Synthetic Human Action Video Data Generation with Pose Transfer

Vaclav Knapp
SSPS
Knapp.Va.2022@skola.ssps.cz

Matyas Bohacek
Stanford University
maty@stanford.edu

Figure 1. Examples of synthetic video frames showing 12 distinct human actions, generated using our method as extensions to the Toyota Smarthome (above) and the NTU RGB+D (below) datasets. The presented identities are from our *RANDOM People* dataset.

Abstract

In video understanding tasks, particularly those involving human motion, synthetic data generation often suffers from uncanny features, diminishing its effectiveness for training. Tasks such as sign language translation, gesture recognition, and human motion understanding in autonomous driving have thus been unable to exploit the full potential of synthetic data. This paper proposes a method for generating synthetic human action video data using pose transfer (specifically, controllable 3D Gaussian avatar models). We evaluate this method on the Toyota Smarthome and NTU RGB+D datasets and show that it improves performance in action recognition tasks. Moreover, we demonstrate that the method can effectively scale few-shot datasets, making up for groups underrepresented in the real training data and adding diverse backgrounds. We open-source the method along with RANDOM People, a dataset with videos and avatars of novel human identities for pose transfer crowd-sourced from the internet.

1. Introduction

As large-scale training datasets proved essential for generalization in AI models [49]—be it for language, image, or video tasks—many internet-scraped datasets emerged. In domains where data at such a scale is unavailable, data augmentation and synthetic data stepped in to fill the gap [46]. The recent advent of generative AI models has only accelerated this trend [50]. However, in video understanding tasks, data augmentation offers only limited gains, and so far, approaches to whole-cloth synthetic data generation have fallen short of quality and overall usability [25]. As a result, synthetic data is not widely used to train video classification or understanding models [20].

Much of the recent work in synthetic video data generation has focused on videos with human actions [54]. This restricts the problem scope while still offering many use cases to downstream applications involving human action understanding and generation (e.g., sign language translation, gesture recognition, and human motion understanding in autonomous driving). In particular, two lines of work

have prevailed in the literature over recent years: (1) approaches based on classic computer graphics, often using simulated rendering, and (2) generative AI models [27]. While the first approach preserves semantic attributes of the action (in that, the pose of protagonist in the video is fully controlled, and so the desired action is, too), it has not yet passed the uncanny valley [21]. The second approach, on the other hand, can generate photorealistic human videos but is unstable in that it lacks full pose control and still contains physical errors as well as artifacts [51]. Therefore, neither of these approaches has been widely adopted as a source of synthetic data for training video classification or understanding models.

Recent work has, however, focused on bridging these two approaches by leveraging the versatility and ability to generate photorealistic content that comes with generative AI while grounding it in a physical model of the human body that gives one full control over the resulting action [26, 34]. While these models still have some issues (namely, high demands for reference identity videos and computational demands, as further discussed in Section 2.2), we found them to be effective enough to open the inquiry into using such models for synthetic video generation of human action for model training.

In particular, we devise a method that, given reference videos, reenacts the human actions shown in these reference videos using novel human identities in novel settings. To do so, we employ a modified ExAvatar [34] 3D Gaussian framework as our avatar animation backbone. We evaluate this method on action recognition using the Toyota Smarthome [8] and NTU RGB+D [43] datasets, improving the performance of two baseline models while proving particularly efficient for few-shot learning.

The primary contributions of this paper include:

1. proposing a method for synthetic human action video data generation using pose transfer;
2. open-sourcing this method with ExAvatar [34] as the avatar animation backbone;
3. collecting and open-sourcing the *RANDOM People* dataset with (a) novel human identity videos, (b) their avatars for pose reenactment, and (c) background images.

The dataset, code, and additional resources are available at `synthetic-human-action.github.io`.

2. Related Work

We open this section by reviewing the existing scholarship on synthetic data generation for action recognition model training. We continue with an overview of pose transfer methods for generating novel images and videos of people reenacting reference poses. We also enumerate the most prominent datasets for video action recognition. We close with a broader discussion of video classification models.

2.1. Synthetic data for Action Recognition

Early work on synthetic video data generation for action recognition model training exploited fully simulated environments powered by classic CGI methods. One of these methods is ElderSim [19], specifically designed to generate videos of elderly individuals performing day-to-day tasks. Due to the uncanny features of the simulation, the resulting videos lacked photorealism and detail. Moreover, as these are fully simulated actions, they lack much of the imperfections of real-world human motion.

Another approach built around CGI simulation was employed in the creation of the Robot Control Gestures (RoCoG-v2) dataset [39]. This dataset comprises both real and synthetic videos of seven gesture classes for human-robot teaming. Similar to ElderSim, the method behind this dataset lacks options for adding external identities or matching real-world human actions, as it is powered by a CGI simulator.

Over time, methods based on classic CGI technologies have been extended and combined with AI techniques. This was the case with SURREACT [48], a model for generating synthetic videos from unseen viewpoints. While this approach has been shown to improve the robustness of action recognition models to diverse viewpoints, the method cannot generate completely new renditions of the actions. The target identities match those from the videos, but additional external identities are not supported.

More recently, BEDLAM [3] showed that a purely synthetic dataset built using SMPL-X [36] can be used to achieve state-of-the-art performance on human pose estimation. SynthAct [42] further showed how Unity can be leveraged to create synthetic human action video data suitable for robotics applications.

While these recent advancements highlight the potential of synthetic data for certain computer vision tasks, most existing methods that generate synthetic data for action recognition still struggle with photorealism and generalization to new target identities. Although they have been shown to enhance human action video classifiers in some scenarios (e.g., handling novel viewpoints), achieving this often requires substantial adaptation to the task at hand. As described Section 3, our method overcomes many of these shortcomings while utilizing the advantages of explicit human models, which ensure full control of the generated human motion. Furthermore, it is versatile and can be applied to many contexts without modifications.

2.2. Pose Transfer

The task of pose transfer entails generating images or videos of a given human identity (target) in new poses (source). Formulations of this task differ in the input (target identity and source pose) and output modality.

Earlier approaches focused only on generating static im-

ages. The first successful works in this domain broke ground around 2017 when a two-stage neural network for generating images of people conditioned on pose was introduced [31]. From then on, the task has been approached by emerging neural network architectures, as seen in other areas of computer vision: most prominently, GANs [28, 32, 44] and attention-equipped methods [40, 55].

Recent approaches focused on generating videos. These include Magic Animate [52], Animate Anyone [18], Champ [56], and ExAvatar [34], which we review in more detail.

Magic Animate extracts a DensePose [16] representation of each frame in the source video. The target identity is provided as a single image. To generate novel videos of the target identity in the source poses, Magic Animate employs a video diffusion model with an appearance encoder.

Animate Anyone extracts body keypoints of each frame in the source video. The target identity is provided as a single image. To generate novel videos of the target identity in the source poses, Animate Anyone trains a 3D U-Net model for video denoising and incorporates an additional reference network to extract features from the reference image.

