Supplement for Physically Plausible Animation of Human Upper Body from a Single Image

Ziyuan Huang
National Taiwan University
r09922031@ntu.edu.tw

Zhengping Zhou
Stanford University
zpzhou.cs@gmail.com

Yung-Yu Chuang
National Taiwan University
cyy@csie.ntu.edu.tw

Jiajun Wu
Stanford University
jiajunwu@cs.stanford.edu

Karen Liu
Stanford University
karenliu@cs.stanford.edu

1. Detail Settings

In this section, we show the detail settings of our experiments, including the setting of our human character and training hyperparameters.

Body Structure Figure. 1 illustrates the upper body of the human character used in our experiments. There are eight active joints (shown in red) and each joint has different numbers of DoFs and rotation axes (shown in the text box next to each joint). Rotations of all joints are not limited. This setting allows our system to generalize among all human subjects. Though this compromises some physical plausibility when it comes to people with different shapes, it greatly increases the feasibility and versatility of our application.

Experimental Setup We run all of our experiments on a machine with one TITAN RTX GPU using CUDA version 10.0. Our system consists of three major components: (1) *dynamic model*. We split 80% of our dataset as training data, 10% for validation, and 10% for testing. For hyperparameters, we use Adam as our optimizer with the initial learning rate 0.001. The learning rate will be decayed by 0.1 after 10 epochs without improvement in the validation loss; (2) *policy model*. The weights of the reward terms are $w_{\text{task}} = 0.1, w_{\text{upright}} = w_{\text{ctrl}} = 0.01, w_{\text{alive}} = 1$ for the tracking task, as mentioned Section 3.2. For the reaching model, we set $w_{task} = 1$ to encourage our model to better reach the target point. To stable the training process, we set the random seed of the environment to 666; (3) *pose-to-image generator*. Most settings of our model are the same as [1]. The rendered image size is 256×256, and the learning rates of the generator and discriminator are both set as 0.0001.

Testing Algorithm After obtaining the dynamic model, the policy model and the photorealistic generator, we can infer the desired video through Algorithm 1. For each step t, the policy model derives the optimal $a^{(t)}$ according to the current state $s^{(t)}$ and the goal g. Afterward, our dynamic model can generate the new state $s^{(t+1)}$ via the output of the policy model. Note that a state $\boldsymbol{s}^t = (\boldsymbol{x}^t, \triangle \boldsymbol{x}^t)$ is a concatenation of a 2D pose and its offset from the 2D pose at the previous time step. With the new pose $\boldsymbol{x}^{t+1}$, the photorealistic renderer ψ can generate $(t+1)th$ image I^{t+1} in the output video.

Full Body Experiments In Figure 13 of the paper, we show an extension of our method for the full-body tracking task. The model can generate results for human subjects tracking wrist and ankle target points specified by the user. The dynamic model is trained on the full Human 3.6M dataset [3] with all joints free except the root pelvis. Since the increase in DOF will make the policy model more likely to generate unreal poses, we add a reward term r_{imitate} to encourage the model to minimize the difference between the generated pose and the ground-truth pose:

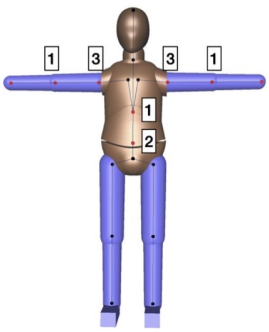


Figure 1: The body structure of our model.

$$r^t_{\text{imitate}} = -w_{\text{imitate}} \left(\| \boldsymbol{x}^{t+1} - \boldsymbol{g}^{t+1} \| \right), \tag{1}$$

where $\boldsymbol{x}^t$ and $\boldsymbol{g}^t$ are the character pose and the ground-truth pose at time t respectively. The weights of the reward terms are $w_{\text{task}} = 0.05, r_{\text{imitate}} = 0.05, w_{\text{upright}} = w_{\text{ctrl}} = 0.01, w_{\text{alive}} = 0.5$. Other settings of the policy are identical to those of the upper-body tracking model described in Section 4.1.

