

A. Supplementary Materials

A.1. InfantAction Dataset Creation

We recruited infant subjects and collected clips from home-based monitoring sessions, capturing moments when the infants were either playing or sleeping. This process was conducted with IRB approval and parental permission. Our infant participants ranged in age from 3 to 12 months, which introduced significant variation in their motion capabilities. To create the final InfantAction dataset, we undertook the following steps:

- **Video Clips Cropping:** We reviewed lengthy video recordings to extract short clips (each approximately 4-5 seconds) that showcased predefined actions such as “Sitting”, “Standing”, “Crawling”, and “Rolling”.
- **Video Selection:** We selected clips that were of high quality and displayed clear action movements, ensuring a variety of movements for each subject. Each clip was manually assigned an action class label.
- **Object Detection:** Using YOLOv8 [31], we automatically detected bounding boxes for each subject. We employed object tracking algorithms to maintain consistency in the bounding boxes, manually correcting any inaccuracies.
- **3D Pose Estimation:** As our videos were solely RGB with no motion capture data, we applied the HW-HuP [20] infant 3D pose estimation model to determine joint locations in each frame.
- **Error Filtering:** After pose estimation, we visualized the predicted 3D poses and removed any clips with incorrect estimations.

Following these processing steps, we compiled a dataset of 273 video clips. The class distribution of these clips is detailed in Tab. 1.

A.2. Implementation Details

Experiments on InfActPrimitive Dataset We deployed our InfAGenC framework on a transformer-based VAE model integrated with a ST-GCN. This model leverages the 6D rotations of SMIL model [12] 24 joints’ as the joints representation, offering a detailed and comprehensive depiction of the dynamic interactions between joints. Each video clip was processed to consist of 60 frames. During the training phase, we utilized the Adam optimizer with a learning rate set to 0.0001 and a batch size established at 16 for epochs. Initially, to ensure accurate performance evaluation, our action recognition component underwent training for 15 epochs using the InfActPrimitive dataset.

To ensure effective evaluation of generated samples, we pre-train an action recognition model. However, it’s essential to strike a balance in training this model. Over-training can reduce synthetic data diversity due to overfitting, while under-training may lead to inaccurate action classification. To address this, we halt the pre-trained model’s training once it achieves 85% accuracy, ensuring both model performance and synthetic data quality.

For the action generation component, we adjusted the loss term weights— λ_{KL} , λ_{rec} , and λ_{vel} —to 1.0, 1.0, and 0.001, respectively. This component was pre-trained for 1100 epochs on the training set of InfActPrimitive, aiming to enrich the generated samples with temporal details beyond mere static poses. Subsequently, we initiated the synthetic data recycling phase for an additional 200 epochs. In our strategy for filtering and selecting generated samples, we set the confidence threshold (θ) at 0.75 and the weights for within-class distance (w_{within_i}) and between-class distance ($w_{between_i}$) at 0.6 and 0.4, respectively.

Upon completing the training of our infant action generative model, we successfully generated 1275 synthetic samples, which were then incorporated into the training set for the action recognition models. This synthetic data was further incorporated into the training set for training the action recognition models up to 100 epochs but ceasing upon model convergence, specifically targeting the infant action recognition task.

Experiments on InfantAction Dataset For the experiments conducted on the InfantAction dataset, we followed a similar configuration to that of the InfActPrimitive dataset, with the main difference being the adjustment of the number of frames per video clip to 90 instead of 60. This adjustment was made to accommodate the longer duration required to capture complex actions accurately. The remaining settings, including the model architecture, optimizer, loss term weights, training duration, and synthetic data recycling strategy, remained consistent. Following the training process, we successfully generated 816 synthetic samples, which were seamlessly integrated into the training set for action recognition models.

