# GRIP: Generating Interaction Poses Using Spatial Cues and Latent Consistency
## **Supplemental Material*

Omid Taheri[1] Yi Zhou[2] Dimitrios Tzionas[3] Yang Zhou[2]
Duygu Ceylan[2] Soren Pirk[4] Michael J. Black[1]

[1]Max Planck Institute for Intelligent Systems, Germany [2]Adobe Research [3]University of Amsterdam [4]Kiel University

In this supplemental material, we provide additional information about GRIP as mentioned in the main paper; this includes details of the method, more qualitative results, grasp analysis, and the details of the cross-grasp transfer application. In addition, to better showcase the realism of the generated hand motions and interaction with 3D objects, we provide a **Supplemental Video**. The video summarizes: (1) the problem and our motivation, (2) our method and key ideas, and (3) shows many qualitative motions generated with our method. The video results make our contribution clear in a way that is hard to capture in print.

## 1. Data Preparation

The GRAB dataset [51] is used to train our GRIP model. It is a MoCap dataset that accurately captures whole-body motions involving the manipulation of 3D objects. The body is parameterized with SMPL-X [41]. The motions are performed by $10$ participants on $51$ objects with different shapes and sizes. We withhold 5 objects for the test-set and use the rest for training and validation of the networks.

**CNet data:** CNet generates hand interaction motion based on the body and object motion in a sequence. We use all the training and test sequences from GRAB for training and testing CNet, respectively. In addition to hand-object grasp frames, we consider other motion frames of each sequence to generalize to pre-grasp and post-grasp hand poses. In total, we use $1335$ motion sequences, performed on $51$ 3D objects. To split the dataset, we use the motions performed on "mug", "apple", "camera", "binoculars", and "toothpaste" as the test set,"fryingpan", "toothbrush", "elephant", and "hand" as the validation set, and the rest as the training set. In total, we have $329K$, $52K$, and $24K$ motion frames for the training, testing, and validation set, respectively.

**RNet data:** RNet refines the motions generated from CNet, therefore, we use the output of CNet as the main data source for RNet. In addition, to model more severe penetration and interaction artifacts, we prepare a synthetic dataset by perturbing the ground-truth data in GRAB. For this, we add Gaussian noise with a standard deviation of $0.3$ to the axis-angle rotation representation of the hand poses.

**Contact Consistency Details:** Here we provide more details about the contact consistency metric. To compute this, we proceed as follows: Let $F$ denote the set of grasp frames selected from the ground truth motions. For each grasp frame $f \in F$, Let $C_g(f)$ denote the contact area on the object for the generated grasp motion and $C_t(f)$ denote the contact area on the object from the ground truth. The deviation distance for frame $f$ is computed as:

$$D(f) = \text{distance}(C_g(f), C_t(f))$$

Then the Contact Consistency is computed as the average deviation distance over all grasp frames in $F$:

$$CC = \frac{1}{|F|} \sum_{f \in F} D(f)$$

Where:

- $|F|$ represents the number of grasp frames.
- $D(f)$ represents the deviation distance for frame $f$.

**ANet data:** ANet is trained to refine noisy arm motion. To prepare the training data, we add Gaussian noise to the shoulder and elbow joints of the ground-truth motion data. The noise is added to the axis-angle rotation of the joints and has $0.01$ and $0.03$ standard deviations for the shoulder and elbow joints, respectively.

## 2. Arm Denoising Network (ANet)

For an architectural overview of ANet see Fig. S.1. As input, A-Net takes the arm motion and hand sensor features of the current Ground Truth frame along with five noisy future frames. As output it gives the denoised arm poses for the five future frames, following [52]. To ensure motion consistency between the successive frames of the denoised motions, we use the LTC algorithm similar to CNet, as explained in the main manuscript (Sec. 3.4). For this, the encoder, $E^A$, maps the input to five latent representations

