.999	-.200	.999	-.200	.868	.800	.900	-.733	.999	.000	.745	1.000	.967	-.533	.349	-1.533
Re.	*All*	.000	-.549	.000	-.768	.000	-1.217	.000	-.705	.387	-.219	.000	-.668	.986	.063	.000	.513
Ac.	*All*	.144	-.049	.016	-.066	.000	-.136	.271	-.042	.928	-.017	.000	-.087	.796	.024	.000	.094

Table 7.1: Tukey HSD adjusted p-values (ρ) and the mean differences (Δ) between the different models for each object category and OCEAN factor. We display all values, including the insignificant results. Each column compares different model pairs: **Base** for original animation, **ANeg** and **APos** represent negative and positive changes using the augmentation framework, respectively. **NNeg** and **NPos** represent the same changes applied using the motion authoring network, respectively. We examine the factor that the system aims to alter for each personality group; for example, when the system aims to change Openness, we evaluate the performance based on perceived Openness. Realism (Re.) and Accuracy (Ac.) scores are calculated across all objects. Significant results are colored with Green for $\rho < 0.05$, and Blue for $\rho < 0.1$. We did not get any significant results for Agreeableness.

We observe that the network's effect is limited in altering the personality significantly; however, this results in less disruption in motion realism. The data-driven nature of the network helps the synthesized animation to appear more realistic and accurate. The network can alter the personality best for *conscientiousness*.

7.3 Multi Factor Personality Authoring

Although the OCEAN input can be a mix of different factors, we focused on single-trait changes to limit the number of animations in our user studies. For a single animation regarding "pour" action with "cup" object, the effect of altering multiple factors simultaneously can be seen in Figure 7.5. We altered multiple OCEAN factors across columns on all *red* agents, where each row represents a distinct keyframe. On the first column, we set *Openness*, *Extraversion*, and *Agreeableness* to 1, while setting other factors to 0 **(OEA=1 and CN=0)**. Notice how the body is positioned and the object is moved compared to the *green* agent in the rightmost column, which is the original motion. The agent leans forward, spreading its legs and arms. We set all the aforementioned factors on the second column to -1, while setting other factors to 0 **(OEA=-1 and CN=0)**. The posture becomes more upright, and the object moves within a limited space. In the third column, we also changed the rest of the factors, *Conscientiousness* and *Neuroticism*, to 1 **(OEA=-1 and CN=1)**. Due to changes in both factors, which have conflicting effects, the object trajectory becomes bizarre, yet more similar to the original trajectory. It is clear how the motion can be authored in noticeable ways when the provided factors are aligned towards similar effects.

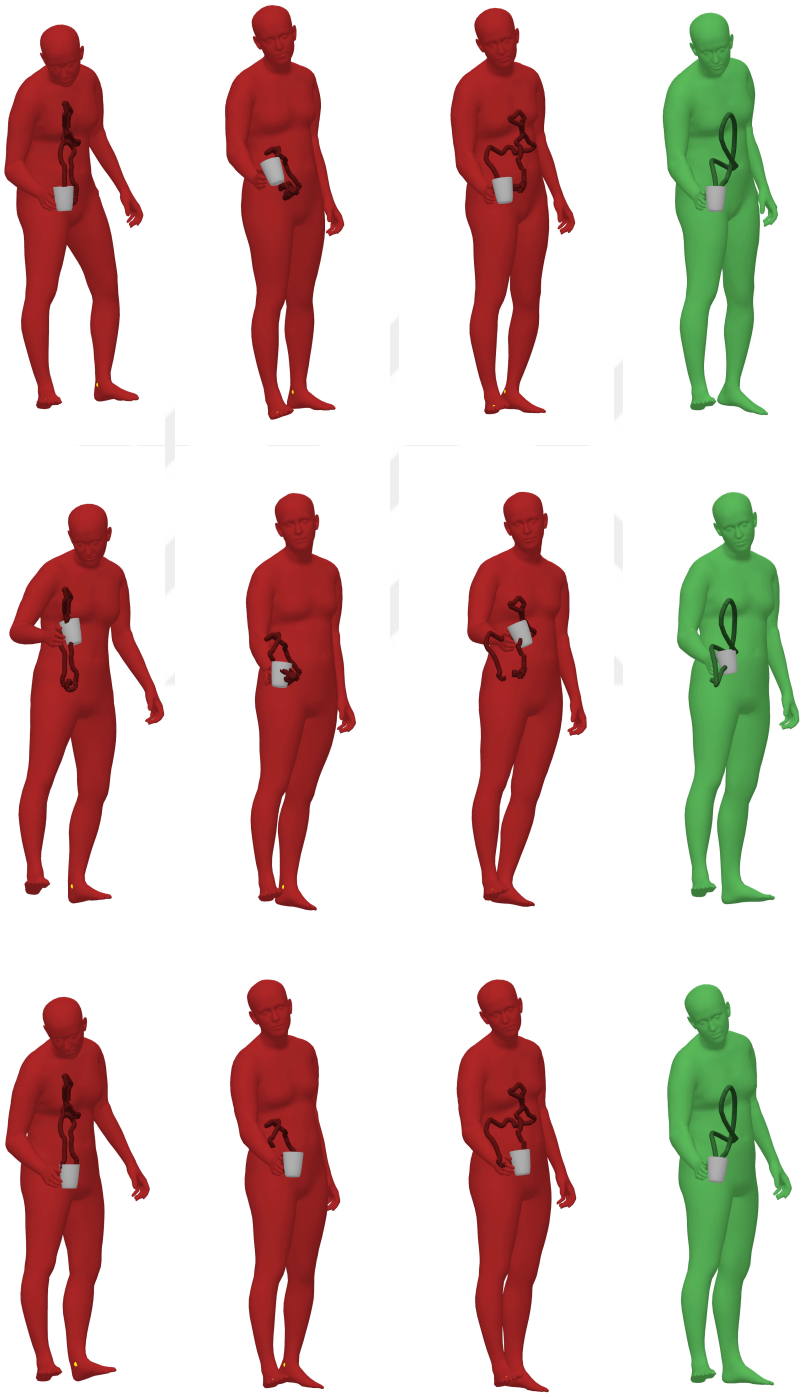

Figure 7.5: Altering multiple OCEAN factors can affect the original motion differently. From left to right as altered OCEAN factors: **(OEA=1 and CN=0)**, **(OEA=-1 and CN=0)**, **(OEA=-1 and CN=1)**, and the original unaltered motion.