Experiments on Prepared NTU Dataset We adopted a different data representation due to the availability of relatively accurate joint location annotations of NTU data. Instead of relying on 24 joints’ 6D rotations, we directly utilized the 25 joints’ 3D coordinates as the data representation for both the generative and recognition models. The duration of videos in this dataset was set to 90 frames as well. Similar to the previous experiments, we generated synthetic samples during the training process. Upon completion, we successfully generated 1288 synthetic samples, which were then incorporated into the training set for further analysis and evaluation of the action recognition models.

All the experiments utilize a robust compute environment featuring the NVIDIA v100-pcie GPU from the Volta generation. This GPU comes equipped with 32GB of memory, enabling substantial data processing capabilities.

Subject ID	Supine	Prone	Sitting	Standing	All-fours	Total
D01	23	75	0	0	0	98
D02	1	1	34	21	22	79
D03	0	6	138	37	70	251
D04	0	45	0	0	0	45

Table S1. The distribution of action classes within the “in-the-wild” segment of InfActPrimitive varies for each infant participant.

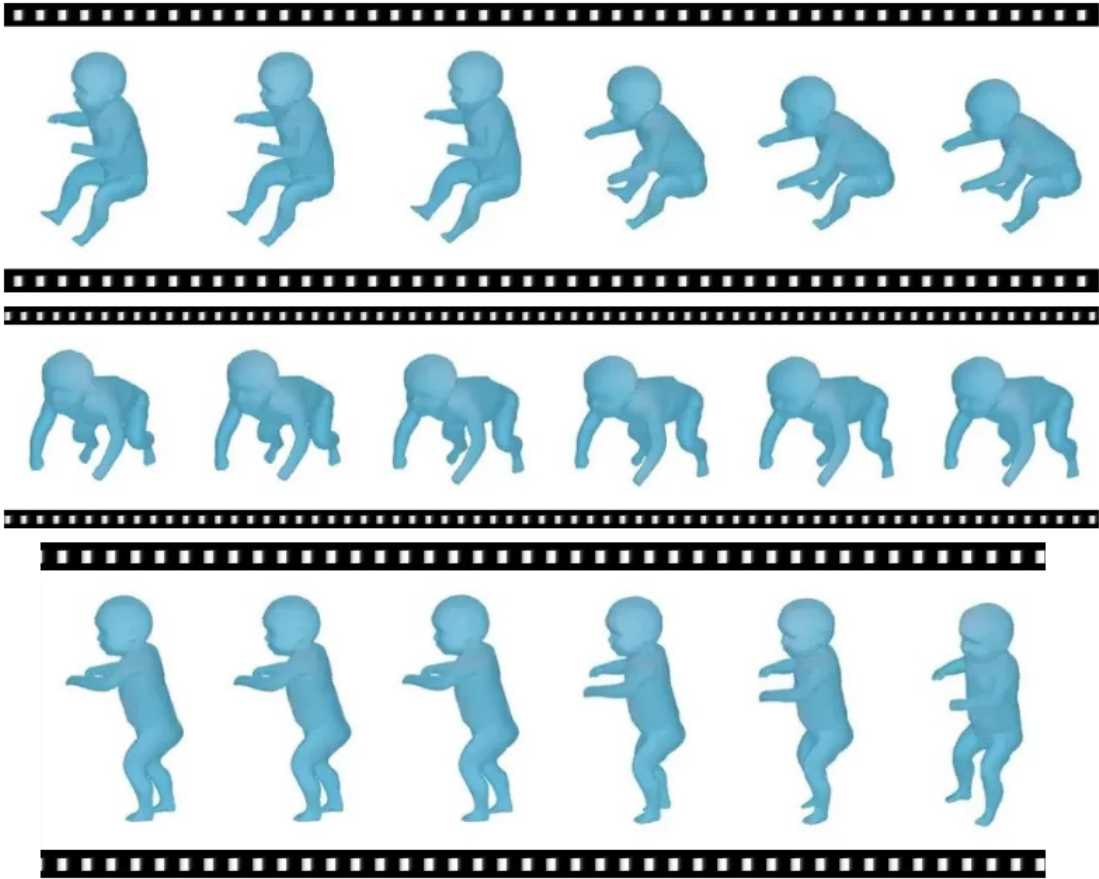


Figure S1. Sample Sequences of Generated Actions. Each sample’s frames are extracted from a generated action sequence spanning 3 seconds. The actions, displayed sequentially from top to bottom, are: Sitting, Crawling, Standing.