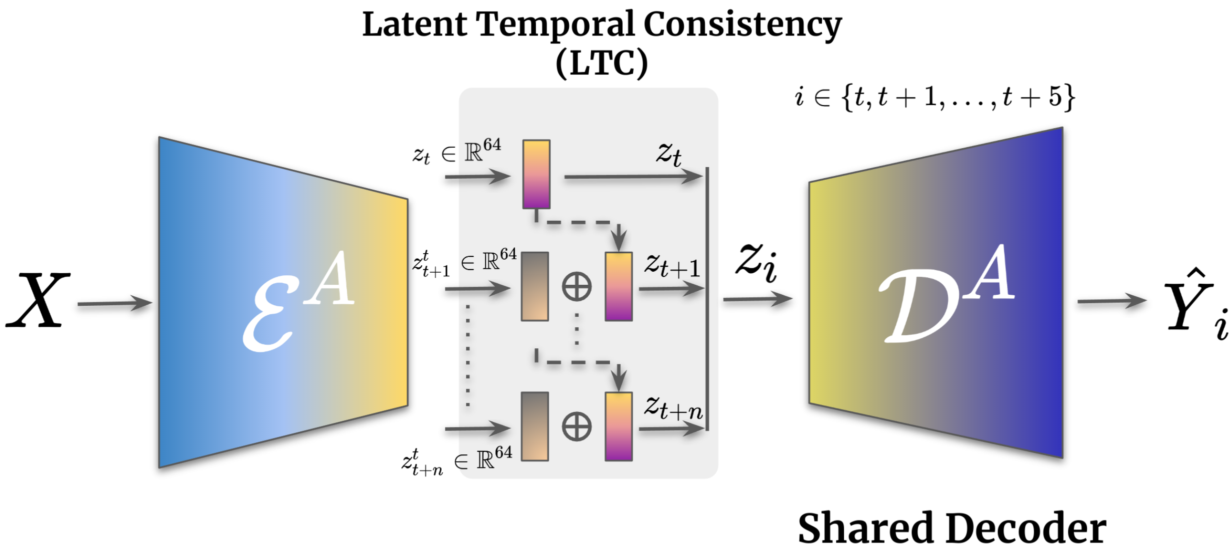


Figure S.1. Architecture overview of ANet. Similar to CNet, we use the LTC algorithm to ensure motion consistency of the denoised arm motions. For this, the encoder maps the input to a global latent code in the current frame and relative latent codes in the future frames. Then a shared decoder is used to generate the denoised motions.

for each arm pose, as shown in Fig. S.1. Then we apply the latent temporal consistency algorithm by adding the residual latent codes, $z_i^t$, to the global latent code, $z_t$. Finally, we use a shared decoder, $D^A$, to decode the denoised motions. Both encoder and decoder have $4$ fully-connected residual layers with skip connections in between.

## 3. RNet Network

For the architecture overview of RNet please see Fig. S.2. RNet takes, as input, hand poses and proximity sensor values of a motion frame and, as output, generates the refined hand poses for both left and right hand. The network consists of $4$ residual blocks with skip connections and an output linear layer.

## 4. Grasp Transfer (Application)

To test whether our method generalizes well to different object shapes and motions, we use GRIP to transfer the input interaction motion from a source object to a target object. Given a sequence of body and object motion without hand poses, we replace the source object with a target object that is roughly of the same size. We then compute the hand sensor features for the new object geometry and use GRIP to generate hand interaction poses for the new object.

Qualitative results show that our method is able to generate realistic hand motions for the target object and generalizes well to the new object's shape and motion. In Fig. S.3 we show two examples of the grasp transfer application. The top row shows that the hands adapt well to the target object geometry, "elephant", and the bottom row shows a change in the grasp type (e.g., thumb contact area) due to the smaller size of the target object, "sphere". This is useful for synthetic data generation because a single motion capture sequence can be repurposed to generate many different synthetic human-object interactions. This is also useful for FX where actors are captured handling a "dummy" object that is replaced by a 3D graphics object; this is a common scenario in film production.

## 5. Runtime

Due to its pure learning-based pipeline, GRIP is able to generate hand poses rapidly. We find that a full forward pass of our method (without ANet) on a single V100-16GB GPU, including the CNet inference, recomputing proximity sensor values, and RNet forward pass, takes $0.022$ seconds, which is equivalent to $45$ fps. Therefore, GRIP can be used to synthesize hands for avatars in interactive applications like video games and mixed reality settings, which are mostly running at 30 fps. Please notice that our network still relies on mean hand-to-object distance in the future $10$ frames, which causes a fixed 10-frame latency ($1/3$ of a second) in real-time applications. This is the trade-off to have more accurate poses with latency instead of real-time performance with lower accuracy, as shown in Tab. 4-right in the main paper.