Methods	realism	precision
Ground truth	24%	24%
Hao et al. [2]	86%	72%
3D simulation	58%	57%

(a) User assessment on tracking task

Methods	realism	precision
Ground truth	34%	53%
Hao et al. [2]	91%	88%
3D simulation	62%	65%

(b) User assessment on reaching task

Table 1: The results of the user study. The matrices include the percentage (%) of users that prefer our method over the left method. The following columns are the user assessment of video realism, and the tracking precision, in terms of how accurately the wrists track the provided targets. (a) User assessment of tracking videos. (b) User assessment of reaching videos.

Algorithm 1 Controlling 2D poses using trained π_θ

▷ **Input:** Initial 2D keypoint coordinates $\boldsymbol{x}^0$, dynamics model ϕ, policy network π, photorealistic renderer ψ, user-specified goal $\boldsymbol{g}$
▷ **Output:** image sequence $\{I^{1:T}\}$

$\boldsymbol{s}^0 \leftarrow [\boldsymbol{x}^0, \boldsymbol{0}]$
for $t \leftarrow 0 : T-1$ **do**
  $\boldsymbol{a}^t \sim \pi(\cdot|\boldsymbol{s}^t, \boldsymbol{g})$
  if user adds perturbation $\tilde{\boldsymbol{a}}$ **then**
    $\boldsymbol{a}^t += \tilde{\boldsymbol{a}}$
  end if
  $\boldsymbol{s}^{t+1} \leftarrow \phi(\boldsymbol{s}^0, \boldsymbol{a}^{0:t})$
  $I^{t+1} = \psi(\boldsymbol{x}^{t+1})$
end for

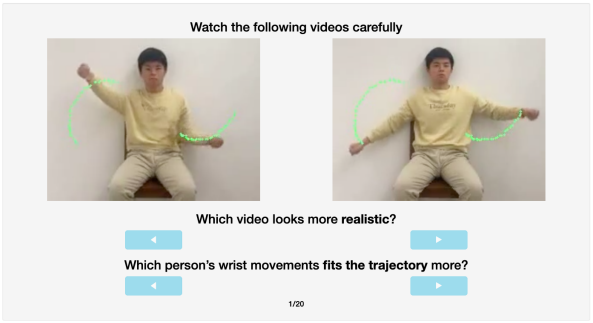
2. More Results

2.1. Human Preference Study

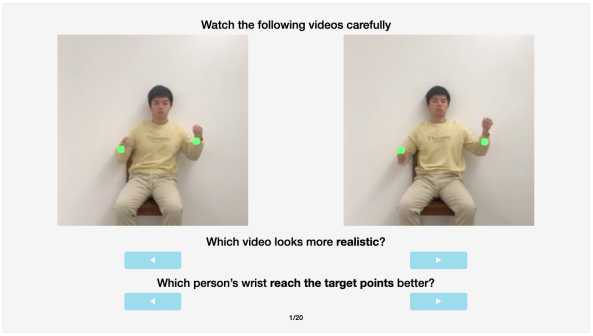
Procedure and interface The user study includes three types of tasks: tracking, reaching, and recovery from perturbation. The instructions of the user study are described as follow:

Task

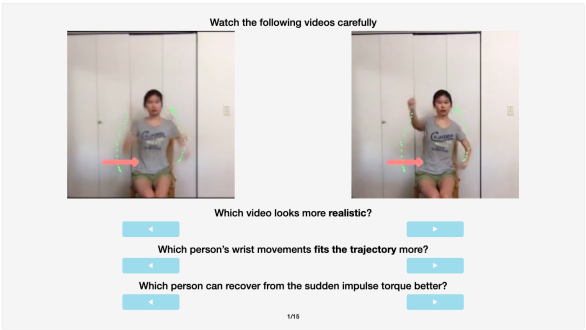
- We are analyzing various computer (automatically) generated human videos with *physic simulated body and controllable wrists*. And we are trying to understand which one is better.
- We will show you two videos side by side for comparisons of different methods. Please choose which one looks more *realistic* and whose wrists *fit the target trajectories* better. The target trajectories will be shown as green dots.
- (Optional, only shown for the perturbation recovery task) We will show you two videos of people tracking target trajectories *with a force to push him from the*



(a) tracking



(b) reaching



(c) perturbation recovery

Figure 2: Screenshots for the user study. The buttons will become opaque once the participants choose them.