7.4 Ablation and Comparison Studies

To evaluate the contribution of each component in our network, we performed ablation studies on quantitative metrics in Table 7.2. For reconstruction performance, we calculated *Joint Rotation Error* **(JRE)** and *Joint Position Error* **(JPE)**. JRE is calculated for all body joints, except the hands, as geodesic loss and is reported in degrees. JPE is calculated as the mean square distance between joints and reported in centimeters. For assessing the motion synthesis performance, we utilized *Diversity* [123] and *Fréchet Inception Distance* **(FID)** [124]. The diversity metric is calculated as the average of the latent distances between different motion instances. A high diversity value indicates highly distinct motions. FID is calculated as the Fréchet distance between latents of synthetic and real motions, thus representing the similarity between their distributions. A low FID score indicates that the synthetic motions are similar to real motions. To calculate both metrics, we used a pre-trained NeMF network. We fine-tuned the NeMF using the GRAB [7] dataset following NeMF's original training protocol. As our network only uses local features, we discarded the global motion component of NeMF. As for personality, we calculate the *Mean Square Error* (MSE) difference between the intended personality and the input. The intended personality is calculated using a regressor that is trained together with all components enabled. We also calculate the accuracy of the output motion in terms of inferred object names and action types.

We turned off a distinct module at each network variation, keeping others intact. In addition to disabling modules, we also reversed the order of the scene features provided to the NeMF module to observe the effect of order dependency between latents. The results in Table 7.2 indicate that the cycle consistency is adversarial to our network in terms of both motion reconstruction and personality recognition. However, a higher **FID** and lower *Diversity* value than the *All* configuration indicates that the generator focuses more on reconstruction than synthesis. Every other network configuration, while generating improvement in some areas, fails at generating semantically consistent interaction sequences.

Models	Motion Rec.		Motion Syn.		Personality Recognition						Object Interaction	
	JRE °↓	JPE (cm) ↓	Div ↑	FID ↓	O ↓	C ↓	E ↓	A ↓	N ↓	*All*	Name Acc. ↑	Action Acc. ↑
w/o critic	7.55	0.315	**4.95**	12.24	**0.255**	0.388	0.437	0.393	0.396	0.3738	27.0	30.0
w/o cycle	**1.44**	**0.01**	4.03	4.93	0.369	**0.195**	**0.178**	0.351	**0.362**	**0.291**	99.0	**100.0**
w/o regressor	28.82	1.305	4.11	25.45	0.467	1.048	0.216	0.206	1.079	0.6	29.0	37.0
rev. lat. ord.	7.99	0.19	4.47	10.86	0.313	0.370	0.201	**0.196**	1.938	0.6	28.0	35.0
All	1.69	0.016	4.19	**2.94**	0.340	0.205	0.475	0.279	0.454	0.35	**100.0**	**100.0**

Table 7.2: Ablation results for our NeMF-based network. Except for the "w/o cycle" option, all network versions suffer from semantic consistency regarding object interaction. The last two columns show the accuracy of the output motion in terms of inferred object names ("Name Acc.") and action types ("Action Acc."). The row "reverse latent order (rev. lat. ord.)" shows the results when the order of the scene features provided to the NeMF module is reversed to observe the effect of order dependency between latents.

To our knowledge, no personality authoring networks involving object interaction sequences exist in the literature. Therefore, we chose the closest architecture regarding object interaction generation: IMoS by Ghosh et al. [8]. IMoS generates consecutive object interaction frames over a history buffer for a given object. Even though the object name and the action type are inputs similar to ours, their method does not allow personality authoring. Additionally, their method takes considerable time due to object optimization. Our method takes 45 seconds to alter the personality, and IMoS takes 3-4 minutes to generate a motion on Nvidia 3070Ti, including all optimization protocols. Due to high time consumption, they also require interpolation between the limited number of generated frames. Although our method snaps the object to fingers and generates a fixed number of frames, it allows personality authoring, thus increasing the animation's expressiveness via high-level control. Figure 7.6 demonstrates the motions generated by our network and IMoS. As seen in the figure, the animation generated by IMoS lacks expressive elements.

7.5 Motion Integrity Evaluation

We asked questions regarding motion quality to obtain further insight into the integrity of our human-object interaction animations. These questions include:

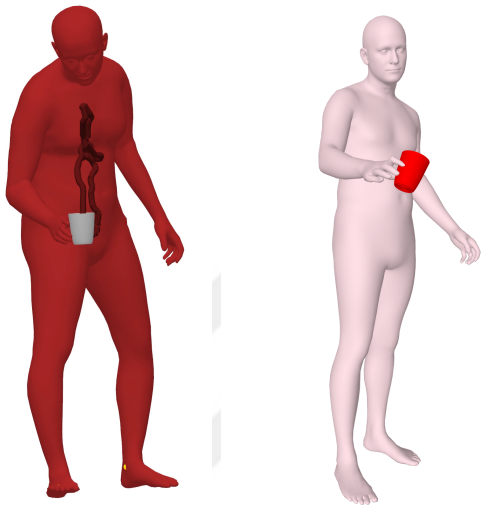

Figure 7.6: Comparison of the motions generated by our network and IMoS [8]. IMoS generates motions that lack personality and take considerable time, considering their focus on object handling. Left: our network, right: IMoS result.

- From scale 0 to 1, how realistic is the motion?
- Is the motion accurate according to its original action label?
- How hot/cold, heavy/light, and soft/hard is the object perceived?
- Which aspects of the motion inspired your decision the most: Body pose, gaze, object trajectory, or finger articulations?

Figure 7.7 depicts the user ratings regarding the perceived realism of the animation for each object category. Although the rated samples belong to the same dataset and are captured using the same setup, they received different realism scores for each object category. This behavior could be due to subjects having different performance qualities regarding objects. The capture setup includes actual objects, but they do not include secondary objects since the focus is on the interacted object. For example, the cup the subject interacts with is empty; thus, the recorded actors perform any drinking or pouring action without caring about spilling. Similarly, there is no other object in the scene for the scissors to cut.