Champ extracts an SMPL model [29] (a parametric 3D pose representation) from each frame in the source video. The target identity is provided as a single image. To generate novel videos of the target identity in the source poses, Champ leverages SMPL as a latent representation within a custom-trained diffusion model.

ExAvatar extracts an expressive whole-body 3D Gaussian representation of each frame in the source video, which is obtained through an ensemble of other body representations, from body keypoints to SMPL-X [37] to expressive head models. The target identity is provided as a video of the person performing a specific set of actions; from this video, the same ensemble of body representations is extracted. Once the avatar and source pose representations are extracted, 3D Gaussian Splatting allows ExAvatar to animate the target identity according to the source poses. In a recent study, human evaluators found ExAvatar to produce human motion that is more consistent with the reference and overall more coherent than its diffusion-based counterparts Magic Animate and Animate Anyone [23].

On the one hand, Magic Animate, Animate Anyone, and Champ are easier to scale because they only analyze a single format of input representations and require only a single image of the target identity, unlike ExAvatar, which analyzes an ensemble of features and requires a specific kind of video of the target. However, constructing 3D Gaussian avatars gives ExAvatar full, guaranteed control of the pose, while Magic Animate, Animate Anyone, and Champ rely on generative approaches like diffusion where the control is not guaranteed. Moreover, these diffusion-based methods often suffer from artifacts, hallucinations, and other deficiencies.

2.3. Video Datasets for Action Recognition

There are multiple popular video datasets for action recognition in the literature. We highlight four most prominent ones, in chronological order of their release.

HMDB51 [24] contains a total of 6,766 videos across 51 classes, collected from YouTube. The actions in this dataset are diverse, including sports, musical instrument playing, and close-up interactions with objects.

UCF101 [45] contains a total of 13,320 videos across 101 classes, from various online sources. Similar to HMDB51, UCF101 also includes a wide range of actions.

NTU RGB+D [43] contains a total of 56,880 videos across 60 classes recorded by the dataset authors with consenting protagonists. The protagonists are shown performing a wide range of regular daily actions (e.g., walking and drinking).

SSV2 [15] contains a total of 220,847 videos across 174 classes. The dataset contains videos of many fine-grained interactions between humans and objects (e.g., pushing, pulling, and sliding).

Toyota Smarthome [8] contains a total of 16,115 videos across 31 classes, with 18 unique participants. This dataset shows mostly senior citizens perform daily activities inside of their homes. This dataset is distinct for its notable intra-class variation and class imbalance.

In our experiments, we chose to use Toyota Smarthome and NTU RGB+D because they include only consenting participants and do not rely on object interactions.

2.4. Video Classification Models

Popular baseline models for video classification include I3D [4], ResNet adapted for video processing [17], SlowFast [13], and, more recently, architectures derived from the Vision Transformer (ViT) [1, 2, 10, 47]. Many ViT-based models are pre-trained on the Kinetics dataset [4] for action recognition, facilitating downstream training on datasets such as Toyota Smarthome and NTU RGB+D. Given this, we opt for older baselines—ResNet and SlowFast—and train them from randomly initialized weights.

3. Methods

In this section, we describe our method for synthetic data generation of human action videos. The method extends existing datasets by using real videos as reference human action that is reenacted by novel identities in new settings.

An overview diagram of the method is shown in Figure 2. At the input, the method receives T, a set of n_T real reference videos with human actions, as well as I, a set of n_I videos with novel human identities for action reenactment. The videos in I show people performing a sequence of actions that was found to be effective for avatar genera-

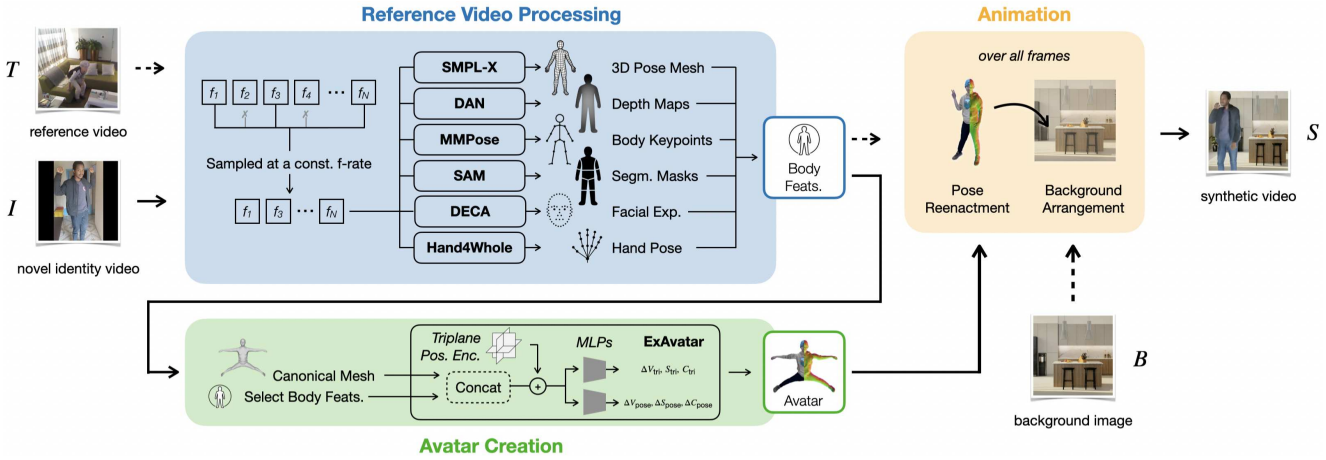


Figure 2. Overview of our method for synthetic human action video data generation. Avatar example taken from [34]. See Section 3.

tion in [34]. Optionally, the method also receives a set of background scene images B. It then generates new videos in which identities in I are animated to reenact the same actions as shown in the real reference videos T, optionally with varying background settings B. The method consists of three main stages: (1) avatar creation, (2) reference video preparation, and (3) animation.

3.1. Avatar Creation

For each novel human identity I_i in $I_1 \dots I_{n_I}$, an expressive whole-body avatar A_i is made using ExAvatar [34], which we use as our avatar animation backbone. This yields fully controllable 3D Gaussian avatars.

To that end, the video I_i is normalized to a constant length and frame rate. Select frames within the normalized video are then used to extract the following features (the frames selection follows ExAvatar's original implementation [34]):

1. 3D pose meshes (SMPL-X [37]);
2. depth maps (DepthAnythingv2 [53]);
3. body keypoints ($133 \cdot 3$, MMPose [6]);
4. identity segmentation masks (Segment Anything [22]);
5. facial expressions (DECA [14]);
6. hand poses (Hand4Whole [33]).

With these features extracted, we train A_i, an animatable 3D Gaussian avatar that preserves the unique identity characteristics of the novel human identity in I_i. In doing so, we follow the implementation of ExAvatar [34] with default parameters unless specified above.