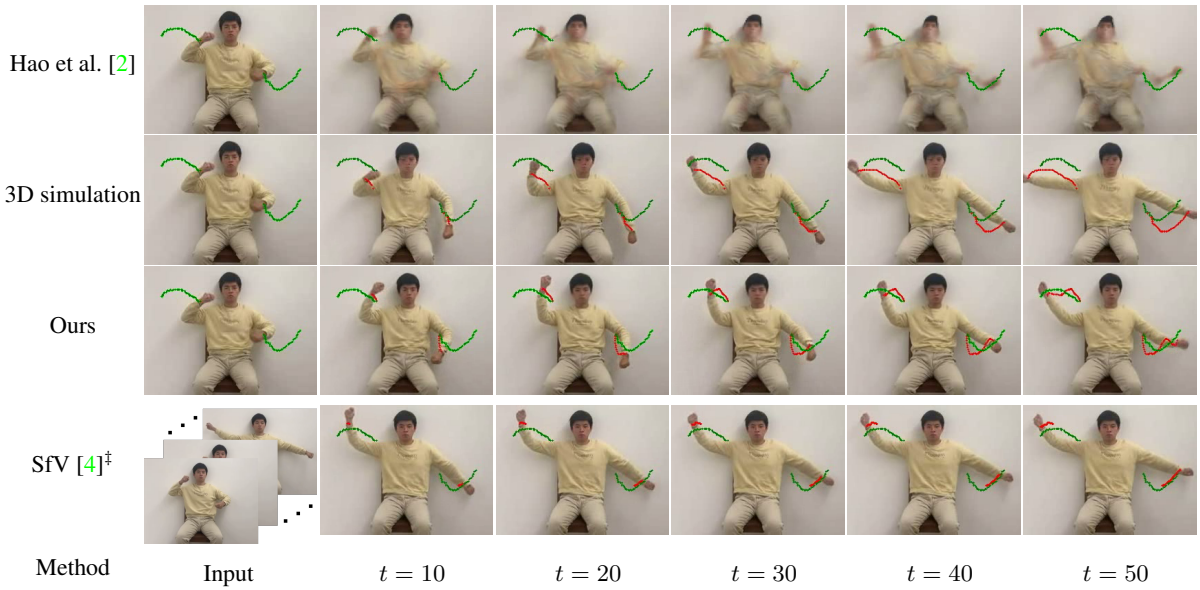


Figure 3: The qualitative comparison with other methods on the tracking task. Our method matches the trajectories better. For each method, we show the input image and the two input wrist trajectories (green) on the leftmost; and the rest columns show the trajectories of the synthesized wrist positions (red). Note that [2] is not a pose-based approach, so the wrist position is not trackable. ‡SfV [4] requires the whole ground truth video as the input for training. See Section 4.2 of the paper for more details.

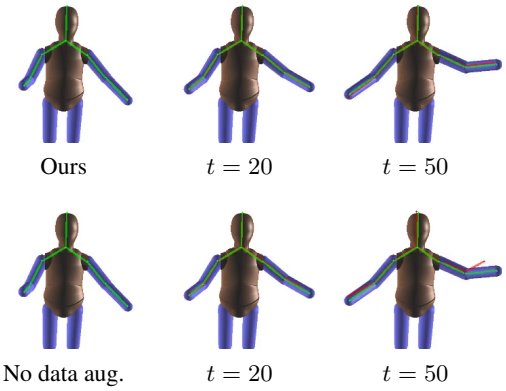


Figure 4: Comparison of our dynamic model against the model without data augmentation. Green keypoints represent the ground truth; red keypoints are the predicted results. Without augmentation, body movement forced by 3D torques produces an error that increases with time.