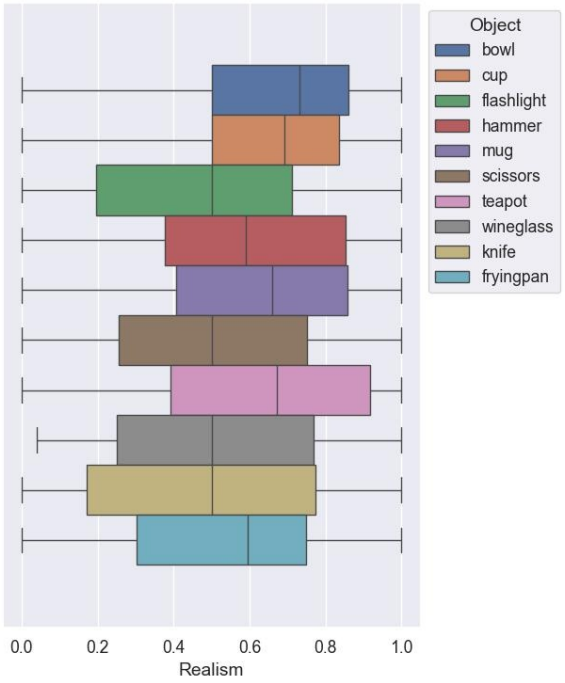

Figure 7.7: Realism score distributions for each object category in the first user study. A high realism score means the participants perceive the sample as realistic and human-like.

Consequently, we observe decreased perceived realism of object categories needing care. We observe that the flashlight and scissors categories received relatively low realism scores, likely because these objects are relatively challenging to act with without a clear objective.

Figure 7.8 shows the objects' perceived temperature, weight, and softness ratings per category. Although the animations use the same object, the participants' perception varies based on the performance. This measurement likely relates to the participant's previous experience with the object. For example, flashlights and scissors are often lighter than hammers, and we observe such results in the answers we receive. On the other hand, specific categories, like bowl and cup, received more varied answers. Our annotations include these measurements per rated sample and the personality annotations for such analysis.

Figure 7.9 shows which aspects of the rated samples were practical in participants' answers to the personality questions. Participants focused on different

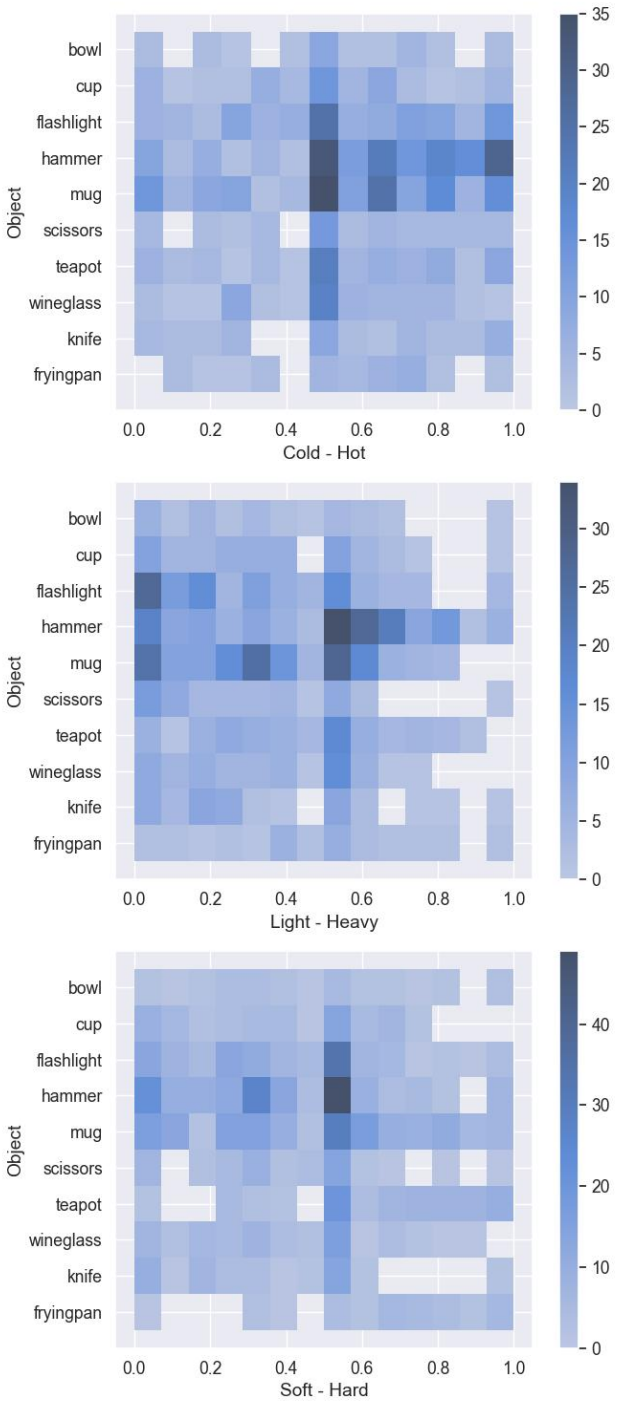

Figure 7.8: Additional object measurements for each object category. The top matrix depicts the perceived temperature (cold or hot) of the object, the middle matrix depicts the perceived weight (light or heavy), and the bottom matrix shows the perceived softness (soft or hard). The ratings do not vary much regarding softness but have distinct results for weight.