3.2. Reference Video Preparation

Each reference video T_i in $T_0 \dots T_{n_T}$ is normalized to a constant length and frame rate. Select frames within the normalized video are then used to extract the same set of features as used for avatar creation (described above in Section 3.1) with the exception of identity segmentation masks and depth maps, which are not extracted. These features are packaged into F_i.

3.3. Animation

With the set of novel identity avatars A and the set of reference video features F representing training videos T, we proceed with the animation stage.

White-Background Videos. For each training video T_i in $T_1 \dots T_{n_T}$, represented by its features F_i, and each novel identity avatar A_j in $A_1 \dots A_{n_A}$, we generate a synthetic video $S_{(i,j)}$. This results in n_A synthetic videos per training video, totaling $n_T \cdot n_A$ synthesized videos. These intermediate videos will be superimposed onto image backgrounds in the next step.

Image-Background Videos. As a final step, image backgrounds are added to the white-background videos. For each training video T_i in $T_1 \dots T_{n_T}$, represented by its features F_i, and each novel identity avatar A_j in $A_1 \dots A_{n_A}$, we randomly select g image backgrounds from a background pool B, denoted as B_k. For each combination of T_i, A_j, and B_k, we synthesize the video $S_{(i,j,k)}$. This results in $g \cdot n_A$ synthetic videos per training video, for a total of $n_T \cdot g \cdot n_A$ synthesized videos.

4. Data

In this section, we describe the data used in our experiments. We first detail the data collection process of novel human identity videos I. We then describe the selection of reference videos T. Finally, we cover the selection of background scene images B.

4.1. Novel Human Identity Videos

Recordings of 188 participants were crowd-sourced through Prolific. These participants were informed about the in-

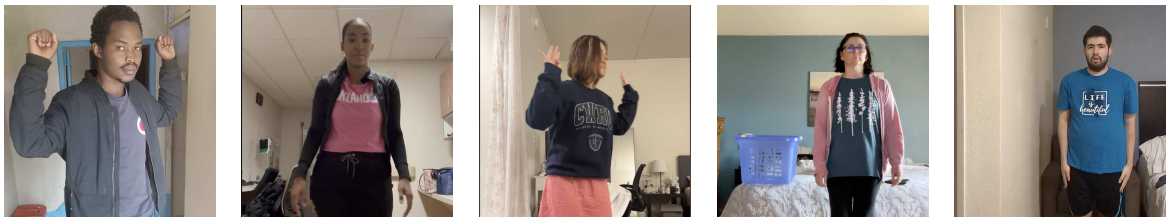


Figure 3. Representative frames from the novel human identity videos I in the *RANDOM People* dataset.

Figure 4. Representative frames from the reference videos T in the *RANDOM People* dataset. These videos, sourced from the Toyota Smarthome dataset, were manually selected following suitability criteria in Section 4.2.

tended use of their recordings and asked to consent to their video—as well as a 3D Gaussian model and other derived artifacts—made publicly available for research purposes prior to entering our interface. The average completion time was 8 minutes, and participants were compensated at a rate of above 8 USD. This data collection was consulted with the IRB office at Stanford University, which concluded that an IRB review is not required.

The participants were instructed to record themselves performing three slow 360 rotations, with a mix of raised and lowered hands. Representative examples of frames from these videos are shown in Figure 3. See Appendix 1 for full recording instructions and acceptance criteria. Based on these criteria, we filtered the 188 collected videos to a set of 100 videos that met our criteria, used as novel human identity videos I.

The participants were chosen from a stratified sample of residents of the United States. Based on self-reported demographics, the dataset (after filtering) contains 41 people who identity as male and 59 people who identify as female. 11 participants are Asian, 10 participants are Black, 11 participants are of mixed race, 66 participants are white, and 2 participants reported a different race. The median age is 38, with a minimum age of 19 and a maximum age of 64. The average response time was 10 minutes.

4.2. Reference Human Action Videos

The following procedure was performed with each evaluated dataset (Toyota Smarthome [7] and NTU RGB+D [43]) individually. To curate a set of reference videos T, we manually selected a subset of 16 human action classes from the dataset (see Appendix 2 for the complete list). The selected classes met criteria outlined in Appendix 2. For each class, we then manually selected 5 videos, yielding a total of $n_T = 16 \cdot 5 = 80$ videos. Representative examples of frames from these videos are shown in Figure 4.

4.3. Background Images

The pool of background images B was created to match typical environments for actions present in each of the evaluated datasets. In total, $n_B = 20$ images were scraped from the internet for each evaluated dataset. Representative examples of background images in B are shown in Figure 5.

4.4. Synthetic Human Action Videos

Using $n_I = 15$ novel human identity videos I, $n_T = 80$ reference human action videos T, and $n_B = 40$ background image pool, sampled at $g = 3$ per video, we applied our synthetic data generation method to produce S: a set of $n_T \cdot g \cdot n_A = 3,600$ image-background synthetic videos for each evaluated dataset. Representative frames from these videos are shown in Figure 1.

Upon manual review, we found that most synthetic videos were consistent with the source videos in terms of pose and scene placement, as illustrated in Figure 10 (Appendix 4). However, in some cases, the generated videos deviated from the source poses and scene alignment, as shown in Figure 11 (Appendix 4).

We used the smaller *RANDOM People 15* subset due to computational constraints. See Appendix 3 for further details on compute considerations.

4.5. Open-sourcing

We call our dataset *RANDOM People* and open-source it for research purposes at `synthetic-human-action.github.io`. This data release includes the synthesized

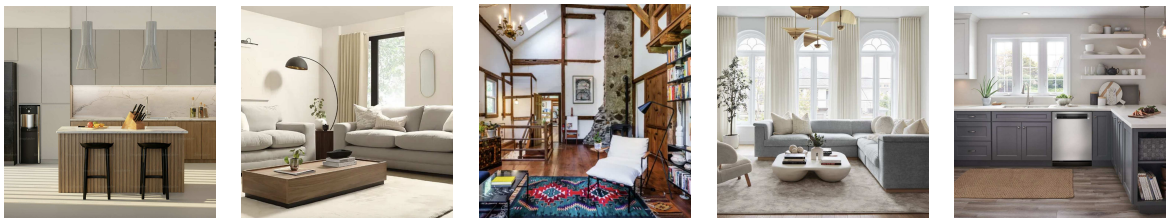


Figure 5. Representative background images from B in the *RANDOM People* dataset.

	Original	Synthetic	Toyota	NTU RGB+D
ResNet	✓		20.46	8.66
	✓	✓	51.15	42.98
SlowFast	✓		38.35	26.75
	✓	✓	55.64	36.29

Table 1. Testing accuracy of ResNet and SlowFast on the Toyota and NTU RGB+D subsets trained in two configurations: with only the original data and both the original and synthetic data.

videos S, the underlying novel human identity videos I along with their avatars A, and the scene background images B. We designate two subsets of the dataset: *RANDOM People 15* with 15 identities, intended for small-scale experiments, and *RANDOM People 100* with 100 identities, intended for large-scale experiments. The dataset is made available under the CC BY-NC 4.0 license[1].