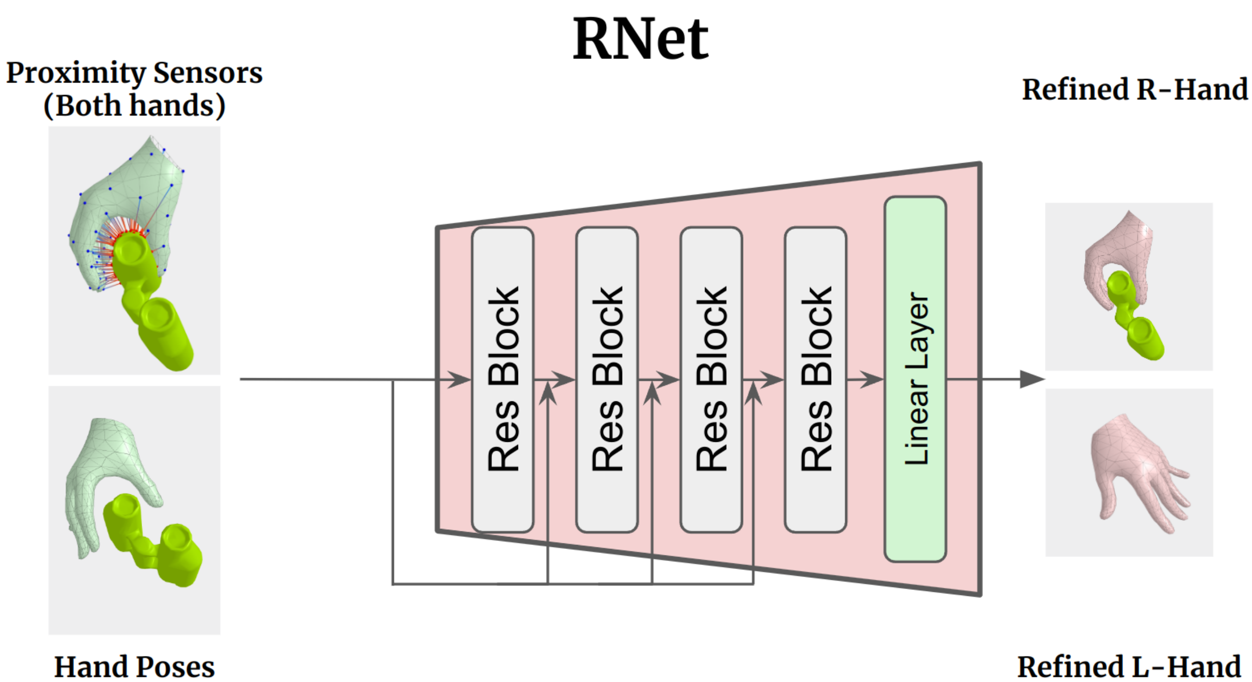


Figure S.2. Architecture overview of RNet. As input, it takes hand poses and proximity sensor values, and generates the refined hand poses. The network consists of 4 residual blocks with skip connections and an output linear layer.

## 6. Physics Simulation

Our main goal is to generate visually plausible hand-object interaction motions, however we also evaluate the physical plausibility of our results, which may be important for the real-world applications. Following prior methods [22, 23, 28], we evaluate the generated grasps in a Bullet physics simulation. We fix the body position and apply gravity to the object. A small object displacement (<1 mm) after 5 physics simulation steps is counted as a "stable" grasp. For all generated grasps, CNet and RNet have 93% and 97% stability, respectively. This suggests that the synthesized hand poses are not just visually pleasing but also physically realistic.

## 7. Performance on Large Objects

In Fig. S.4 we show more qualitative results of our method performance to generate hand grasps for large objects. Note that these objects have extended 3D structure compared with all the training objects in the GRAB dataset. What is important to note here is that our hand sensors are not distracted by the extended objects due to their locality. Thus GRIP is able to generate plausible grasps for such objects. See the **Supplemental Video** for more examples.

## 8. Qualitative Results

In Fig. S.5 we show more qualitative results generated on unseen objects, using GRIP. The top row shows input body and object motion, and the bottom row shows generated hand poses. We show close-ups of the generated hand poses, in single and bimanual scenarios, to show the accuracy of the generated grasps. In Fig. S.6 we provide results for successive frames of a motion sequence to show the consistency of the generated hand poses over time. Additionally, the results show that our method is able to refine the noisy arm poses from the InterCap dataset. For more results, please see our **Supplemental Video**.

In Fig. S.7, we show representative scores for the ManipNet grasps from our user study. These results confirm several limitations of ManipNet which GRIP addresses these, making it easy to apply in real-world scenarios.

## 9. Grasp Analysis

To further evaluate the quality of the generated grasps from GRIP, we compare the aggregated contact heatmaps from our method with GRAB [51]. For each motion frame in the test set, we compute the contact vertices on both hands based on their distance to the object surface, similar to GRAB. We then aggregate the contact maps across all frames to compute the overall contact heatmap. Figure S.8

