Synthetic Human Action Video Data Generation with Pose Transfer

Supplementary Material

1. Prolific Participant Instructions

As described in Section 4.1, to create the *RANDOM People* dataset, we crowd-sourced novel human identity videos using the Prolific data platform. Before entering the recording interface and seeing any instructions, the participants were informed about the intended use of the dataset, and asked whether they consent to their video—as well as a 3D Gaussian model and other derivate artifacts—being publicly available for research purposes.

Once the users agreed, they advanced to the recording interface, where they were presented with a video showing the action to perform and the following instructions:

Participant Instructions

Watch [this video] on YouTube. You will use your phone or tablet to record yourself performing the same sequence of actions.

First, prepare the recording. Place your phone or tablet approximately 7-8 feet (2-2.5 meters) away on an elevated surface. The phone should be positioned at a height above your waist level. Ensure that you are fully visible and approximately in the center of the frame.
Next, proceed with recording yourself while performing the following sequence of actions:

1. a slow 360 rotation with your hands down;
2. a slow 360 rotation with your hands up in a double L shape as shown below;
3. a slow 360 rotation with your hands down.

Importantly, the recording must meet the following **criteria**:

- Your whole body, head to feet, is visible in the video at all times.
- You must be well-lit.
- Besides you, no other people, animals, or moving objects appear in the video. This includes statues, posters, and TV.
- Your camera is positioned on an elevated surface, such as a table or wardrobe—do not record with a phone placed on the ground.

When you're ready, upload the video below. By uploading, you agree to [these terms].
Thank you!

2. Selected Action Classes

As described in Section 4, we manually selected a subset of 16 action classes within the Toyota Smarthome [7] and NTU RGB-D [43] based on the following criteria: (1) Minimal Use of External Objects, (2) Consistent Camera Angles, and (3) Distinctive Actions. In particular, these subsets include:

Selected Action Classes: Toyota Smarthome

1. Cook.cut
2. Cook.stir
3. Cook.Usestove
4. Drink.Frombottle
5. Drink.Fromcan
6. Drink.Fromcup
7. Eat.snack
8. Getup
9. Laydown
10. Pour.Fromkettle
11. Pour.Frombottle
12. Sitdown
13. Walk
14. Usetelephone
15. Maketea.Insertteabag
16. Enter

Selected Action Classes: NTU RGB-D

1. drink water (A1)
2. eat meal (A2)
3. brush teeth (A3)
4. pick up (A6)
5. throw (A7)
6. sit down (A8)
7. stand up (A9)
8. clapping (A10)
9. hand waving (A23)
10. kicking something (A24)
11. jump up (A27)
12. point to something (A31)
13. nod head/bow (A35)
14. salute (A38)
15. put palms together (A39)
16. cross hands in front (A40)

3. Compute Considerations

This appendix section discusses compute considerations surrounding our experimental setup. Our aim is to provide an intuition for the computational demands of this process and to explain the parameters we chose, which were largely constrained by our computing capacity. Due to limited GPU access, we were only able to perform the experiments on the *RANDOM People 15* subset with 15 novel human identities instead of the complete set of 100 novel human identities.

These identity videos I were standardized to 18 seconds at 18 FPS; the reference videos T were normalized to 20 seconds at 25 FPS. While the statistics reported below, which informed this parameter choice, have been measured precisely, this is not meant to constitute a formal analysis of the running time and optimization; rather, we aim to equip the reader with an understanding of the approximate computing complexity and the rationale behind our parameter decisions.

Most identity videos in I collected for *RANDOM People* were between 40 to 60 seconds in length, containing approximately $1,200$ frames. Creating an avatar (as described in Section 3) from a single identity video in I on an NVIDIA RTX 4090 GPU took approximately six hours. By normalizing the videos to 18 seconds at 18 FPS, the avatar creation time was reduced by a factor of four, down to approximately 1.5 hours.

We explored additional configurations as well. When normalizing to 20 seconds at 25 FPS, the processing time was approximately 2.5 hours. At 20 seconds and 20 FPS, the processing time was around 2.3 hours, and at 20 seconds and 18 FPS, it was roughly 1.8 hours.