5. Experiments

We conducted three sets of experiments—baseline, one-shot, and few-shot—on both Toyota Smarthome and NTU RGB+D.

5.1. Baseline Experiments

We performed the baseline experiments on two standard video classification architectures: ResNet adapted for video processing [17] and SlowFast [13]. In one case, the model was trained only on the original data from the respective dataset; in the other, it was trained on both the original data and our synthetic data. The training set consisted of $n_{\text{real}} = n_{\text{background}} = 225$ videos per class[2]. The testing set, composed only of original videos, included $n_{test} = 50$ videos per class.

The model was trained for 5 epochs with a learning rate of 1×10^{-4} and a batch size of 4. The input frames were resized to 224×224. The ResNet model had a depth of 50, with the number of frames set to 16. For SlowFast, the number of frames was set to 32. If not otherwise specified, default hyperparameter values from PyTorchVideo [12] were used (a snapshot is included in the open-source release).

5.2. One-shot Experiments

We further evaluated the effectiveness of our synthetic data generation method in a one-shot learning scenario on ResNet. With $n_{\text{real}} = 1$, we trained five models while increasing the sample of synthetic videos from $n_{\text{background}} = 0$ to $n_{\text{background}} = 200$ by 50 videos at a time. The testing set was the same as in the baseline configuration; the one-shot example was selected at random.

5.3. Few-shot Experiments

Finally, we evaluated the effectiveness of our synthetic data generation method in a few-shot learning scenario. The setup is identical as in the previous one-shot experimental setup, except $n_{\text{real}} = 5$. These few-shot samples, too, are selected at random.

6. Results

In this section, we report the results of our experiments, implemented using the PyTorch [35] and PyTorchVideo [12] libraries. The source code is fully open-sourced at `synthetic-human-action.github.io`.

6.1. Baseline Experiments

Table 1 presents the classification accuracy for each model when trained with only original data and with original plus synthetic data.

When trained solely on the original data, ResNet achieved an accuracy of 20% on Toyota and 9% on NTU RGB+D.[3] When synthetic data was added, the accuracy increased to 51% and 43%, respectively.

[1] This license allows use, adaptation, and sharing of the dataset provided the use is non-commercial and appropriate credit is given. See https://creativecommons.org/licenses/by-nc/4.0/deed.en for details.

[2] We capped the number of samples across data sources at 225 (i.e., $n_{\text{real}} = n_{\text{background}} = 225$), despite having more synthetic samples, to match the minimum per-class repetition count in the Toyota Smarthome dataset. This ensured that the ablation compared the contributions of each component rather than reflecting imbalanced data source ratios.

[3] We investigated this performance outside of the baseline methodology to understand why it was below chance (6.25% for this 16-class task). We found that the model struggled to converge, even with a hyperparameter

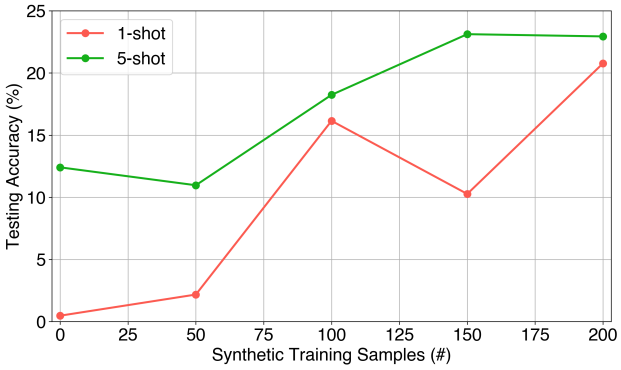


Figure 6. Testing accuracy on real videos from the Toyota Smarthome of two ResNet models trained in a one-shot (green) and few-shot (red) manner with an increasing amount of synthetic samples per class $n_{\text{background}} = 0 \ldots 200$ by steps of 50.

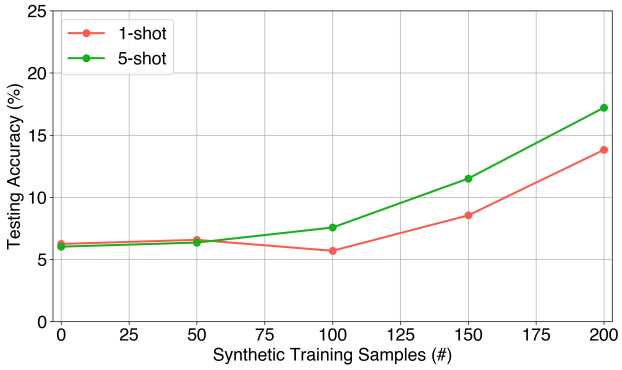


Figure 7. Testing accuracy on real videos from the NTU RGB+D of two ResNet models trained in a one-shot (green) and few-shot (red) manner with an increasing amount of synthetic samples per class $n_{\text{background}} = 0 \ldots 200$ by steps of 50.

Similarly, when trained solely on the original data, SlowFast achieved an accuracy of 38% on Toyota and 27% on NTU RGB+D. With the addition of synthetic data, performance improved to 56% and 36%, respectively.

These results demonstrate that incorporating our synthetic data enhances the performance of video classification models in action recognition.

6.2. One-shot Experiments

Shown in Figures 6 and 7 (red curves) are the testing accuracy advantages gained when training one-shot models with our synthetic data on Toyota Smarthome and NTU RGB+D, respectively, plotted as a function of the number of synthetic samples.

With Toyota Smarthome, shown as the red curve in Figure 6, the model starts below the accuracy of chance (6.25% for this 16-class problem)[4]. Once more than 100 synthetic samples were included, the performance grew to 16%, 11%, and 21%, at $n_{\text{background}} = \{100, 150, 200\}$.

With NTU RGB+D, shown as the red curve in Figure 7, the performance improved from 7%, the chance of luck, to 14%, doubling the original accuracy.

These results suggest that our synthetic data can meaningfully improve the performance of one-shot action recognition models. The trend observed in Figures 6 and 7 suggests that extending the synthetic data sample beyond $n_{\text{background}} = 200$ may hold further improvement.

6.3. Few-shot Experiments

Shown in Figures 6 and 7 (green curves) are the testing accuracy advantages gained when training few-shot models with our synthetic data on Toyota Smarthome and NTU RGB+D, respectively, plotted as a function of the number of synthetic samples.

With Toyota Smarthome, shown as the green curve in Figure 6, the performance improved from 13% at $n_{\text{background}} = 0$ to 23% at $n_{\text{background}} = 200$.

With NTU RGB+D, shown as the green curve in Figure 7, the performance improved from 7%, the chance of luck, at $n_{\text{background}} = 0$ to 23% at $n_{\text{background}} = 200$.

