GOAL: Generating 4D Whole-Body Motion for Hand-Object Grasping
Supplemental Material

Omid Taheri Vasileios Choutas Michael J. Black Dimitrios Tzionas
Max Planck Institute for Intelligent Systems, Tübingen, Germany
{otaheri, vchoutas, black, dtzionas}@tue.mpg.de

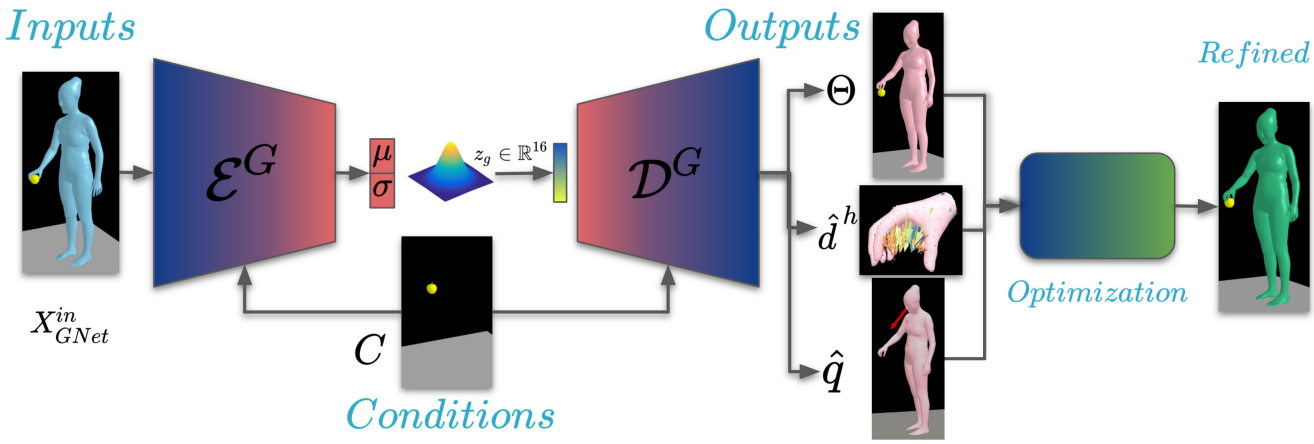

Figure S.1. Architectural overview of the GNet network, as well as the optimization post-processing step (right-most part).

The supplemental material includes this document. However, our results involve human motion and interaction with 3D objects. Therefore, we also provide a video on our website to showcase the realism of our generated grasping motions. Our video: (1) explains the problem and our motivation, (2) explains our method and key ideas, and (3) shows many results, including qualitative motion results.

1. Data Preparation

GNet – Data preparation: GNet generates static whole-body grasps. Therefore, from the GRAB dataset, we collect all frames with right-hand grasps, for which subjects grasp the object in a stable way. For this, we follow the selection criteria used for GrabNet's [3] training data. We then center the object at the origin along the horizontal plane, i.e., while preserving its height. This is important as the object height changes the body pose for grasping. In total, we collect $160K$, $26K$, and $12.5K$ frames for the training, testing, and validation set, respectively.

MNet – Data preparation: On the other hand, since MNet generates motion, from each sequence of GRAB, we gather all frames from the starting one up to the frame where the right hand first establishes a stable grasp. For this, we use the same selection criteria as above for GNet. We then create several sub-sequences by sliding a 21-frame long window over each sequence with a stride of 1 frame. For each sub-sequence, we consider the first 10 frames as "past" motion, the last 10 frames as "future" motion, and the middle one as the "current" frame. Then, following Starke et al. [2], we make all "past" and "future" frames relative to the body coordinate system of the "current" frame, while keeping the gravity direction always upward. In total, we collect roughly $40K$, $7K$, and $3K$ motion sub-sequences for the training, testing, and validation sets, respectively.