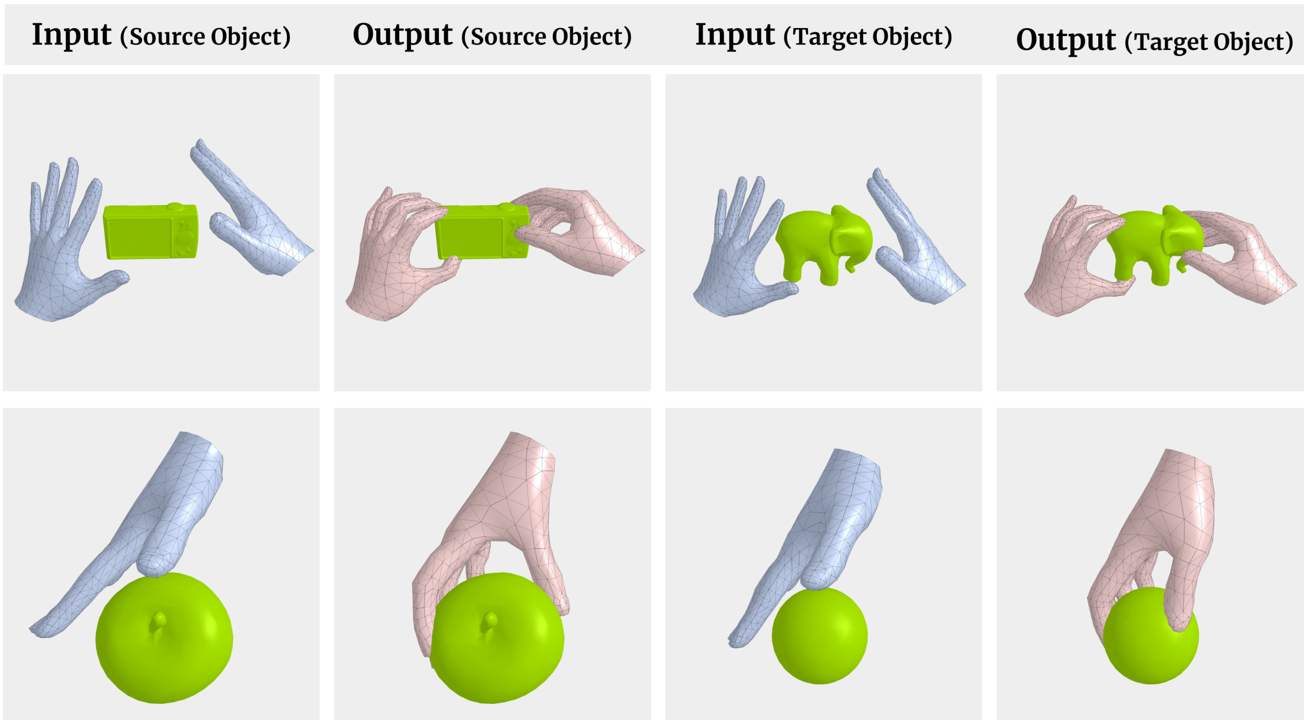


Figure S.3. Grasp transfer from a source object to a target one. Given a sequence of body and object motion without hand poses, we replace the source object with a target one and use GRIP to generate hand interaction poses for the new object. The top row shows grasp transfer from "camera" to an "elephant" geometry, and the bottom row shows grasp transfer from an "apple" to a small "sphere". Notice how the hands adapt to the new object shape (top row) and the change in the grasp type (bottom row).

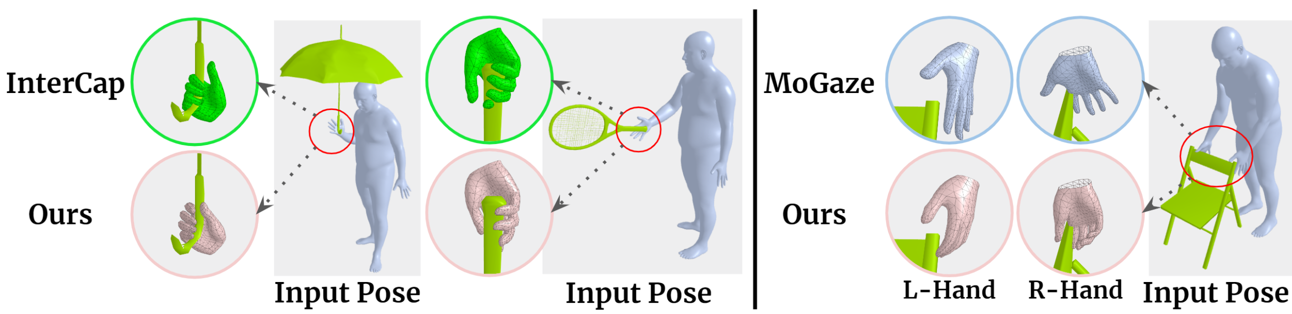


Figure S.4. GRIP's performance to generate hand grasps for large objects. We generate hand poses on the unseen large objects from Intercap (left) and MoGaze (right) datasets. These objects have larger 3D structures compared to the 3D objects during training, however, our hand sensors are not distracted by the extended objects due to their locality. Thus, GRIP is able to generate plausible grasps for such objects.

(top) shows the contact heatmap from GRAB and (bottom) shows the heatmaps for GRIP. Areas with a high likelihood of contact are shown with "hot" (red) colors and with a low likelihood of contact are shown with "cool" (blue) colors. We see that GRIP contact maps follow a similar pattern to GRAB, and have higher contact likelihood on the fingertips. The similarity suggests that generated grasps exhibit similar contacts as real grasps.