These results indicate that our synthetic data can meaningfully improve the performance of few-shot action recognition models. The trend observed in Figure 6 suggests that, for Toyota Smarthome, the performance gains saturate around $n_{\text{background}} = 150$. For NTU RGB+D, shown in Figure 7, there may be additional gains to be attained beyond $n_{\text{background}} = 200$.

7. Limitations

In this section, we enumerate the primary limitations of our method for synthetic video data generation. These limitations can be divided into two categories: those stemming from the employed pose transfer framework (in our case, ExAvatar) and those introduced directly by our method.

7.1. Introduced by the Pose Transfer Framework

The following limitations can be addressed in future work by utilizing newer pose transfer frameworks, which, we expect, will focus on addressing these shortcomings:

L1. Videos with multiple people cannot be generated. This is because ExAvatar only supports pose transfer of videos with a single protagonist.

L2. Actions that involve interaction with objects cannot be generated convincingly, as the interaction with the object will not be transferred. This is because ExAvatar does not support object interactions.

search. While the model occasionally exceeded 10% accuracy on the validation set after some epochs, it consistently regressed below this threshold.

[4]This was caused by a tendency for over-predicting one class. We repeated this experiment on different seeds but this phenomenon persisted.

7.2. Introduced by Our Method

The following limitations can be addressed in future work by further improving our method (possible directions are discussed in Section 8):

L3. In some videos generated with our method, the pose-transferred human action, superimposed atop the background image, may not be physically plausible in space. This is because our method currently does not reflect the semantics and depth element of the scene depicted in the background image. See examples in Figure 12.

L4. Action classes that are associated with a particular setting may be generated in unnatural scenes or parts of scenes. Consider, for example, any cooking-related action classes (*Cook: cut* or *Cook: stir* in Toyota Smarthome), which would usually occur in the kitchen. Our method may generate new videos of these actions in inappropriate settings (e.g., the living room). See examples in Figure 13.

8. Discussion

Our results demonstrate that the method presented in this paper enhances human action video classification model performance in both baseline, one-shot, and few-shot learning scenarios. By providing additional training samples that increase the identity diversity of the training data, our approach improves the model's ability to generalize from limited real data. Furthermore, it enhances background diversity, which proved particularly effective for NTU RGB+D, where all videos are captured in just two locations.

We expect this to be especially beneficial in applications where collecting large amounts of labeled video data is impractical or cost-prohibitive. Examples include sign language translation for low-resource sign languages and video processing for autonomous driving in challenging areas—such as cities and countries that have not been extensively mapped or scanned.

We believe that our method will play an important role in mitigating the bias of computer vision datasets and algorithms across various video understanding tasks. It has been shown that bias towards certain demographic groups in computer vision systems often lies in underrepresentation in the underlying training datasets [9, 11]. This is where our method steps in: by generating new videos with protagonists of underrepresented demographic categories, it can balance such datasets.

Bias in computer vision systems is not limited to demographic groups, however. In the context of action recognition, for example, methods may suffer from a background bias, where the background is learned as a more predictive signal, leading the systems to ignore the actual human action [5]. Our method can be used to generate new training samples with varying backgrounds to mitigate this bias.

In terms of the employed pose transfer framework, we posit that methods combining 3D representations with recent generative AI techniques, similar to ExAvatar, will continue to be critical for synthetic video data generation. These approaches can leverage the fine-grained control and physical realism that come with 3D representations while enjoying the improved photorealism and scale of generative AI.

This, we believe, will also play an important role in addressing limitation L3: ensuring physical realism of the generated action in the new scene. Recent advancements in scene depth estimation [30, 38, 41] will allow one to place the generated human motion (with a precise 3D representation) in a plausible part of the scene.

Limitation L4 may be addressed with image understanding techniques such as multimodal large language models, VQA models, or multimodal embeddings. These could quantify the appropriateness of a scene for a particular action.

9. Conclusion

In this paper, we introduce a novel framework for synthetic data generation of human action video. By leveraging a modified ExAvatar framework for 3D Gaussian avatar animation, our method reenacts human actions from reference videos using novel human identities in varied settings. By combining computer graphics and generative AI frameworks, this approach addresses limitations in photorealism and semantic control that have hindered previous synthetic data generation methods for video understanding tasks.

We evaluate our method on a subset of the Toyota Smarthome and NTU RGB+D dataset, demonstrating notable improvements in action recognition performance across two video classification architectures. In particular, our method yields significant improvements in baseline, one-shot, and few-shot learning scenarios.

Finally, we present the *RANDOM People* dataset, which contains synthetic videos, novel human identity videos along with their avatars, and scene background images.

The dataset, code, and additional resources are open-sourced at `synthetic-human-action.github.io`. In future work, we aim to resolve limitations L3 and L4 pertaining suitable background selection and adjustment, extend our system's ability to capture and render human interactions with objects and further improve the photorealism of the resulting videos.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 3

[3] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 3

[5] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019. 8

[6] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 4

[7] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2022. 5, 1

[8] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 833–842, 2019. 2, 3

[9] Sepehr Dehdashtian, Ruozhen He, Yi Li, Guha Balakrishnan, Nuno Vasconcelos, Vicente Ordonez, and Vishnu Naresh Boddeti. Fairness and bias mitigation in computer vision: A survey. *arXiv preprint arXiv:2408.02464*, 2024. 8

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223: 103552, 2022. 8

[12] Haoqi Fan, Tullie Murrell, Heng Wang, Kalyan Vasudev Alwala, Yanghao Li, Yilei Li, Bo Xiong, Nikhila Ravi, Meng Li, Haichuan Yang, et al. Pytorchvideo: A deep learning library for video understanding. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3783–3786, 2021. 6

[13] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 3, 6

[14] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images, 2021. 4

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 3

[16] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. DensePose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3

[17] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555, 2018. 3, 6

[18] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 3

[19] Hochul Hwang, Cheongjae Jang, Geonwoo Park, Junghyun Cho, and Ig-Jae Kim. Eldersim: A synthetic data generation platform for human action recognition in eldercare applications. *IEEE Access*, 11:9279–9294, 2021. 2

[20] Michael Jones and Neel Patel. Ai-generated images as data sources: The dawn of synthetic era. *arXiv preprint arXiv:2310.01830*, 2023. 1

[21] Michael Jones and Neel Patel. Ai-generated images as data sources: The dawn of synthetic era. *arXiv preprint arXiv:2310.01830*, 2023. 2

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 4

[23] Vaclav Knapp and Matyas Bohacek. Can pose transfer models generate realistic human motion? *arXiv preprint arXiv:2501.15648*, 2025. 3

[24] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 3

[25] Ziqi Li and Hao Chen. Data augmentation techniques for the video question answering task. *arXiv preprint arXiv:2008.09849*, 2020. 1