2. Network Architectures

2.1. GNet Architecture

For an architectural overview of GNet and its optimization-based post processing, see Fig. S.1. GNet has a cVAE architecture that generates a static whole-body grasp, conditioned on the given object and its location. To do this, the encoder first encodes whole-body grasps into an embedding space. Then, the decoder takes a sample from this space and outputs SMPL-X parameters, $\hat{\boldsymbol{\Theta}}$, the head direction vector, $\hat{\boldsymbol{q}}$, and hand offset vectors, $\hat{\boldsymbol{d}}^{h\rightarrow o}$, shown in the figure. We then use the interaction features, $\hat{\boldsymbol{q}}$ and $\hat{\boldsymbol{d}}^{h\rightarrow o}$, to refine the predicted SMPL-X parameters $\hat{\boldsymbol{\Theta}}$ to get a more realistic whole-body grasp.

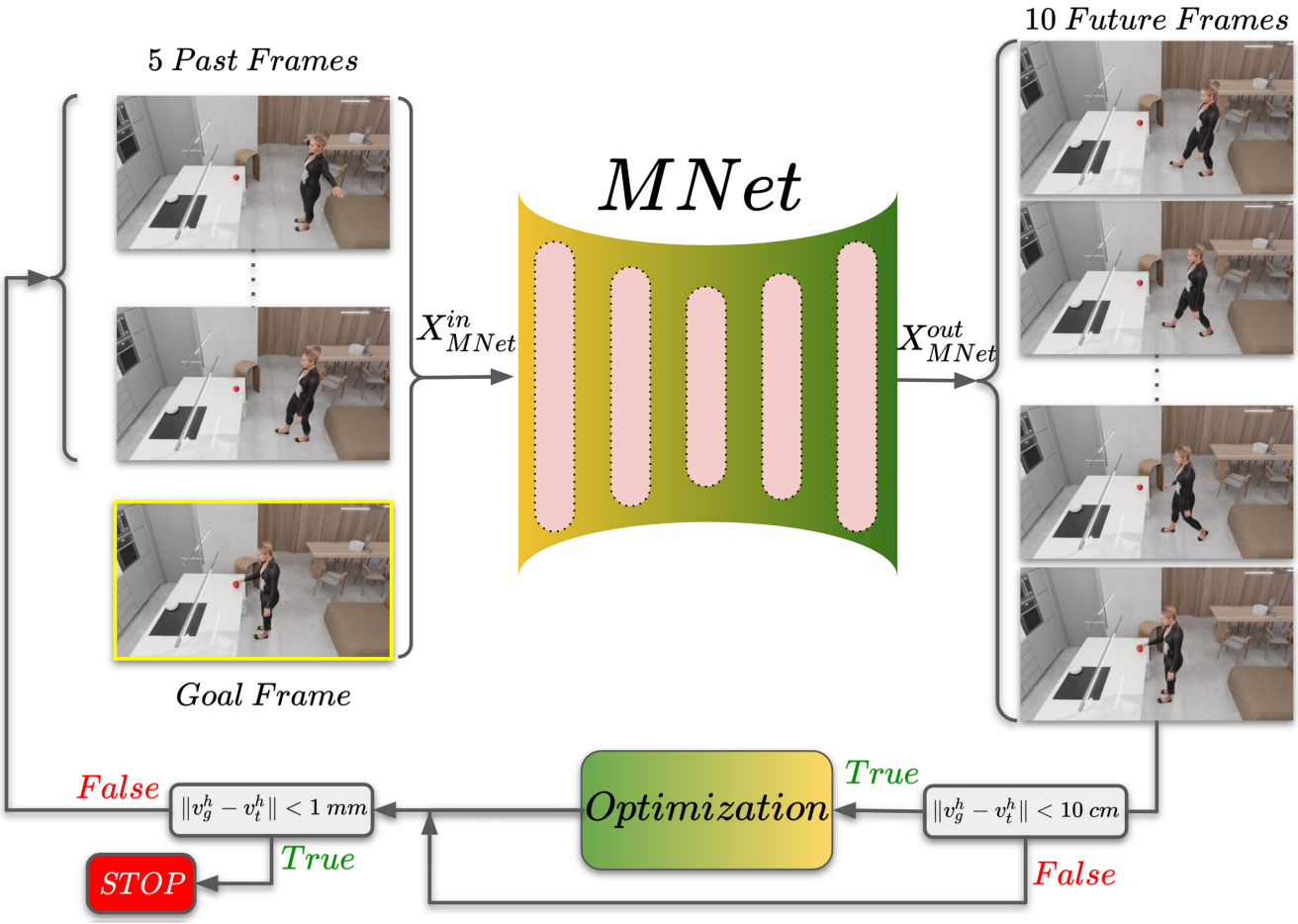

Figure S.2. Architectural overview of the MNet network, as well as the optimization post-processing step (bottom part).