However, when reducing the frame count further, we observed a decline in the quality of the final avatar. Ultimately, we found that the optimal balance between model accuracy and processing time was achieved with approximately 320 training frames per identity.

4. Qualitative Evaluation

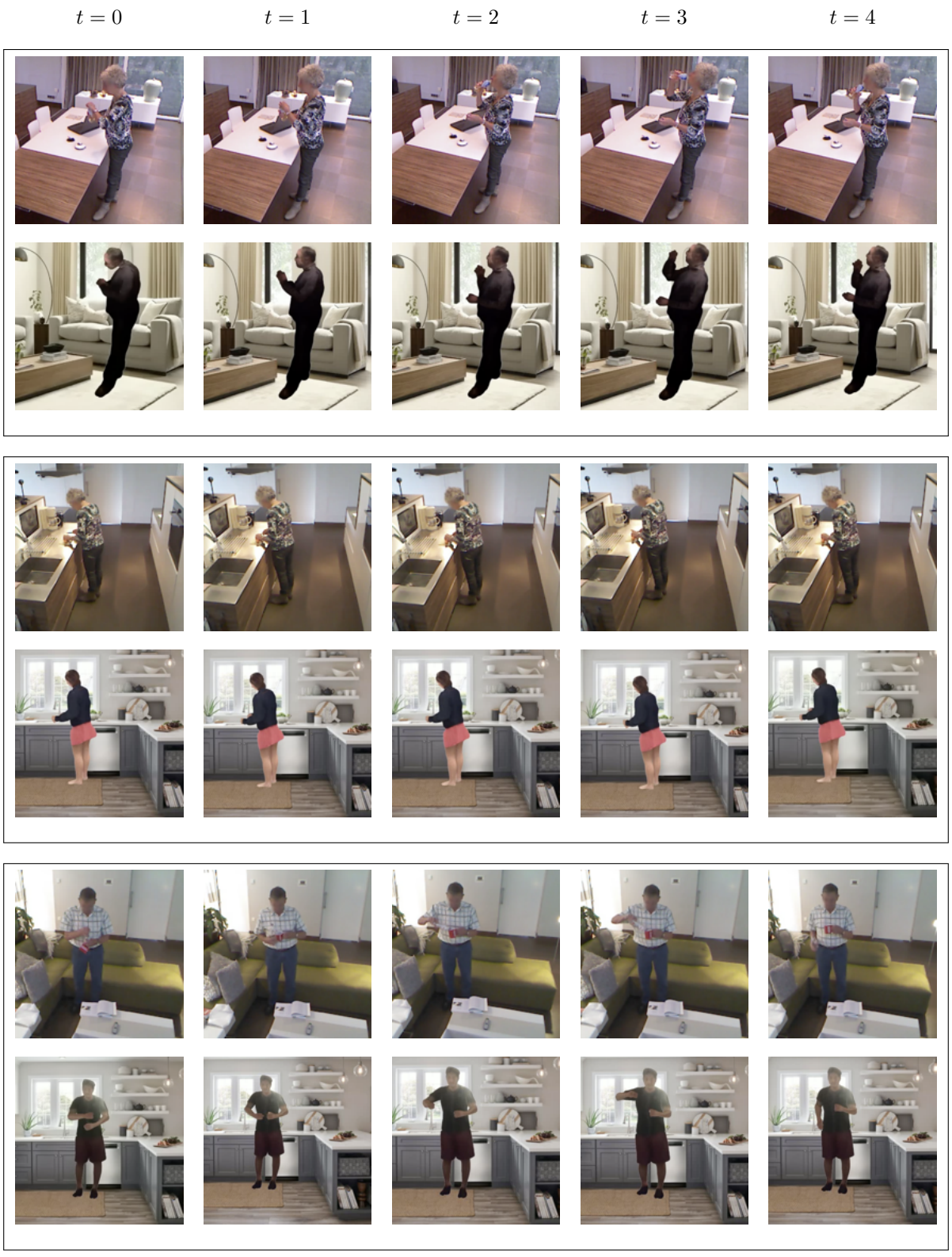

Figure 8. Examples of video frames at $t = \{0, 1, 2, 3, 4\}$ seconds from the source video (top), taken from Toyota Smarthome, and the target video (bottom), generated by our synthetic data generation method, where the pose alignment is consistent.