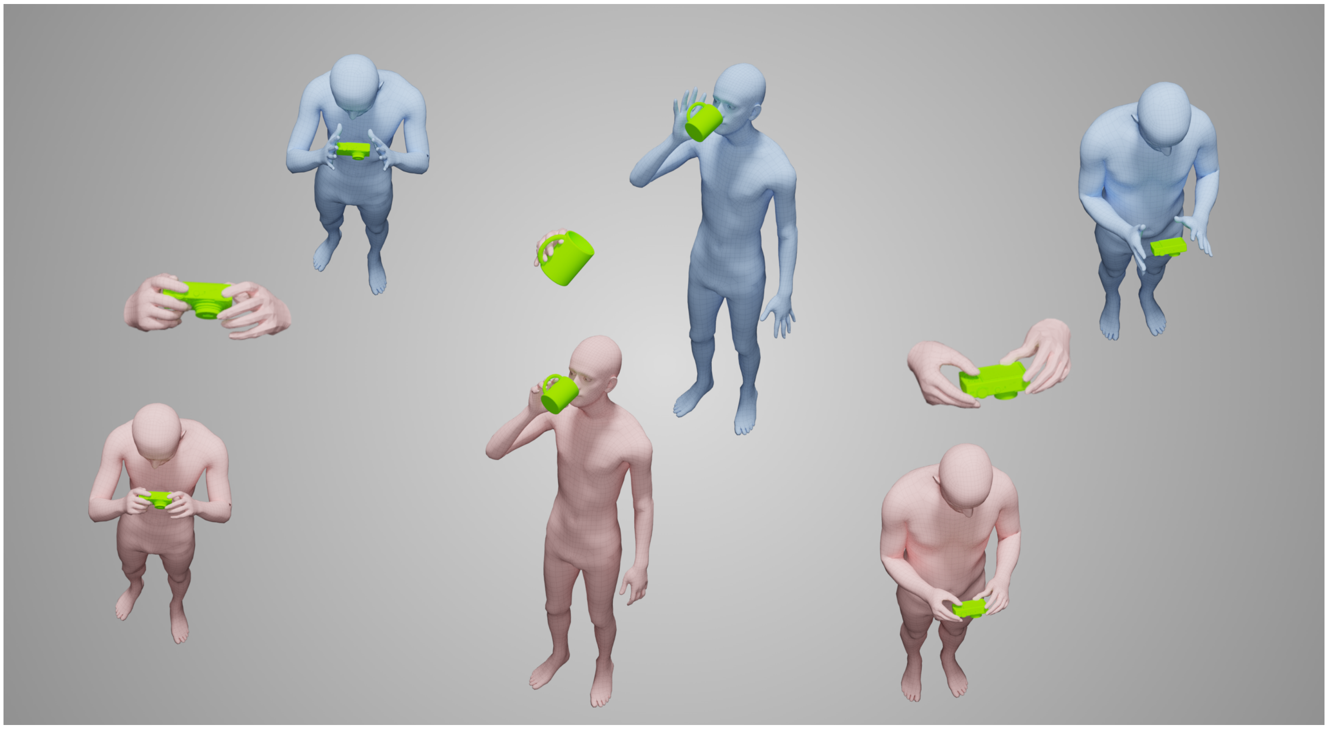

Figure S.5. Generated results with GRIP for unseen objects. (Top row) input body and object, (bottom row) generated hand poses. We show close-ups of the generated hand poses in single and bimanual scenarios, to show the accuracy of the generated grasps.