2.2. MNet Architecture

In Fig. S.2 we show the architectural overview of MNet and its optimization-based post processing. MNet is an auto-regressive network that takes in each iteration 5 past frames, $\boldsymbol{X}_p$, and generates the next 10 frames, $\boldsymbol{X}_f$. The optimization process refines the motion to better "reach" the "goal" grasp (generated by GNet). Note that the optimization step is activated only when MNet's estimated hand vertices get closer than 10 cm to the "goal" hand vertices.

3. Qualitative Results

In Fig. S.3 we show more qualitative results of GOAL from different views and with close-ups on hands.

4. Failure Cases

Despite generating mostly realistic motions, the MNet optimization sometimes results in small hand-object penetration before the "goal" grasping frame; we show two examples in Fig. S.4. This is due to linearly interpolating the motion between the "current" and "goal" frames during optimization, and could be solved in future work by adding a penetration loss, and potentially by replacing the linear interpolation with a more involved approach.

In addition, in some cases we observe "foot sliding", especially when the "starting" body is placed further than $1.5m$ from the object. Figure S.5 shows some "foot-sliding" cases in comparison to the ground-truth motion. While our main focus here is to generate grasping motion, future work should look into combining GOAL with longer walking-motion generation methods [1, 4].

5. Social Impact

While realistic motion generation has mostly positive use cases in AR/VR, games, and movies, with the recent advances in neural rendering and deepfakes, we see a possibility that our results could be used for full-body deepfakes. Being aware of this, we will make our models available with an appropriate license.

Figure S.3. More qualitative results generated by GOAL and showed from different views with close-ups on the hand grasps.