[26] Hao Liu and Ting Wu. Promptonomyvit: Multi-task prompt learning improves video transformers using synthetic scene data. *arXiv preprint arXiv:2212.04821*, 2022. 2

[27] Hao Liu and Ting Wu. Promptonomyvit: Multi-task prompt learning improves video transformers using synthetic scene data. *arXiv preprint arXiv:2212.04821*, 2022. 2

[28] Wayne Wu Liu, Xian Zhang, Cheng Li, and Chen Change Loy. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-

person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[30] Zhengyang Lu and Ying Chen. Self-supervised monocular depth estimation on water scenes via specular reflection prior. *Digital Signal Processing*, 149:104496, 2024. 8

[31] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017. 3

[32] Yuming Men, Liming Jiang, Jianmin Zhang, Shuaicheng Liu, and Ming-Hsuan Yang. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 3

[33] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation, 2022. 4

[34] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3d gaussian avatar, 2024. 2, 3, 4

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019. 2

[37] 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, 2019. 3, 4

[38] Zihan Qin, Jialei Xu, Wenbo Zhao, Junjun Jiang, and Xianming Liu. Adaptive stereo depth estimation with multi-spectral images across all lighting conditions. *arXiv preprint arXiv:2411.03638*, 2024. 8

[39] Arun V Reddy, Ketul Shah, William Paul, Rohita Mocharla, Judy Hoffman, Kapil D Katyal, Dinesh Manocha, Celso M De Melo, and Rama Chellappa. Synthetic-to-real domain adaptation for action recognition: A dataset and baseline performances. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11374–11381. IEEE, 2023. 2

[40] Prasun Roy, Saumik Bhattacharya, Subhankar Ghosh, and Umapada Pal. Multi-scale attention guided pose transfer. *arXiv preprint arXiv:2202.06777*, 2022. 3

[41] Mohamed Sayed, Filippo Aleotti, Jamie Watson, Zawar Qureshi, Guillermo Garcia-Hernando, Gabriel Brostow, Sara Vicente, and Michael Firman. Doubletake: Geometry guided depth estimation. In *European Conference on Computer Vision*, pages 121–138. Springer, 2025. 8

[42] David Schneider, Marco Keller, Zeyun Zhong, Kunyu Peng, Alina Roitberg, Jürgen Beyerer, and Rainer Stiefelhagen. SynthAct: Towards generalizable human action recognition based on synthetic data. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13038–13045. IEEE, 2024. 2

[43] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2, 3, 5, 1

[44] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 3

[45] Khurram Soomro, Amir Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *ArXiv*, abs/1212.0402, 2012. 3

[46] DataCamp Team. Synthetic data generation: A hands-on guide in python, 2024. 1

[47] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3

[48] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287, 2021. 2

[49] Xin Wang, Kai Xu, et al. Large-scale deep learning optimizations: A comprehensive survey. *arXiv preprint arXiv:2111.00856*, 2021. 1

[50] Liyuan Xu and Yifan Zhang. Generative ai for synthetic data generation: Methods and applications. *arXiv preprint arXiv:2403.04190*, 2024. 1

[51] Liyuan Xu and Yifan Zhang. Generative ai for synthetic data generation: Methods and applications. *arXiv preprint arXiv:2403.04190*, 2024. 2

[52] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3

[53] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. 4

[54] Wei Zhang and Yao Liu. Learn2augment: Learning to composite videos for data augmentation in action recognition. *arXiv preprint arXiv:2206.04790*, 2022. 1

[55] Hao Zhu, Xiaoqian Huang, Hongwei Shi, Xiaoguang Li, Ran He, and Zhenan Wang. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 3

[56] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*, 2024. 3

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

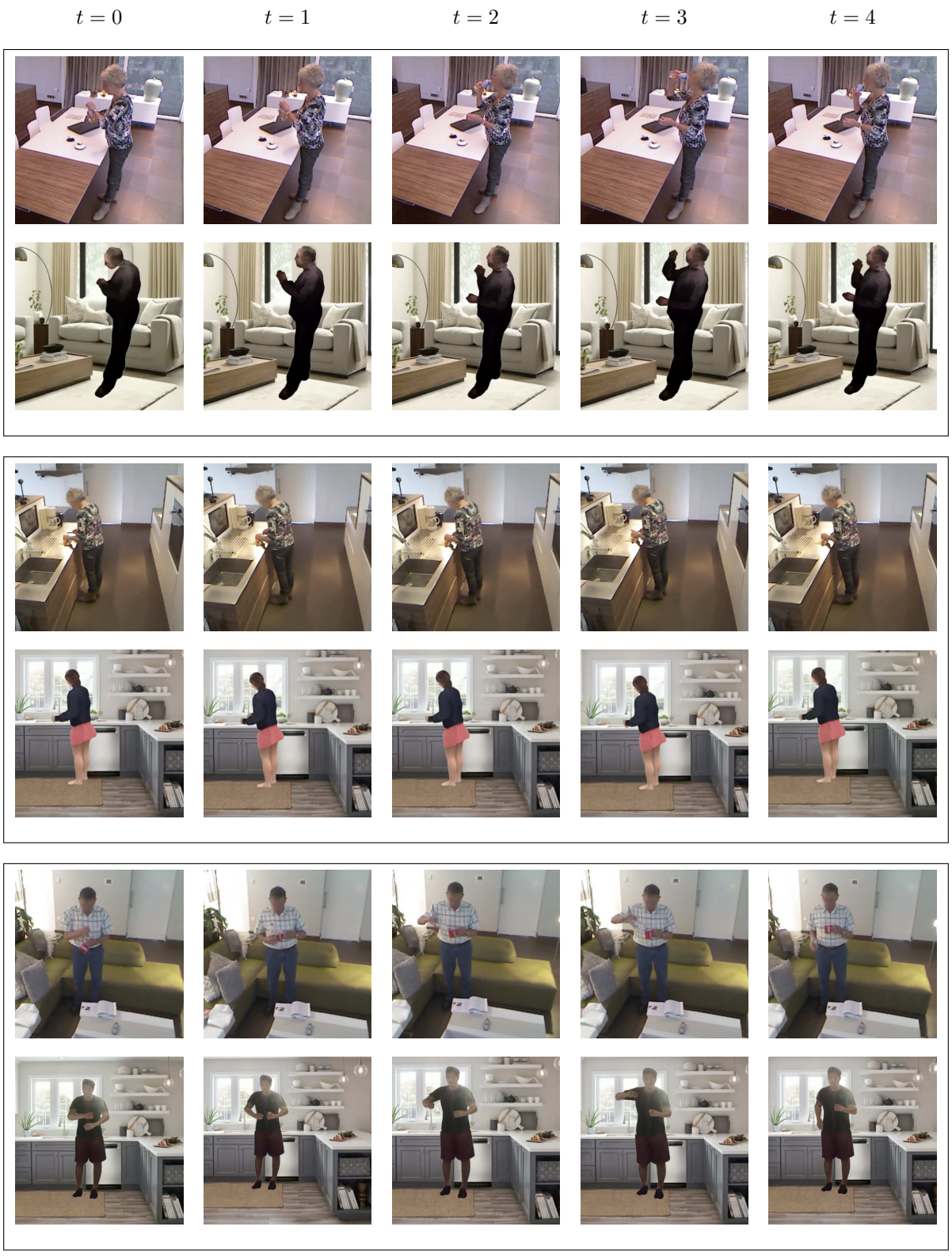


Figure 8. Examples of video frames at $t = \{0, 1, 2, 3, 4\}$ seconds from the source video (top), taken from Toyota Smarthome, and the target video (bottom), generated by our synthetic data generation method, where the pose alignment is consistent.

