Infant Action Generative Modeling

Xiaofei Huang[1★], Elaheh Hatamimajoumerd[1★], Amal Mathew[1], Sarah Ostadabbas[1*]
[1]Augmented Cognition Lab (ACLab), Department of Electrical and Computer Engineering,
Northeastern University, Boston, MA, USA
★These authors contributed equally to this work.
*Corresponding author: ostadabbas@ece.neu.edu

Abstract

Despite advancements in human motion generation models, their performance drops in infant motion generation due to limited data available and lack of 3D skeleton ground truth. To address this, we introduce the infant action generation and classification (InfAGenC) pipeline, which combines a transformer-based variational autoencoder (VAE) with a spatial-temporal graph convolutional network (ST-GCN) to create synthetic infant action samples. By iterative refinement of the generative model with diverse and accurate data, we improve the realism of synthetic data, leading to more precise infant action recognition models. Our results show significant improvements in action recognition performance on real-world data, demonstrating that synthetic data can enhance small training datasets and advance infant action recognition. Our pipeline increases action recognition accuracy up to 88.58% on the infant action dataset and up to 98% on an adult action dataset[1].

1. Introduction

Despite the significant advancements in current vision-based human action recognition (HAR) models, which leverage extensive datasets like NTU RGB+D 120 [21], Human3.6M [17], and N-UCLA [36] with abundant samples per class, infant activity recognition is still in its infancy. High costs associated with collecting and labeling infant data, coupled with concerns regarding security and privacy, have resulted in a significantly limited pool of infant data available for model training. Recent research [12] highlights a challenge in transferring knowledge from adult-based HAR models to infant datasets, resulting in poor performance. This discrepancy stems from differences in action types and settings between adults and infants. Adult datasets [19, 21, 31, 33] often feature actions like taking selfies or shaking hands, absent in infant behavior. Conversely, infant actions like crawling are not typically found in adult datasets. Variations in how common actions are performed further hinder generalization. Unlike controlled lab setups for adults, obtaining precise infant datasets is challenging due to infants' uncooperative nature.Consequently, most current accessible infant datasets like InfAct [16] and InfActPrimitive [12] lack 3D skeleton ground truth and are sourced from diverse camera angles without control, often collected through YouTube.

Recent advances in generative models [10, 11, 35, 40] have facilitated the creation of synthetic motion data to augment datasets. However, their effectiveness is hindered by a reliance on large datasets for training, posing a challenge when applied to smaller datasets. Models such as generative adversarial networks (GANs), variational autoencoders (VAEs) struggle to generate sufficient data for limited datasets due to their need for extensive training data. This issue is notably pronounced in specialized areas such as infant motion recognition, where accurately capturing domain-specific interactions and knowledge from sparse real data is exceptionally challenging [12, 15].

Given the limited infant action data and domain gap between real and synthetic datasets, we successfully generate high-quality skeleton-based infant motions by introducing our infant action generation and classification (InfAGenC) pipeline depicted in Fig. 1. InfAGenC pipeline harnesses the synergistic capabilities of a transformer-based VAE [27] for data generation, combined with a spatial-temporal graph convolutional network (ST-GCN) [38] for action recognition. Drawing inspiration from active learning's pool-based sampling method [39], our approach selectively enriches the dataset. Every ten iterations, we identify and incorporate high-quality synthetic samples—evaluated on accuracy and diversity—using between- and within-class distance measurements. These carefully chosen samples, alongside real data, are reintroduced to the system, creating a more robust and diverse dataset for model training.

[1] The InfAGenC code and our infant skeletal data available at https://github.com/ostadabbas/Infant-Action-Generative-Modeling.

Supported by NSF-CAREER Grant #2143882.