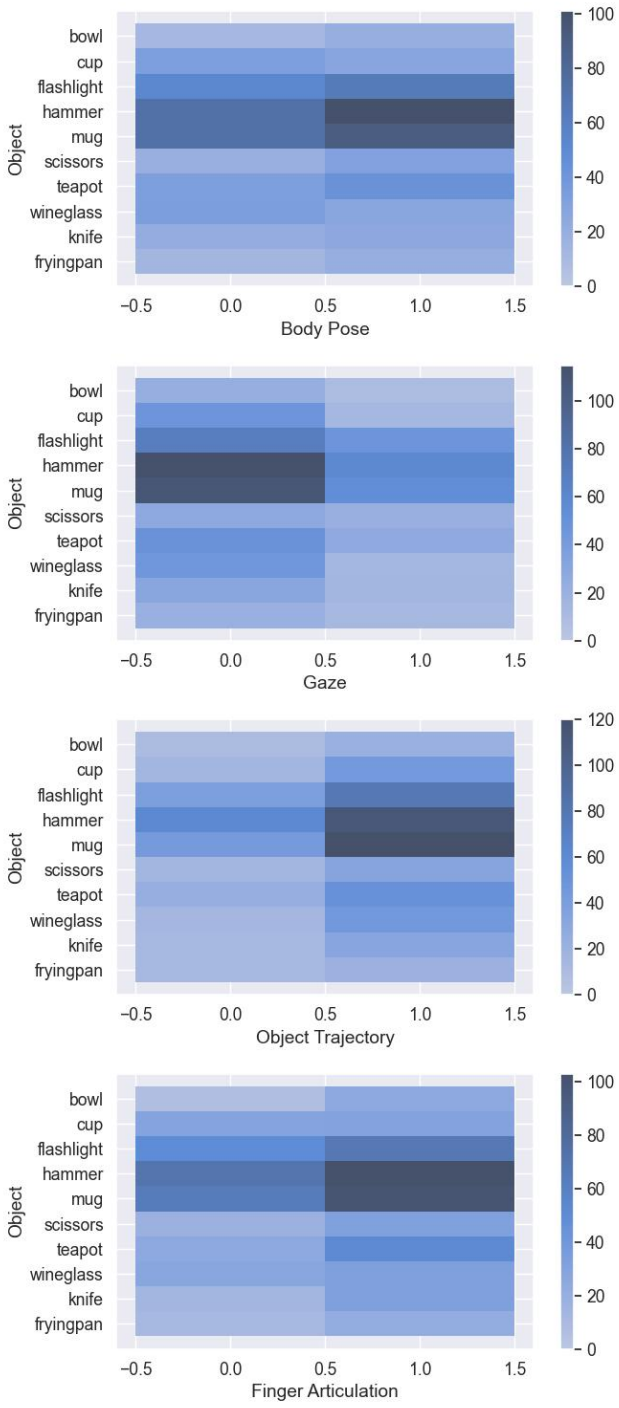

Figure 7.9: The participants' answers to the questions that ask which aspects of the motion they consider influential on their ratings. For different object categories, the animation elements that participants find influential vary.

animation elements when describing the subject's personality, which also depends on the object category. For example, body pose is more pronounced when the object interaction includes more dramatic pose changes, as in the hammer category. Gaze receives more attention when arm movements are limited. Participants found the object trajectory to be significant regardless of the object type. We also observe that finger articulation is pretty essential. Consequently, any personality manipulation in animations with object interaction should focus on redesigning the body pose and object trajectory. Our annotations include per-sample ratings for these answers so that future work can examine if the participants' personality ratings are influenced by the elements they focus on. For example, participants who concentrate on different elements may perceive the same animation as expressing different traits.

Chapter 8

Conclusions and Future Work

In this thesis, we explored the recognition and authoring of personality expression for three-dimensional human-object interaction sequences. We investigated multiple three-dimensional datasets and stated our custom subset of motions. As there are no three-dimensional object interaction datasets with personality annotations, we performed an annotation user study and quantitatively analyzed its outcomes.

Our motion augmentation protocol provides controllable scaling for our limited set of annotated motion data via LMA Effort parameters. The augmented dataset is used to train our NeMF [12] based network. Our network captures object intent, type, and the target OCEAN personality to alter a given motion's overall expression. *Critic* and *regressor* modules within the network ensure the resulting animation is plausible, realistic, and semantically consistent with the input motion.

We performed a comparison user study to assess the performance of our augmentation and network module. Our quantitative analysis revealed that network outputs tend to be more subtle yet realistic, in contrast to the erratic yet noticeable alterations of the augmentation module. Differences between altered

versions of motions are perceived more frequently across object types for *Conscientiousness*, *Extraversion*, and *Neuroticism* compared to other factors. Our visualization for network outputs aligns with our quantitative results, but also shows that our method can generate more expressive motions when controlled with multiple factors.

As we used the GRAB [7] dataset, the objects were 3D printed and lacked their counterparts' weight, temperature, and softness features. Using datasets such as Intercap [51] with real objects with accurate properties, increased diversity, and translation would further improve our interaction authoring framework. Handling grasping behavior via finger articulations could cover "pick up", "put down" and "drop" actions, thereby increasing the realism and completeness of the interaction.

As we removed gender-specific components and facial expressions from subjects in our dataset, our findings relied only on local motions. Future research should be conducted while including such distinguishing aspects of humans to cover a greater portion of personality-relevant features.

Unlike IMoS [8], our method generates a predefined number of frames, thus limiting its real-time applications. Additionally, there is no history logic, where previous frames dictate the next frame, together with features from the updated environment. Incorporating history would enable our method to "react" to events in the scene, which can also lead to emotional expression. This extension, however, requires incorporating the Pleasure-Arousal-Dominance (PAD) model [125] into human-object interaction scenarios. As literature lacks personality expression regarding object interaction, emotional behavior requires further investigation and the possibility of a novel dataset.

Bibliography

[1] J. Volanti, "The Last of Us Ashley Johnson Hotel Scene." Pinterest, 2016. Available at `https://tr.pinterest.com/pin/460563499373877566/`, Accessed: 2025/08/09.

[2] Naughty Dog, "Cinematic Process in The Last of Us - Gamescom 2012." Blog Post on Naughty Dog Website, August 2012. Available at `https://www.naughtydog.com/blog/cinematic_process_in_the_last_of_us_gamescom_2012`, Accessed: 2025/08/09.

[3] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, 2017. Article no. 194, 17 pages.

[4] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostrikov, and A. Zisserman, "The AVA-Kinetics Localized Human Actions Video Dataset," *CoRR*, vol. abs/2005.00214, 2020. Available at `https://arxiv.org/abs/2005.00214`, Accessed: 2025/08/09.

[5] Y.-W. Chao, Z. Wang, Y. He, J. Wang, , and J. Deng, "HICO: A benchmark for recognizing human-object interactions in images," in *Proceedings of the IEEE International Conference on Computer Vision*, ICCV '15, (Piscataway, NJ, USA), pp. 1017–1025, IEEE, 2015.

[6] F. Krebs, A. Meixner, I. Patzer, and T. Asfour, "The KIT bimanual manipulation dataset," in *Proceedings of the IEEE-RAS 20th International Conference on Humanoid Robots*, Humanoids '21, pp. 499–506, 2021.

[7] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *European Conference on Computer Vision - ECCV 2020, Lecture Notes in Computer Science*, vol. 12349, pp. 581–600, Springer, Cham, 2020.

[8] A. Ghosh, R. Dabral, V. Golyanik, C. Theobalt, and P. Slusallek, "IMoS: Intent-driven full-body motion synthesis for human-object interactions," *Computer Graphics Forum*, vol. 42, no. 2, 2023. Article no. cgf.14739, 12 pages.

[9] M. Mori, K. F. MacDorman, and N. Kageki, "The uncanny valley [from the field]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 2, pp. 98–100, 2012.

[10] J. J. Gibson, *The Senses Considered as Perceptual Systems*. London, UK: George Allen & Unwin Ltd., 1966.