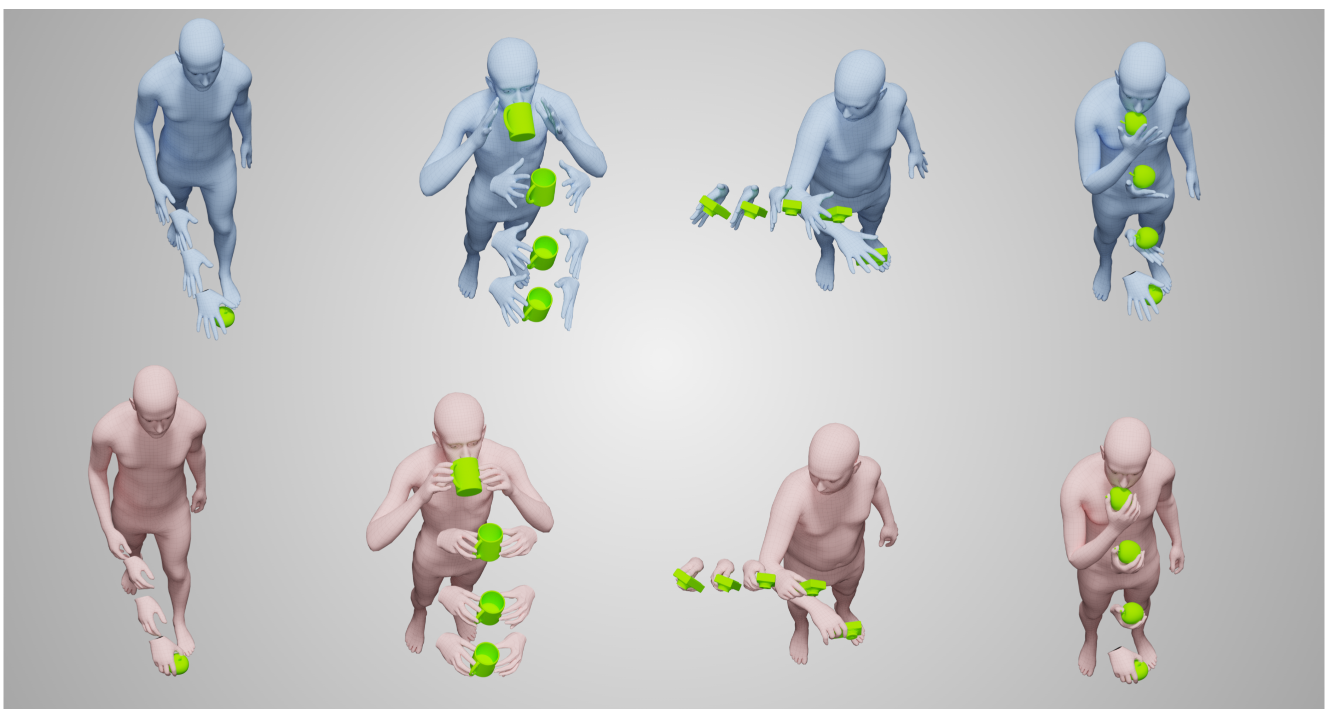

Figure S.6. Generated hand motions using GRIP. (Top row) input body and object motion. (Bottom row) generated hand poses. We provide results for successive frames of the same motion to show the consistency of the generated motions over time. Please see the Supplemental Video.

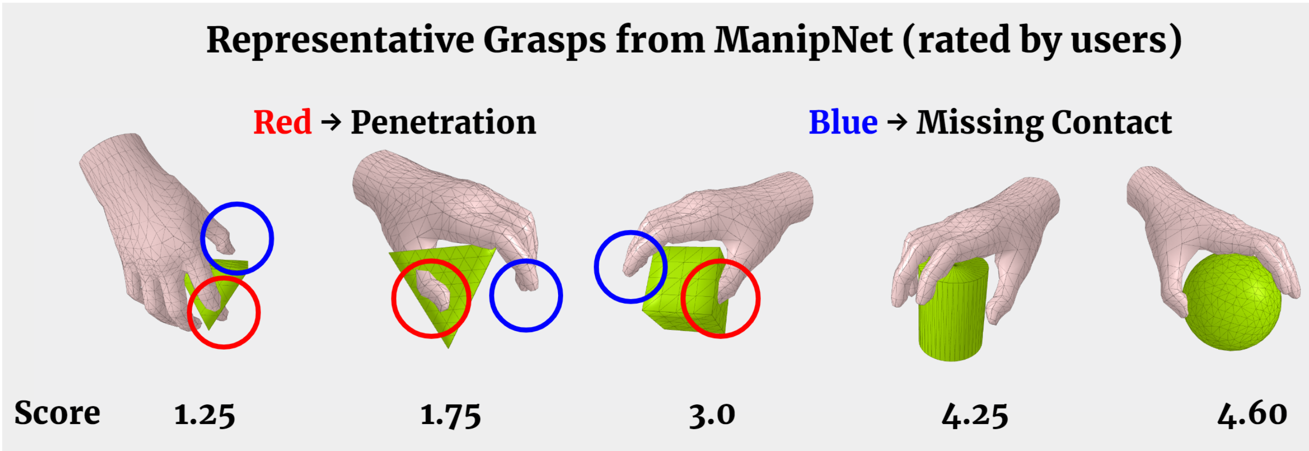


Figure S.7. representative scores for ManipNet [58] grasps from our user study.