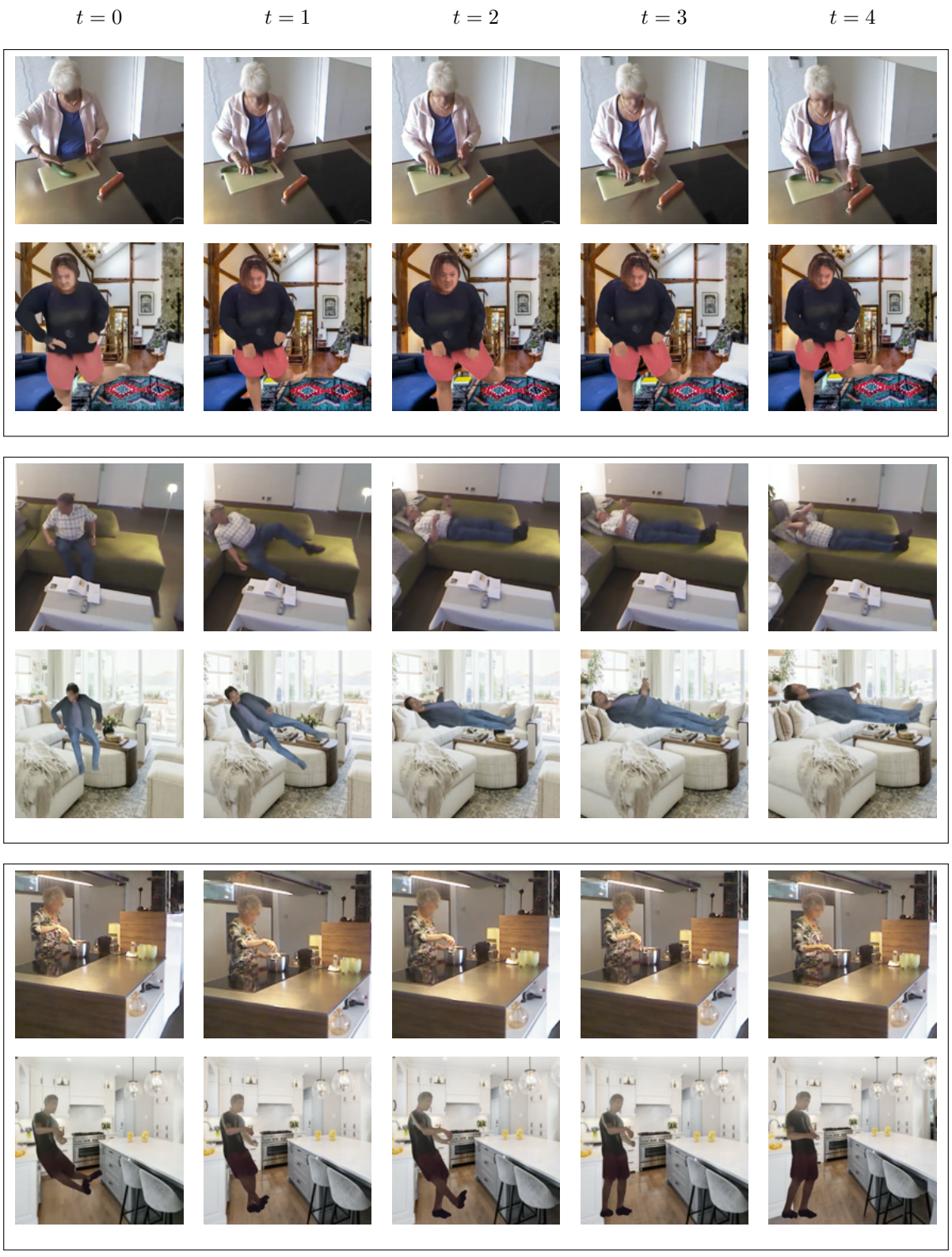

Figure 9. Examples of video frames at $t = \{0, 1, 2, 3, 4\}$ seconds from the source video (top), taken from Toyota Smarthome, and the target video (bottom), generated by our synthetic data generation method, where the pose alignment is inconsistent.