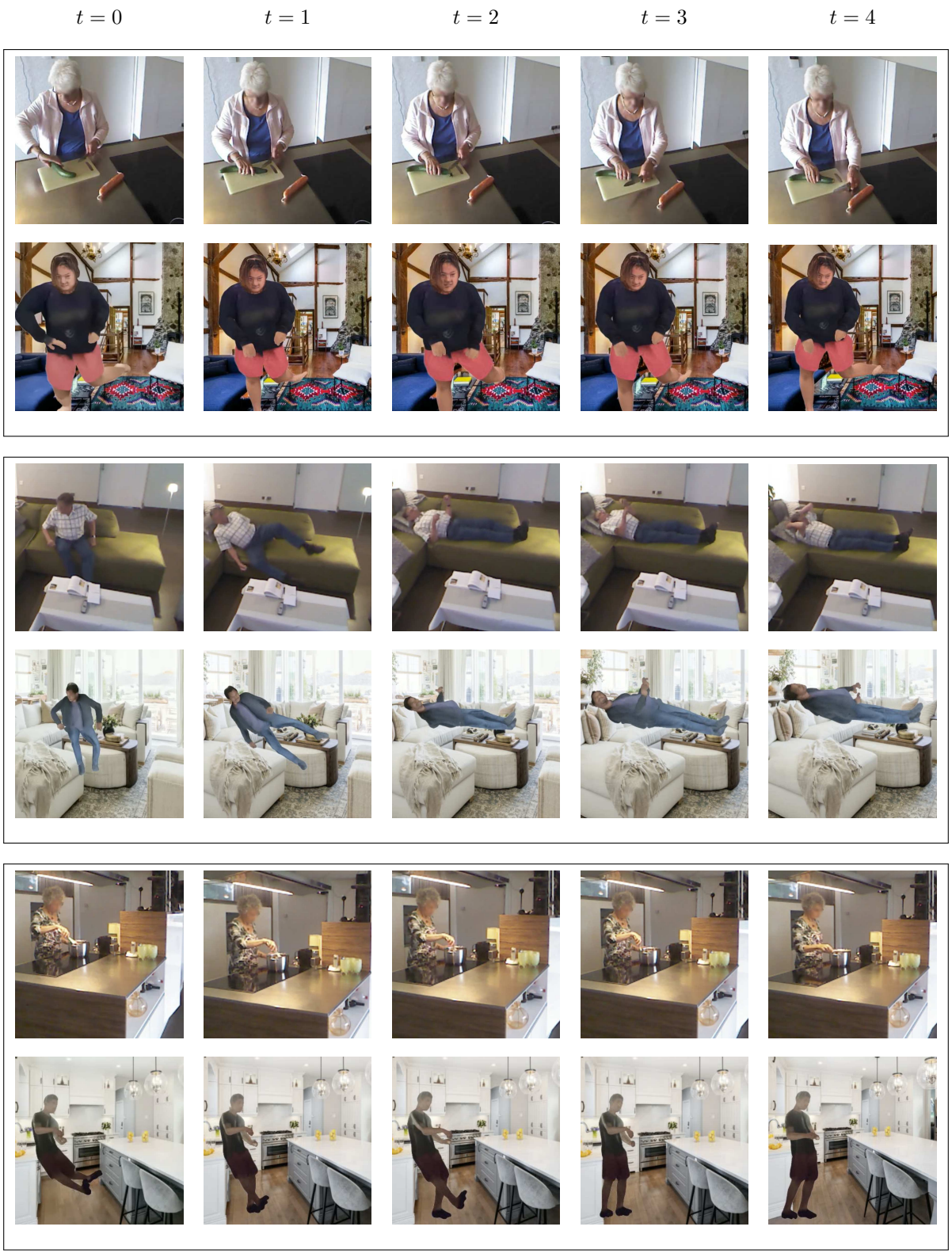


Figure 9. Examples of video frames at $t = \{0, 1, 2, 3, 4\}$ seconds from the source video (top), taken from Toyota Smarthome, and the target video (bottom), generated by our synthetic data generation method, where the pose alignment is inconsistent.