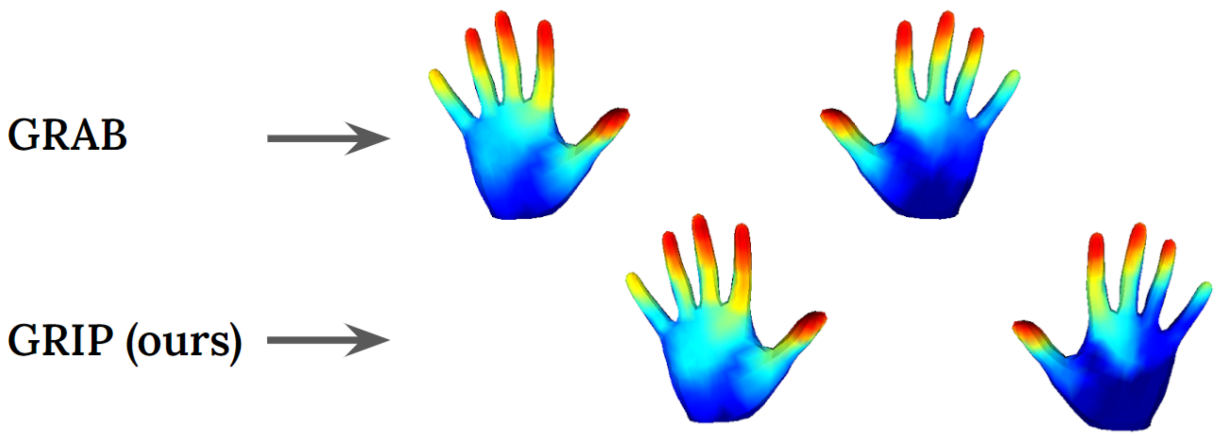


Figure S.8. Comparison of the contact heatmaps from GRAB and GRIP. We compute contact vertices on both left and right hand and aggregate them across all frames. Results show that GRIP contact maps are similar to GRAB, which is indicative of the realism of the generated hand grasps.

## References


[1] Rami Ali Al-Asqhar, Taku Komura, and Myung Geol Choi. Relationship descriptors for interactive motion adaptation. In *Symposium on Computer Animation (SCA)*, pages 45–53, 2013.

[2] Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub W. Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *International Journal of Robotics Research (IJRR)*, 39(1), 2020.

[3] Kevin Bergamin, Simon Clavet, Daniel Holden, and James Richard Forbes. DReCon: Data-driven responsive control of physics-based characters. *Transactions on Graphics (TOG)*, 38(6), 2019.

[4] Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 15935–15946, 2022.

[5] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotics*, 30:289–309, 2014.

[6] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision (ECCV)*, pages 361–378, 2020.

[8] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *Conference on Artificial Intelligence (AAAI)*, pages 5887–5895, 2021.

[9] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. *Conference on Robot Learning (CoRL)*, 2021.

[10] Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically plausible dynamic grasp synthesis for hand-object interactions. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[11] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Gregory Rogez. GanHand: Predicting human grasp affordances in multi-object scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5040, 2020.

[12] Sai Kumar Dwivedi, Cordelia Schmid, Hongwei Yi, Michael J. Black, and Dimitrios Tzionas. POCO: 3D pose and shape estimation using confidence. In *International Conference on 3D Vision (3DV)*, 2024.

[13] Sahar El-Khoury, Anis Sahbani, and Philippe Bidaud. 3D objects grasps synthesis: A survey. In *IFToMM World Congress on Mechanism and Machine Science*, 2011.

[14] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2023.

[15] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[16] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. In *International Conference on Intelligent Robots and Systems (IROS)*, pages 9561–9568, 2020.

[17] Michael Gleicher. Retargetting motion to new characters. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 33–42, 1998.

[18] Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1471–1481, 2021.

[19] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020.

[20] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael J. Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, pages 11374–11384, 2021.

[21] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 14708–14718, 2021.

[22] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11807–11816, 2019. 3

[23] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 568–577, 2020. 3

[24] Edmond S. L. Ho, Taku Komura, and Chiew-Lan Tai. Spatial relationship preserving character motion adaptation. *Transactions on Graphics (TOG)*, 29(4):33:1–33:8, 2010.

[25] Yinghao Huang, Omid Taheri, Michael J. Black, and Dimitrios Tzionas. InterCap: Joint markerless 3D tracking of humans and objects in interaction. In *German Conference on Pattern Recognition (GCPR)*, pages 281–299, 2022.

[26] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *International Conference on Computer Vision (ICCV)*, pages 11087–11096, 2021.

[27] Mubbasir Kapadia, Xu Xianghao, Maurizio Nitti, Marcelo Kallmann, Stelian Coros, Robert W. Sumner, and Markus

Gross. Precision: Precomputing environment semantics for contact-rich character animation. In *Symposium on Interactive 3D Graphics (SI3D)*, 2016.

[28] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, pages 333–344, 2020. 3

[29] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision (3DV)*, pages 11–21, 2021.

[30] Philipp Kratzer, Simon Bihlmaier, Niteesh Balachandra Midlagajni, Rohit Prakash, Marc Toussaint, and Jim Mainprice. MoGaze: A dataset of full-body motions that includes workspace geometry and eye-gaze. *Robotics and Automation Letters (RA-L)*, 6(2):367–373, 2021.

[31] Paul G. Kry and Dinesh K. Pai. Interaction capture and synthesis. *Transactions on Graphics (TOG)*, 25(3):872–880, 2006.

[32] Jehee Lee, Jinxiang Chai, Paul S. A. Reitsma, Jessica K. Hodgins, and Nancy S. Pollard. Interactive control of avatars animated with human motion data. *Transactions on Graphics (TOG)*, 21(3):491–500, 2002.

[33] Kang Hoon Lee, Myung Geol Choi, and Jehee Lee. Motion patches: Building blocks for virtual environments annotated with motion data. *Transactions on Graphics (TOG)*, 25(3): 898–906, 2006.