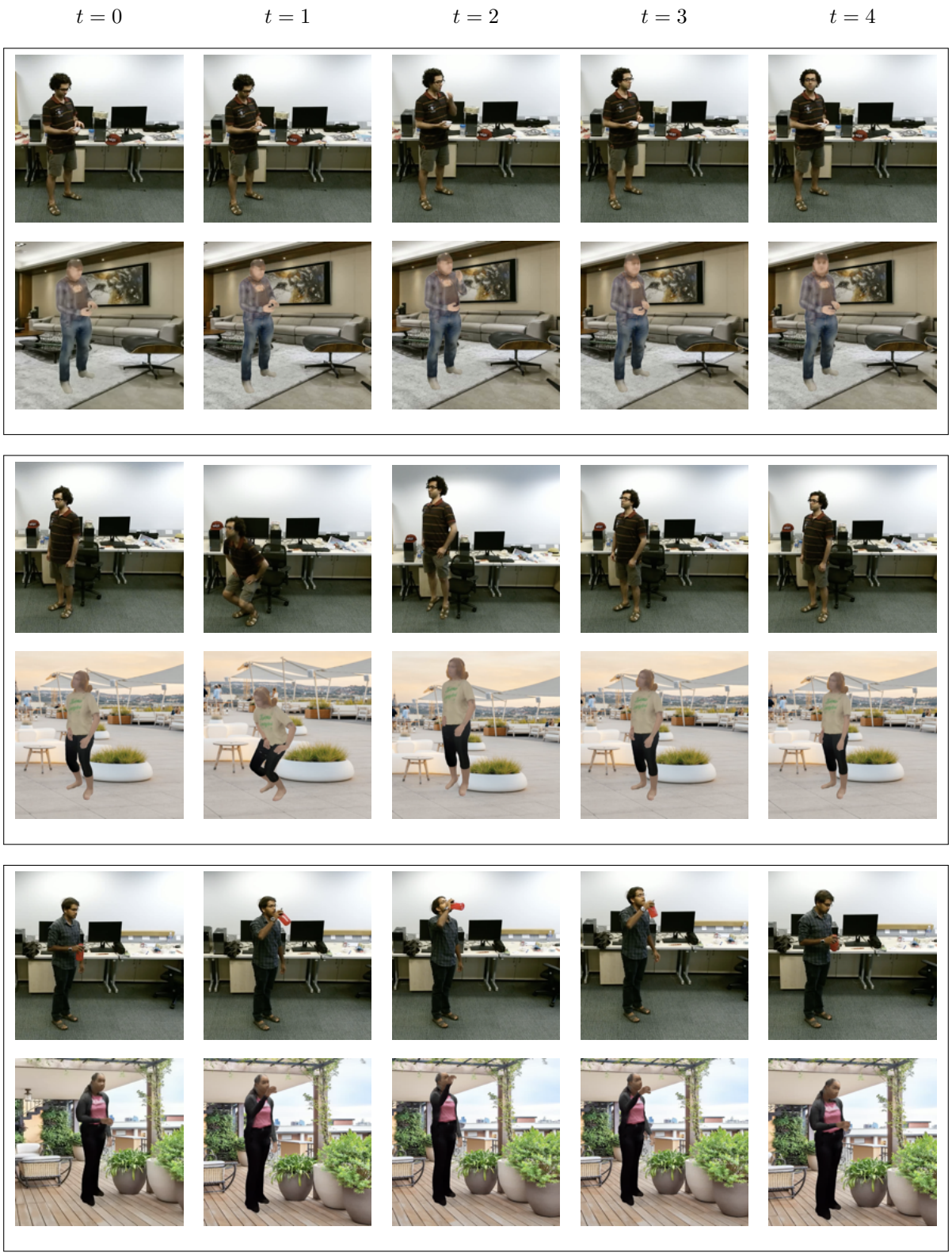

Figure 10. Examples of video frames at $t = \{0, 1, 2, 3, 4\}$ seconds from the source video (top), taken from NTU RGB+D dataset, and the target video (bottom), generated by our synthetic data generation method, where the pose alignment is consistent.