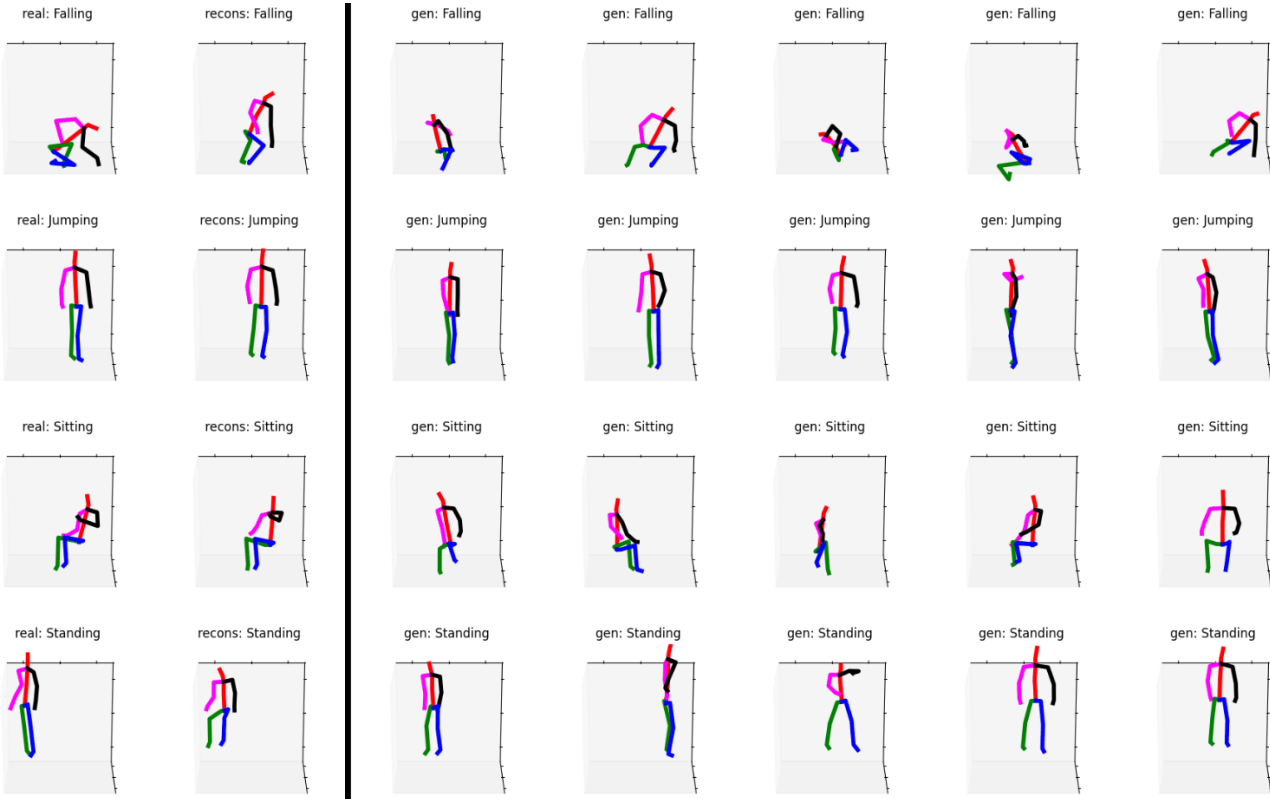


Figure S2. Snapshots of Generated Adult Action Samples. The generated samples are produced by trained our generative model on our prepared small NTU Dataset with four action classes: Falling, Jumping, Sitting, and Standing. Each row shows one action class samples.