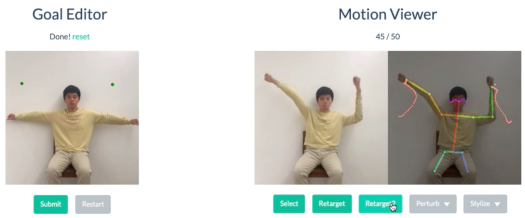


Figure 5: Screenshots for the user interface. User can specify the reaching point by clicking arbitrary positions on the left picture (green points). After submitting the task, our application will automatically generate a video of the human reaching the target points.

pelvis. Please choose which video can recover from the sudden impulse torque better.

- The chosen videos should look more *realistic* and *physically reasonable*!

Rules

- Every page will have two videos at the top and two questions at the bottom like the ones shown above.
- For each question, you have to choose one of the two options at the bottom as your answer.
- There will be 15 to 20 (depending on different tasks) pages, and every page will roughly take you 10-30 seconds to analyze and answer.
- Once you select your preferred video, click the blue Next button on the bottom right of the page.
- The blue Next button will not be activated for the first few seconds. Please take the time to check out the video.
- You will be paid ($0.5 to $3.5, depending on different tasks) in total compensation.
- Please observe the videos carefully as your responses are critical to us! Thank you!

After the participants read the instructions, they are asked to answer 15 to 20 questions. Figure 2 shows the screenshots

of the questions for three different tasks. We compare our method with the baseline methods ([2] and 3D simulation) on the three tasks. For tracking and reaching, since it is possible to have ground truth, we also compare the ground truth with our method. Each questionnaire takes roughly 5 to 10 minutes to finish, and the participants will approximately be paid $10 per hour. In total, we spent $120 on participant compensation for the user study.

2.2. Detailed analysis of different tasks

In Section 4.2.1 of the paper, we present the result of the user study in Table 1. It is the combined result for tracking and reaching tasks. This section shows the results of the user study for different tasks. Table 1(a) gives the preference matrices for the tracking task, while Table 1(b) reports for the reaching task. By comparing Table 1(a) and Table 1(a), it is clear that our method can synthesize more realistic videos for the reaching tasks. The preference rate of our method against the ground truth raises from 24% for tracking to 33% for reaching. Similarly, in terms of precision, our method performs better on the reaching task than tracking. Table 1(b) shows that the supporters for our results and the ground truth are even equally matched for reaching.

2.3. Other Qualitative Comparison

Figure 6 of the paper shows the qualitative comparison of the perturbation task. Figure 3 shows the qualitative comparison on a tracking task without perturbation. As shown in the figure, although the 3D simulation method can trigger the wrists to move in the same direction as the target trajectories, 2D-to-3D projection generates inevitable errors that make the wrists shift. As for SfV [4], the synthesized motion cannot track as precisely as our method does.

2.4. Ablation on Data Augmentation

Section 3.1 of the paper describes data augmentation techniques for increasing training data. Figure 4 provides a qualitative comparison of our dynamic model and the model without data augmentation. With data augmentation, our dynamic model performs more accurately.

2.5. Application

To demonstrate the usage of our model, we design an interface for the users to generate the controllable video on their own. Figure 5 shows the screenshot of our application. The users can edit the reaching goal point on the left part of the interface, and the result will be displayed on the right part. Also, the users can choose to add a perturbation force or retarget the result to another human subject by clicking the buttons under the result. In the future, we will design a more flexible interface in which users can generate target tracking points by dragging their mouses and can push any amount of external force at arbitrary positions of external force at arbitrary positions.

References

[1] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942, 2019.

[2] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, 2013.

[4] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfv: Reinforcement learning of physical skills from videos. *ACM Transactions On Graphics*, 37(6):1–14, 2018.