[11] R. R. McCrae and P. T. Costa, *Personality in Adulthood: A Five-factor Theory Perspective*. New York, NY, USA: Guilford Press, 2005.

[12] C. He, J. Saito, J. Zachary, H. Rushmeier, and Y. Zhou, "NeMF: Neural motion fields for kinematic animation," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, (Red Hook, NY, USA), Curran Associates Inc., 2022. Article no. 307, 13 pages.

[13] Y. Doğan, S. Sonlu, S. Demirci, A. Ü. Ergüzen, and U. Güdükbay, "Exploring personality in human-object interactions," *Signal, Image and Video Processing*, vol. 19, no. 12, 2025. Article no. 999, 11 pages.

[14] A. Yurtoğlu, S. Sonlu, Y. Doğan, and U. Güdükbay, "Personality perception in human videos altered by motion transfer networks," *Computers & Graphics*, vol. 119, 2024. Article no. 103886, 11 pages.

[15] S. Sonlu, Y. Doğan, A. Ü. Ergüzen, M. E. Ünalan, S. Demirci, F. Durupinar, and U. Güdükbay, "İnsan hareketi kişilik tespit parametreleri (Human movement personality detection parameters) (in Turkish)," in *Proceedings*

of the IEEE Signal Processing and Communications Applications- Sinyal İşleme ve Uygulamaları Kurultayı, SİU '24, (Mersin, Türkiye), pp. 1–4, May 2024.

[16] S. Sonlu, Y. Doğan, A. Ü. Ergüzen, M. E. Ünalan, S. Demirci, F. Durupinar, and U. Güdükbay, "Towards understanding personality expression via body motion," in *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces, Workshop on Multi-modal Affective and Social Behavior Analysis and Synthesis in Extended Reality*, MASSXR '24, (Piscataway, NJ, USA), pp. 628–631, IEEE, 2024.

[17] H. Goldstein, C. Poole, and J. Safko, *Classical Mechanics*. Boston, MA: Addison Wesley, 3rd ed., 2002.

[18] K. Shoemake, "Animating rotation with quaternion curves," *ACM SIGGRAPH Computer Graphics*, vol. 19, no. 3, pp. 245–254, 1985.

[19] O. Rodrigues, "Des lois géométriques qui régissent les déplacements d'un système solide dans l'espace, et de la variation des coordonnées provenant de ces déplacements considérés indépendamment des causes qui peuvent les produire (Geometric laws that govern the displacements of a solid system in space, and the variation of coordinates resulting from these displacements considered independently of the causes that may produce them)," *Journal de Mathématiques Pures et Appliquées*, vol. 5, pp. 380–440, 1840.

[20] R. P. Paul, *Robot Manipulators: Mathematics, Programming, and Control: The Computer Control of Robot Manipulators*. Cambridge, MA, USA: MIT Press, 1981.

[21] Blender Online Community, *Blender - A 3D Modelling and Rendering Package*. Stichting Blender Foundation, Amsterdam, 2018.

[22] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. A. Osman, D. Tzionas, and M. J. Black, "Expressive body capture: 3D hands, face, and body from a single image," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '19, (Piscataway, NJ, USA), pp. 10975–10985, IEEE, 2019.

[23] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Transactions on Graphics*, vol. 34, no. 6, 2015. Article no. 248, 16 pages.

[24] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, vol. 36, no. 6, 2017. Article no. 245, 17 pages.

[25] CMU Graphics Laboratory, "CMU Motion Capture Database." Available at `http://mocap.cs.cmu.edu/`, Accessed: 2025/08/09.

[26] Movella, Inc., "Xsens 3D Motion Tracking." Available at `https://www.xsens.com/`, Accessed: 2025/08/09.

[27] Rokoko Community, "Rokoko - Studio-grade Motion Capture Tools for All Creators." Available at `https://www.rokoko.com/`, Accessed: 2025/08/06.

[28] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: realtime multi-person 2D pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2019.

[29] Q. Shuai, C. Geng, Q. Fang, S. Peng, W. Shen, X. Zhou, and H. Bao, "Novel view synthesis of human interactions from sparse multi-view videos," in *ACM SIGGRAPH Conference Proceedings*, (New York, NY, USA), ACM, 2022. Article no. 57, 10 pages.

[30] H. Lin, S. Peng, Z. Xu, Y. Yan, Q. Shuai, H. Bao, and X. Zhou, "Efficient neural radiance fields for interactive free-viewpoint video," in *ACM SIGGRAPH Asia Conference Proceedings*, (New York, NY, USA), ACM, 2022. Article no. 39, 9 pages.

[31] J. Dong, Q. Fang, W. Jiang, Y. Yang, H. Bao, and X. Zhou, "Fast and robust multi-person 3D pose estimation and tracking from multiple views," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6981–6992, 2022.

[32] J. Dong, Q. Shuai, Y. Zhang, X. Liu, X. Zhou, and H. Bao, "Motion capture from Internet videos," in *Proceedings of the European Conference on Computer Vision*, ECCV '20, pp. 210–227, Springer, 2020.

[33] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '21, (Piscataway, NJ, USA), IEEE, 2021.

[34] Q. Fang, Q. Shuai, J. Dong, H. Bao, and X. Zhou, "Reconstructing 3D human pose by watching humans in the mirror," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '21, (Piscataway, NJ, USA), pp. 12809–12818, IEEE, 2021.

[35] Y. Rong, T. Shiratori, and H. Joo, "FrankMocap: A monocular 3D whole-body pose estimation system via regression and integration," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, ICCVW '21, (Piscataway, NJ, USA), pp. 1749–1759, IEEE, 2021.

[36] H. Joo, N. Neverova, and A. Vedaldi, "Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3D human pose estimation," in *Proceedings of the International Conference on 3D Vision*, 3DV '21, (Piscataway, NJ, USA), pp. 42–52, IEEE, 2021.

[37] Y. Sun, Q. Bao, W. Liu, Y. Fu, B. Michael J., and T. Mei, "Monocular, one-stage, regression of multiple 3D people," in *Proceedings of the International Conference on Computer Vision*, ICCV '21, (Piscataway, NJ, USA), pp. 11159–11168, IEEE, 2021.

[38] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black, "Putting people in their place: Monocular regression of 3D people in depth," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '22, (Piscataway, NJ, USA), pp. 13233–13242, IEEE, 2022.