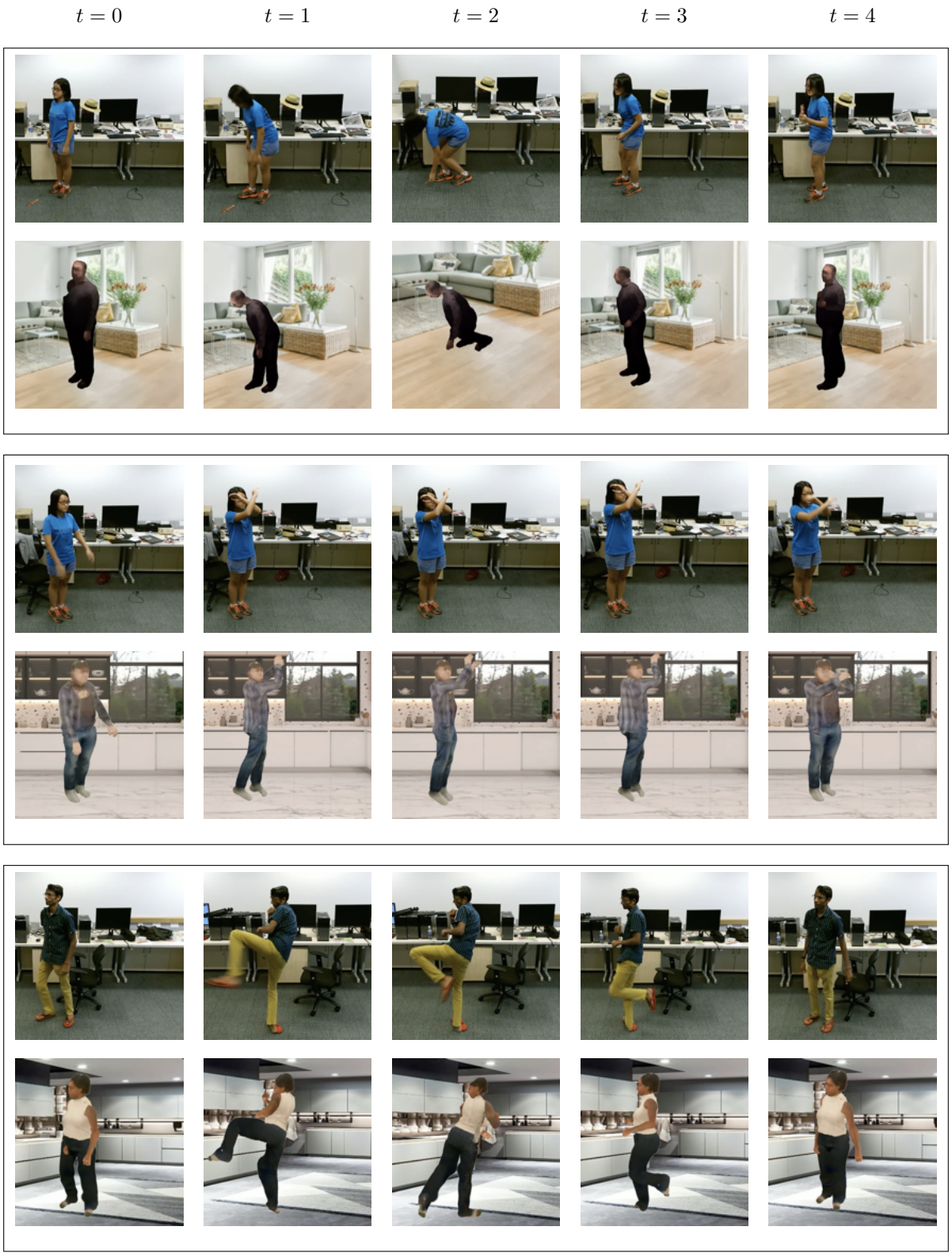

Figure 11. Examples of video frames at $t = \{0, 1, 2, 3, 4\}$ seconds from the source video (top), taken from NTU RGB+D dataset, and the target video (bottom), generated by our synthetic data generation method, where the pose alignment is inconsistent.

Figure 12. Example video frames illustrating limitation L3 (see Section 7).

Cook.stir Cook.Usestove Laydown Cook.cut Maketea.Insertteabag

Figure 13. Example video frames illustrating limitation L4 (see Section 7). The shown action classes are from Toyota Smarthome.