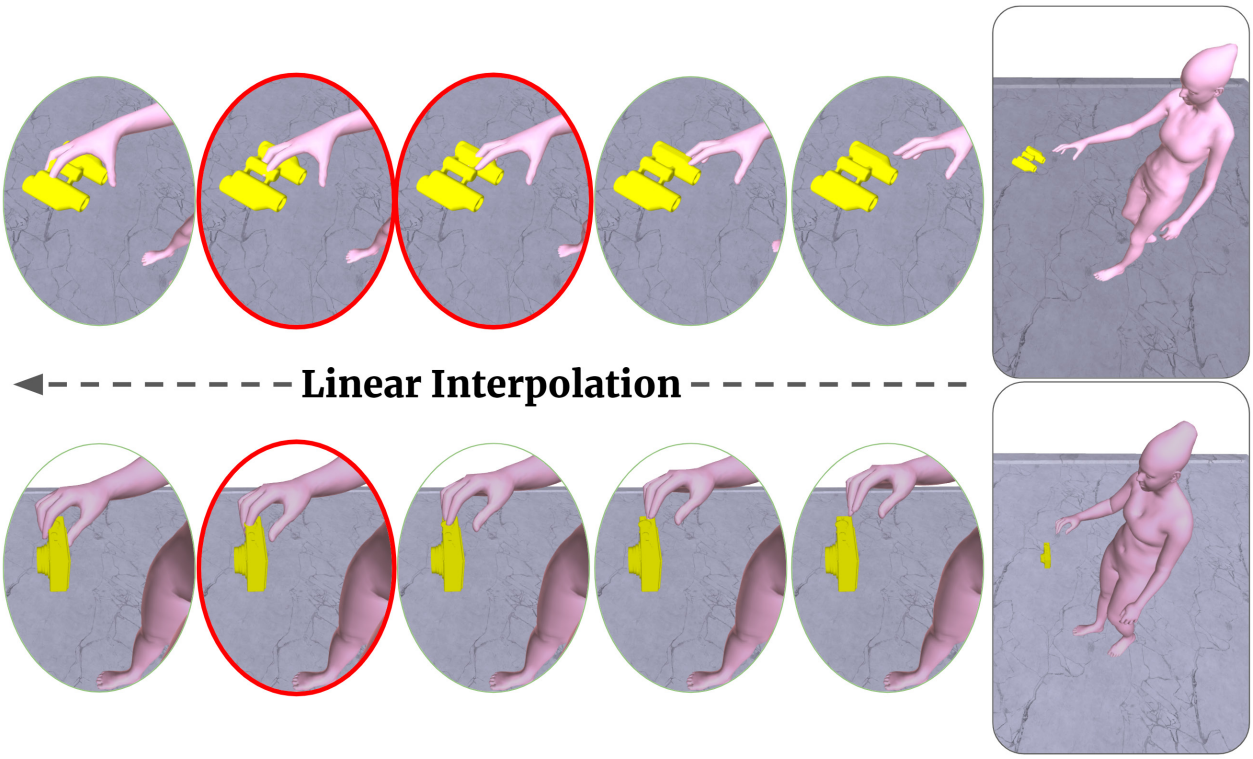

Figure S.4. Two penetration failure cases during MNet's optimization post-processing with linear interpolation. In the figure the hand approaches the object from right to left, and the red ovals highlight hand-object penetrations.

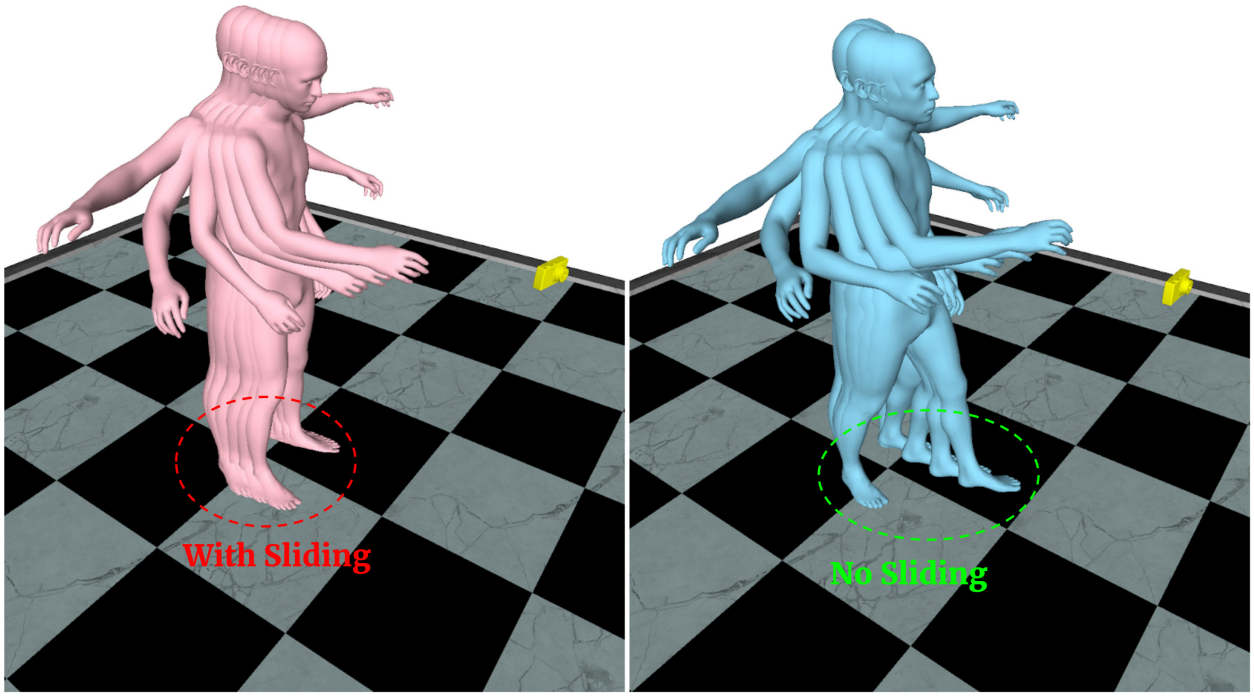

Figure S.5. A failure case of "foot sliding" generated by MNet (left), and compared to the corresponding ground-truth motion (right). Note that for the ground truth (right) the right foot maintains contact with the floor, while the left foot moves in the air for walking.

References

[1] 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.

[2] 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.

[3] 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)*, volume 12349, pages 581–600, 2020.

[4] Yan Zhang, Michael J. Black, and Siyu Tang. We are more than our joints: Predicting how 3D bodies move. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3372–3382, 2021.