[34] Beatriz León, Stefan Ulbrich, Rosen Diankov, Gustavo Puche, Markus Przybylski, Antonio Morales, Tamim Asfour, Sami Moisio, Jeannette Bohg, and James Kuffner. Opengrasp: A toolkit for robot grasping simulation. In *Simulation, Modeling, and Programming for Autonomous Robots SIMPAR*, pages 109–120, 2010.

[35] Ying Li, Jiaxin L. Fu, and Nancy S. Pollard. Data-driven grasp synthesis using shape matching and task-based pruning. *Transactions on Visualization and Computer Graphics (TVCG)*, 13(4):732–747, 2007.

[36] Karen C. Liu. Dextrous manipulation from a grasping pose. *Transactions on Graphics (TOG)*, 28(3):59, 2009.

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. *Transactions on Graphics (TOG)*, 34 (6):248:1–248:16, 2015.

[38] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The KIT whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015.

[39] Igor Mordatch, Zoran Popovic, and Emanuel Todorov. Contact-invariant optimization for hand manipulation. In *Symposium on Computer Animation (SCA)*, pages 137–144, 2012.

[40] Soohwan Park, Hoseok Ryu, Seyoung Lee, Sunmin Lee, and Jehee Lee. Learning predict-and-simulate policies from unorganized human motion data. *Transactions on Graphics (TOG)*, 38(6), 2019.

[41] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1

[42] Xue Bin Peng, Glen Berseth, and Michiel Van de Panne. Terrain-adaptive locomotion skills using deep reinforcement learning. *Transactions on Graphics (TOG)*, 35(4):81:1–81:12, 2016.

[43] Xue Bin Peng, Glen Berseth, Kangkang Yin, and Michiel Van De Panne. DeepLoco: Dynamic locomotion skills using hierarchical deep reinforcement learning. *Transactions on Graphics (TOG)*, 36(4), 2017.

[44] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *Transactions on Graphics (TOG)*, 37(4):143:1–143:14, 2018.

[45] Sören Pirk, Olga Diamanti, Boris Thibert, Danfei Xu, and Leonidas J. Guibas. Shape-aware spatio-temporal descriptors for interaction classification. In *International Conference on Image Processing (ICIP)*, pages 4527–4531, 2017.

[46] Sören Pirk, Vojtech Krs, Kaimo Hu, Suren Deepak Rajasekaran, Hao Kang, Yusuke Yoshiyasu, Bedrich Benes, and Leonidas J. Guibas. Understanding and exploiting object interaction landscapes. *Transactions on Graphics (TOG)*, 36 (3):31:1–31:14, 2017.

[47] Nancy S. Pollard and Victor Brian Zordan. Physically based grasping control from example. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 311–318, 2005.

[48] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. In *Robotics: Science and Systems (RSS)*, 2018.

[49] Qijin She, Ruizhen Hu, Juzhan Xu, Min Liu, Kai Xu, and Hui Huang. Learning high-dof reaching-and-grasping via dynamic representation of gripper-object interaction. *Transactions on Graphics (TOG)*, 41(4), 2022.

[50] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *Transactions on Graphics (TOG)*, 38(6):209:1–209:14, 2019.

[51] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, pages 581–600, 2020. 1, 3

[52] Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 13253–13263, 2022. 1

[53] Julian Tanke, Chintan Zaveri, and Juergen Gall. Intention-based long-term human motion anticipation. In *International Conference on 3D Vision (3DV)*, pages 596–605, 2021.

[54] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision (IJCV)*, 118(2):172–193, 2016.

[55] He Wang, Sören Pirk, Ersin Yumer, Vladimir G. Kim, Ozan Sener, Srinath Sridhar, and Leonidas J. Guibas. Learning

a generative model for multi-step human-object interactions from videos. *Computer Graphics Forum (CGF)*, 38(2):367–378, 2019.

[56] Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. Synthesizing long-term 3D human motion and interaction in 3D scenes. In *Computer Vision and Pattern Recognition (CVPR)*, pages 9401–9411, 2021.

[57] Yuting Ye and C. Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *Transactions on Graphics (TOG)*, 31(4):41:1–41:10, 2012.

[58] He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. *Transactions on Graphics (TOG)*, 40(4):121:1–121:14, 2021. 6

[59] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3D people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6193–6203, 2020.

[60] Wenping Zhao, Jianjie Zhang, Jianyuan Min, and Jinxiang Chai. Robust realtime physics-based motion control for human grasping. *Transactions on Graphics (TOG)*, 32(6): 207:1–207:12, 2013.

[61] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. TOCH: Spatio-temporal object correspondence to hand for motion refinement. In *European Conference on Computer Vision (ECCV)*, pages 1–19, 2022.

[62] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5745–5753, 2019.

[63] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *International Conference on Computer Vision (ICCV)*, pages 15721–15731, 2021.