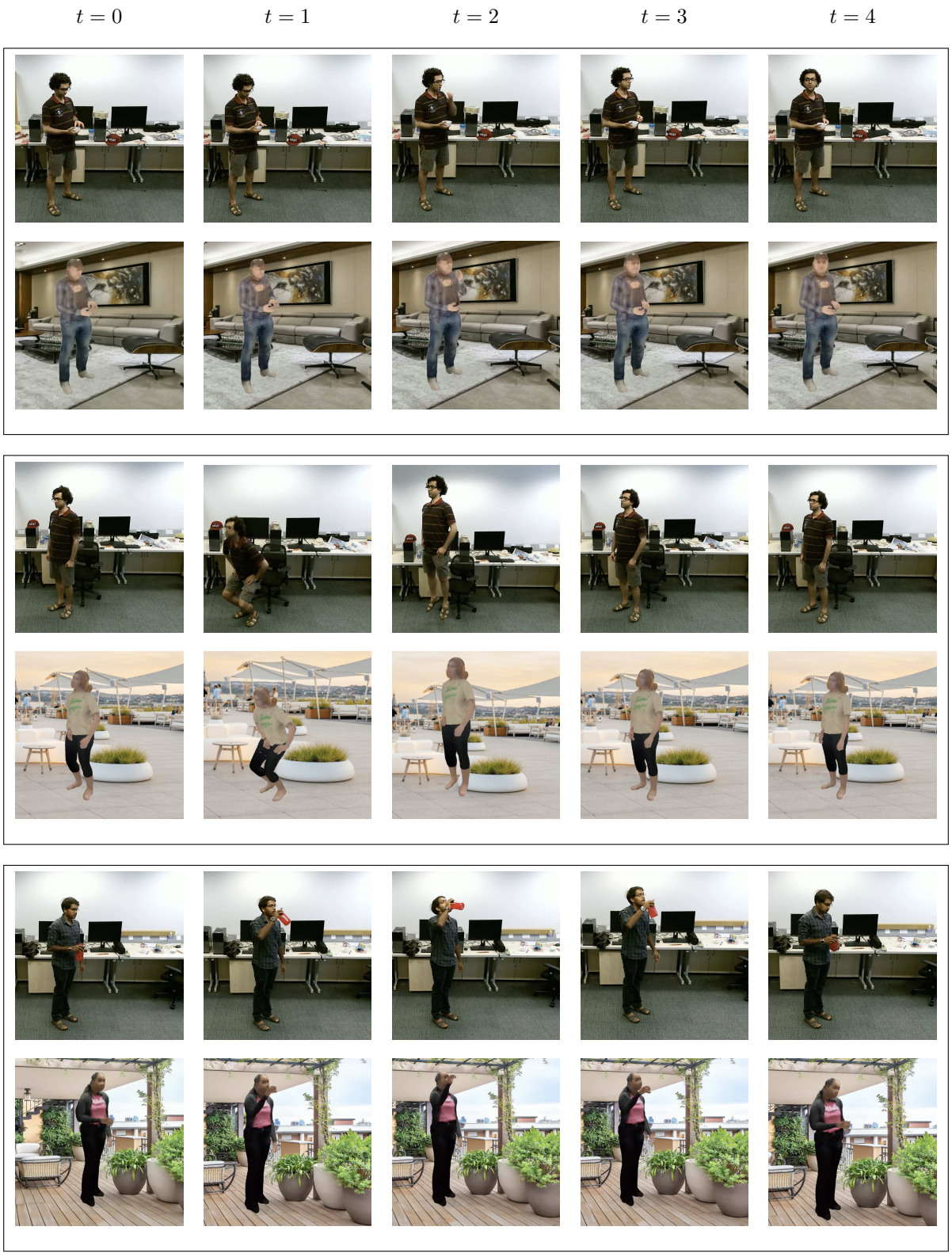


Figure 10. Examples of video frames at $t = \{0, 1, 2, 3, 4\}$ seconds from the source video (top), taken from NTU RGB+D dataset, and the target video (bottom), generated by our synthetic data generation method, where the pose alignment is consistent.

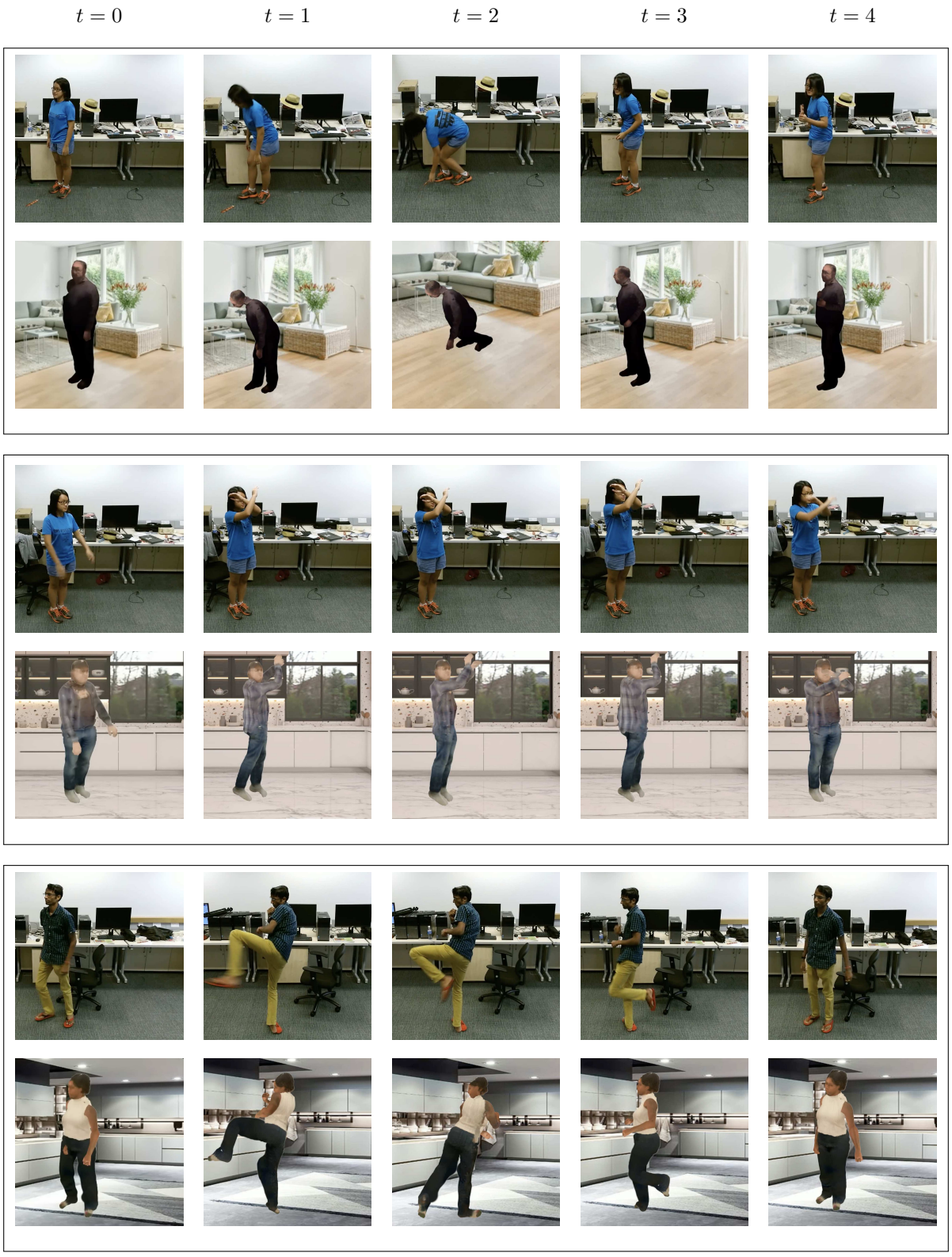
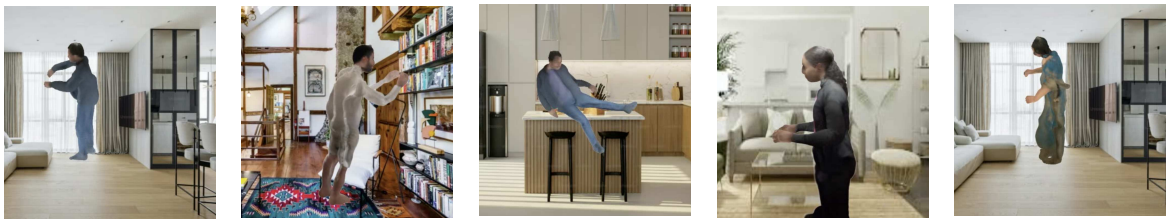


Figure 11. Examples of video frames at $t = \{0, 1, 2, 3, 4\}$ seconds from the source video (top), taken from NTU RGB+D dataset, and the target video (bottom), generated by our synthetic data generation method, where the pose alignment is inconsistent.

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.