[39] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black, "PARE: Part attention regressor for 3D human body estimation," in *Proceedings of the*

International Conference on Computer Vision, ICCV '21, (Piscataway, NJ, USA), pp. 11127–11137, IEEE, 2021.

[40] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black, "Collaborative regression of expressive bodies using moderation," in *Proceedings of the International Conference on 3D Vision*, 3DV '21, pp. 792–804, 2021.

[41] Y. Ye and C. K. Liu, "Synthesis of detailed hand manipulations using contact sampling," *ACM Transactions on Graphics*, vol. 31, no. 4, pp. 1–10, 2012.

[42] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Communications of ACM*, vol. 65, no. 1, pp. 99–106, 2020.

[43] T. Kerbl, M. Kanzler, A. Kurz, and M. Steinberger, "3D Gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, 2023. Article no. 139, 14 pages.

[44] M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, SIGGRAPH '23, (New York, NY, USA), ACM, 2023. Article no. 72, 12 pages.

[45] Polycam, Inc., "Polycam: Reality capture for professionals:." Available at `https://poly.cam/`, Accessed: 2025/08/09.

[46] J. Y. Zhang, S. Pepose, H. Joo, D. Ramanan, J. Malik, and A. Kanazawa, "Perceiving 3D human-object spatial arrangements from a single image in the wild," in *Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, (Berlin, Heidelberg), pp. 34–51, Springer-Verlag, 2020.

[47] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2." Available at `https://github.com/facebookresearch/detectron2`, Accessed: 2025/08/09, 2019.

[48] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '17, (Piscataway, NJ, USA), pp. 605–613, IEEE, 2017.

[49] X. Xie, B. L. Bhatnagar, and G. Pons-Moll, "CHORE: Contact, Human and Object Reconstruction from a Single RGB Image," in *Computer Vision - ECCV 2022* (S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), (Cham), pp. 125–145, Springer Nature Switzerland, 2022.

[50] Y. Hasson, G. Varol, I. Laptev, and C. Schmid, "Towards unconstrained joint hand-object reconstruction from RGB videos," in *Proceedings of the International Conference on 3D Vision*, 3DV '21, (Piscataway, NJ, USA), pp. 659–668, IEEE, 2021.

[51] Y. Huang, O. Taheri, M. J. Black, and D. Tzionas, "InterCap: Joint markerless 3D tracking of humans and objects in interaction," in *Proceedings of the German Conference on Pattern Recognition*, vol. 13485 of *GCPR '22, Lecture Notes in Computer Science*, pp. 281–299, Springer, 2022.

[52] Y. Ye, A. Gupta, and S. Tulsiani, "What's in your hands? 3D reconstruction of generic objects in hands," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '22, (Piscataway, NJ, USA), pp. 3885–3895, IEEE, 2022.

[53] Y. Ye, P. Hebbar, A. Gupta, and S. Tulsiani, "Diffusion-guided reconstruction of everyday hand-object interaction clips," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV '23, (Piscataway, NJ, USA), pp. 19660–19671, IEEE, 2023.

[54] Z. Fan, M. Parelli, M. E. Kadoglou, M. Kocabas, X. Chen, M. J. Black, and O. Hilliges, "HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '24, (Piscataway, NJ, USA), pp. 494–504, IEEE, 2024.

[55] X. Xie, J. E. Lenssen, and G. Pons-Moll, "InterTrack: Tracking, human object interaction without object templates," in *International Conference on 3D Vision*, (Washington DC, USA), IEEE Computer Society, 2025.

[56] S. Starke, H. Zhang, T. Komura, and J. Saito, "Neural state machine for character-scene interactions," *ACM Transactions on Graphics*, vol. 38, no. 6, 2019. Article no. 209, 14 pages.

[57] X. Zhang, B. L. Bhatnagar, S. Starke, I. Petrov, V. Guzov, H. Dhamo, E. Pérez-Pellitero, and G. Pons-Moll, "FORCE: Physics-aware Human-object Interaction," 2024. Available at `https://arxiv.org/abs/2403.11237`, Accessed: 2025/08/08.

[58] M. Hassan, V. Choutas, D. Tzionas, and M. J. Black, "Resolving 3D human pose ambiguities with 3D scene constraints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV '19, (Piscataway, NJ, USA), pp. 2282–2292, IEEE, 2019.

[59] S. Zhang, Y. Zhang, Q. Ma, M. J. Black, and S. Tang, "PLACE: Proximity learning of articulation and contact in 3D environments," in *Proceedings of the International Conference on 3D Vision*, 3DV '20, (Piscataway, NJ, USA), pp. 642–651, IEEE, 2020.

[60] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3D scenes by learning human-scene interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '21, (Piscataway, NJ, USA), pp. 4703–14713, IEEE, 2021.

[61] M. Hassan, D. Ceylan, R. Villegas, J. Saito, J. Yang, Y. Zhou, and M. Black, "Stochastic scene-aware motion prediction," in *Proceedings of the International Conference on Computer Vision*, ICCV '21, (Piscataway, NJ, USA), pp. 11354–11364, IEEE, 2021.

[62] M. Hassan, Y. Guo, T. Wang, M. J. Black, S. Fidler, and X. Bin Peng, "Synthesizing physical character-scene interactions," in *ACM SIGGRAPH Conference Proceedings*, (New York, NY, USA), ACM, 2023.

[63] J. Li, A. Clegg, R. Mottaghi, J. Wu, X. Puig, and C. K. Liu, "Controllable human-object interaction synthesis," in *Computer Vision - ECCV 2024: 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLI*, (Berlin, Heidelberg), pp. 54–72, Springer-Verlag, 2024.

[64] S. Xu, Z. Li, Y.-X. Wang, and L.-Y. Gui, "InterDiff: Generating 3d human-object interactions with physics-informed diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV '23, (Piscataway, NJ, USA), pp. 14882–14894, IEEE, 2023.

[65] Z. Wang, Y. Chen, B. Jia, P. Li, J. Zhang, J. Zhang, T. Liu, Y. Zhu, W. Liang, and S. Huang, "Move as you say, interact as you can: Language-guided human motion generation with scene affordance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '24, (Piscataway, NJ, USA), pp. 433–444, IEEE, 2024.

[66] W. Zhang, R. Dabral, T. Leimkühler, V. Golyanik, M. Habermann, and C. Theobalt, "ROAM: Robust and object-aware motion generation using neural pose descriptors," in *Proceedings of the International Conference on 3D Vision*, 3DV '24, (Piscataway, NJ, USA), pp. 1392–1402, IEEE, 2024.

[67] Y. Wu, J. Wang, Y. Zhang, S. Zhang, O. Hilliges, F. Yu, and S. Tang, "SAGA: Stochastic whole-body grasping with contact," in *Computer Vision - ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI*, (Berlin, Heidelberg), pp. 257–274, Springer-Verlag, 2022.

[68] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas, "GOAL: Generating 4D whole-body motion for hand-object grasping," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '22, pp. 13253–13263, 2022.

[69] H. Zhang, Y. Ye, T. Shiratori, and T. Komura, "ManipNet: Neural manipulation synthesis with a hand-object spatial representation," *ACM Transactions on Graphics*, vol. 40, no. 4, 2021. Article no. 121, 14 pages.

[70] O. Taheri, Y. Zhou, D. Tzionas, Y. Zhou, D. Ceylan, S. Pirk, and M. J. Black, "GRIP: Generating interaction poses using spatial cues and latent consistency," in *Proceedings of the International Conference on 3D Vision*, 3DV '24), 2024.

[71] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *Proceedings of the International Conference on Machine Learning*, ICML '21, pp. 8748–8763, PMLR, 2021.

[72] J. Braun, S. Christen, M. Kocabas, E. Aksan, and O. Hilliges, "Physically plausible full-body hand-object interaction synthesis," in *Proceedings of the International Conference on 3D Vision*, 3DV '24, pp. 464–473, 2024.

[73] S. Christen, S. Hampali, F. Sener, E. Remelli, T. Hodan, E. Sauser, S. Ma, and B. Tekin, "DiffH2O: Diffusion-based synthesis of hand-object interactions from textual descriptions," in *SIGGRAPH Asia 2024 Conference Papers*, SA '24, (New York, NY, USA), Association for Computing Machinery, 2024. Article no. 145, 11 pages.

[74] J. Cha, J. Kim, J. S. Yoon, and S. Baek, "Text2HOI: Text-guided 3D motion generation for hand-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '24, pp. 1577–1585, 2024.

[75] K. Aberman, Y. Weng, D. Lischinski, D. Cohen-Or, and B. Chen, "Unpaired motion style transfer from video to animation," *ACM Transactions on Graphics*, vol. 39, no. 4, 2020. Article no. 64, 12 pages.

[76] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '19, (Piscataway, NJ, USA), pp. 4401–4410, IEEE, 2019.

[77] P. Li, K. Aberman, Z. Zhang, R. Hanocka, and O. Sorkine-Hornung, "GANimator: Neural motion synthesis from a single sequence," *ACM Transactions on Graphics*, vol. 41, no. 4, 2022. Article no. 138, 12 pages.

[78] D.-K. Jang, S. Park, and S.-H. Lee, "Motion puzzle: Arbitrary motion style transfer by body part," *ACM Transactions on Graphics*, vol. 41, no. 3, 2022. Article no. 33, 16 pages.

[79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS '17, (Red Hook, NY, USA), pp. 6000–6010, Curran Associates Inc., 2017.

[80] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '22, (Piscataway, NJ, USA), pp. 10674–10685, IEEE, 2022.

[81] A. Blattmann, R. Rombach, K. Oktay, J. Müller, and B. Ommer, "Retrieval-augmented diffusion models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, (Red Hook, NY, USA), Curran Associates Inc., 2022. Article no. 1114, 16 pages.

[82] N. Athanasiou, A. Cseke, M. Diomataris, M. J. Black, and G. Varol, "MotionFix: Text-driven 3D human motion editing," in *SIGGRAPH Asia 2024 Conference Papers*, SA '24, (New York, NY, USA), Association for Computing Machinery, 2024.

[83] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *Proceedings of the The Eleventh International Conference on Learning Representations*, ICLR '23, 2023.

[84] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of Personality*, vol. 60, no. 2, pp. 175–215, 1992.

[85] R. R. McCrae and P. T. Costa, *Personality in Adulthood: A Five-Factor Theory Perspective*. New York, NY, USA: Guilford Press, 2003.

[86] F. Durupinar, M. Kapadia, S. Deutsch, M. Neff, and N. I. Badler, "PERFORM: Perceptual approach for adding OCEAN personality to human motion using Laban Movement Analysis," *ACM Transactions on Graphics*, vol. 36, no. 1, 2016. Article no. 6, 16 pages.

[87] V. Maletic, *Body - Space - Expression*. Boston, MA, USA: De Gruyter Mouton, 1987.

[88] B. Adrian, *Actor Training the Laban Way: An Integrated Approach to Voice, Speech, and Movement*. New York, NY, USA: Allworth, 2008.

[89] Z. Erkoç, S. Demirci, S. Sonlu, and U. Güdükbay, "Skeleton-based Personality Recognition using Laban Movement Analysis," in *Understanding Social Behavior in Dyadic and Small Group Interactions* (C. Palmero, J. C. S. Jacques Junior, A. Clapés, I. Guyon, W.-W. Tu, T. B. Moeslund, and S. Escalera, eds.), vol. 173 of *Proceedings of Machine Learning Research*, pp. 74–87, PMLR, 16 October 2022.

[90] S. Sonlu, U. Güdükbay, and F. Durupinar, "A conversational agent framework with multi-modal personality expression," *ACM Transactions on Graphics*, vol. 40, no. 1, 2021. Article no. 7, 16 pages.

[91] D. Damen, H. Doughty, G. M. Farinella, , A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100," *International Journal of Computer Vision*, vol. 130, no. 1, pp. 33–55, 2022.

[92] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The EPIC-KITCHENS dataset," in *Computer Vision - ECCV 2018* (V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, eds.), (Cham), pp. 753–771, Springer International Publishing, 2018.

[93] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fründ, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," in *Proceedings of the IEEE International Conference on Computer Vision*, ICCV '17, pp. 5843–5851, IEEE Computer Society, 2017.

[94] F. Mahdisoltani, G. Berger, W. Gharbieh, D. J. Fleet, and R. Memisevic, "On the effectiveness of task granularity for transfer learning," *CoRR*, vol. abs/1804.09235, 2018. Available at `https://arxiv.org/abs/1804.09235`, Accessed: 2025/08/09.

[95] X. Xu, H. Joo, G. Mori, and M. Savva, "D3D-HOI: Dynamic 3D human-object interactions from videos," *CoRR*, vol. abs/2108.08420, 2021. Available at `https://arxiv.org/abs/2108.08420`, Accessed: 2025/08/09.

[96] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "ARCTIC: A dataset for dexterous bimanual hand-object manipulation," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '23, (Piscataway, NJ, USA), pp. 12943–12954, IEEE, 2023.

[97] B. L. Bhatnagar, X. Xie, I. Petrov, C. Sminchisescu, C. Theobalt, and G. Pons-Moll, "BEHAVE: Dataset and method for tracking human object interactions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '22, (Piscataway, NJ, USA), pp. 15914–15925, IEEE, 2022.

[98] X. Lv, L. Xu, Y. Yan, X. Jin, C. Xu, S. Wu, Y. Liu, L. Li, M. Bi, W. Zeng, and X. Yang, "HIMO: A new benchmark for full-body human interacting with multiple objects," in *Computer Vision - ECCV 2024* (A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, eds.), (Cham), pp. 300–318, Springer Nature Switzerland, 2025.

[99] J. Kim, J. Kim, J. Na, and H. Joo, "ParaHome: Parameterizing everyday home activities towards 3D generative modeling of human-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, CVPR '19, (Piscataway, NJ, USA), pp. 1816–1828, IEEE, 2025.

[100] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "ContactDB: Analyzing and predicting grasp contact via thermal imaging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '19, (Piscataway, NJ, USA), IEEE, 2019.

[101] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays, "ContactPose: A dataset of grasps with object contact and hand pose," in *Computer Vision - ECCV 2020* (A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, eds.), (Cham), pp. 361–378, Springer International Publishing, 2020.

[102] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '19, (Piscataway, NJ, USA), pp. 909–918, IEEE, 2019.

[103] F. Xiang, Y. Qin, K. Mo, Y. Xia, H. Zhu, F. Liu, M. Liu, H. Jiang, Y. Yuan, H. Wang, L. Yi, A. X. Chang, L. J. Guibas, and H. Su, "SAPIEN: A SimulAted Part-based Interactive ENvironment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '20, (Piscataway, NJ, USA), IEEE, 2020.

[104] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," *CoRR*, vol. abs/1512.03012, 2015. Available at `http://arxiv.org/abs/1512.03012`, Accessed: 2025/08/09.

[105] L. Yang, K. Li, X. Zhan, F. Wu, A. Xu, L. Liu, and C. Lu, "OakInk: A large-scale knowledge repository for understanding hand-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '22, (Piscataway, NJ, USA), pp. 20921–20930, IEEE, 2022.

[106] C. Mandery, Ö. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "Unifying representations and large-scale whole-body motion databases for studying human motion," *IEEE Transactions on Robotics*, vol. 32, no. 4, pp. 796–809, 2016.

[107] S. Prokudin, C. Lassner, and J. Romero, "Efficient learning on point clouds with basis point sets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV '19, (Piscataway, NJ, USA), pp. 4331–4340, IEEE, 2019.

[108] S. Deng, X. Xu, C. Wu, K. Chen, and K. Jia, "3D AffordanceNet: A benchmark for visual object affordance understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR '21, (Piscataway, NJ, USA), IEEE, 2021.

[109] H. Kim, S. Han, P. Kwon, and H. Joo, "Beyond the contact: Discovering comprehensive affordance for 3D objects from pre-trained 2D diffusion models," in *Computer Vision - ECCV 2024* (A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, eds.), (Cham), pp. 400–419, Springer Nature Switzerland, 2025.

[110] A. Ü. Ergüzen, S. Demirci, S. Sonlu, and U. Gudukbay, "Personality transfer in human animation: Handcrafted versus data-driven approaches," 2025. Available at SSRN: `https://dx.doi.org/10.2139/ssrn.5003652`.

[111] M. Meredith and S. Maddock, "Motion capture file formats explained," tech. rep., Department of Computer Science, University of Sheffield, 2001. Available at `https://staffwww.dcs.shef.ac.uk/people/S.Maddock/publications/MotionCaptureFileFormatsExplained.pdf`, Accessed: 2025/08/09.

[112] N. Mahmood, N. Ghorbani, N. F. Troje, G. Pons-Moll, and M. J. Black, "AMASS: Archive of motion capture as surface shapes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, ICCV '19, (Piscataway, NJ, USA), pp. 5442–5451, IEEE, 2019.

[113] Mwni, "Blender Animation Retargeting," 2021. Available at `https://github.com/Mwni/blender-animation-retargeting`, Accessed: 2025/08/09.

[114] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (Piscataway, NJ, USA), pp. 5745–5753, IEEE, 2019.

[115] L. Fussell, K. Bergamin, and D. Holden, "Supertrack: Motion tracking for physically simulated characters using supervised learning," *ACM Transactions on Graphics*, vol. 40, no. 6, 2021. Article no. 197, 13 pages.

[116] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems*, vol. 30 of *NIPS '17*, (Red Hook, NY, USA), pp. 5769–5779, Curran Associates Inc., 2017.

[117] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Artificial Intelligence and Statistics, AISTATS 2010* (Y. W. Teh and D. M. Titterington, eds.), vol. 9 of *JMLR Proceedings*, pp. 249–256, JMLR.org, 2010.

[118] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Conference Track Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015* (Y. Bengio and Y. LeCun, eds.), 2015. Available at `http://arxiv.org/abs/1412.6980`, Accessed: 2025/08/09.

[119] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating KL vanishing," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, MN, USA), pp. 240–250, Association for Computational Linguistics, 2019.

[120] Unity Technologies, "Unity - Game Development Platform," 2022. Available at `https://unity.com/`, Accessed: 2025/08/08.

[121] S. D. Gosling, P. J. Rentfrow, and W. B. Swann Jr, "A very brief measure of the big-five personality domains," *Journal of Research in Personality*, vol. 37, no. 6, pp. 504–528, 2003.

[122] H. Abdi and L. J. Williams, "Tukey's Honestly Significant Difference (HSD) Test," in *Encyclopedia of Research Design* (N. Salkind, ed.), pp. 1–5, Thousand Oaks, CA, USA: Sage Publications, 2010.

[123] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2Motion: Conditioned generation of 3D human motions," in *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, (New York, NY, USA), pp. 2021–2029, Association for Computing Machinery, 2020.

[124] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a Nash equilibrium," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS '17, (Red Hook, NY, USA), pp. 6629–6640, Curran Associates Inc., 2017.

[125] F. Durupınar, U. Güdükbay, A. Aman, and N. I. Badler, "Psychological parameters for crowd simulation: From audiences to mobs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 9, pp. 2145–2159, 2015.