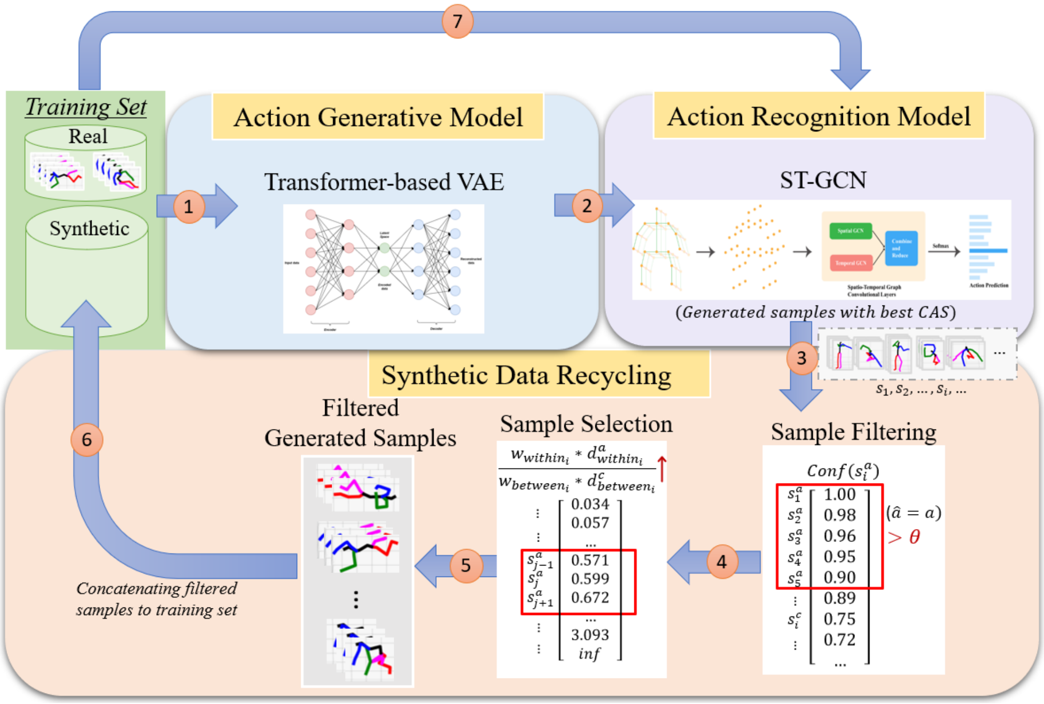


Figure 1. Overview of Our Infant Action Generation and Classification (InfAGenC) Pipeline. The diagram depicts the cyclic pipeline, beginning with Step 1, where real and synthetic data form the initial training set for the transformer-based VAE pose generative model. In Step 2, the action recognition model, an ST-GCN, evaluates the generated samples to identify those with the highest Classification Accuracy Score (CAS). Step 3 applies a sample filtering technique to retain high-confidence samples, while Step 4 involves a sample selection process based on weighted within-class and between-class distance metrics. The central portion of these samples is then selected in Step 5 for synthetic data recycling. In Step 6, the filtered high-quality samples are concatenated back to the training set. Finally, Step 7 closes the loop by updating the action recognition model with the augmented training set, preparing it for the next cycle of evaluation.

Compared to GANs [37], VAEs are known for their more stable training dynamics and the ability to generate interpretable latent spaces, which facilitate a more accurate assessment of diversity scores. Recognizing the variability in infants' movement speeds across different actions, our approach enhances motion consistency between real and synthetic data by integrating velocity as a factor in the loss function through the use of regularization metrics.

Our experimental outcomes indicate a marked improvement in the performance of action recognition models test on "in-the wild" videos of InfActPrimitive dataset, achieving increasing accuracy to more than 15%. In summary, this paper introduces following significant contributions:

- Developing a novel pipeline, called InfAGenC that combines a transformer-based VAE for generation with a spatial-temporal graph convolutional network (ST-GCN) for recognition, aiming to enhance infant action recognition with limited data and bridging gap between real and synthetic data.

- Implementing a pool-based sampling approach to enrich the dataset, selecting samples that offer an optimal balance between accuracy (determined by predicted confidence scores) and diversity (assessed through between- and within-class distances).

- Creating an infant action dataset (InfantAction) with more complex action movements compared to the existing InfActPrimitive data. InfantAction includes daily activity data from 5 infant subjects, covering broader action classes such as Sitting, Standing, Rolling, and Crawling.

2. Related Work

Skeleton-based methods for Human Activity Recognition (HAR) [5, 8, 23, 28, 38]emerged as a prominent choice due to their ability to efficiently represent human movements using joint coordinates, effectively minimizing potential disruptions caused by RGB appearance variations. ST-GCN [38] introduces inter-frame edges, connecting corresponding joints across consecutive frames, enhancing inter-frame relationship modeling and improving temporal dynamics understanding within skeletal data. MS-G3D [23] combines multi-scale graph convolutions into a unified G3D module, enhancing long-range modeling by prioritizing nodes in different neighborhoods. It utilizes dense cross-spacetime edges as skip connections for direct information propagation across the spatial-temporal graph. InfoGCN [5] merges a learning objective with an encoding method utilizing attention-based graph convolution, capturing discriminative information regarding human actions.

However, the above-mentioned models have been trained on the enormous data backbones such as NTU RGB+D 120 [21] containing 120 kinds of actions, a total of 114,480 samples, in the form of depth, 3D skeleton, RGB and infrared sequence, Kinetics-700 [4], a video dataset of 650,000 clips that covers 700 human action classes, and BABEL [29], a large dataset with language labels describing the actions being performed in mocap sequences. There are over 28k sequence labels, and 63k frame labels in BABEL, which belong to over 250 unique action categories. When it comes to infant studies, Recently, several infant-specific image and video datasets have been released, each are tailored for specific applications [1, 2, 41, 42]. SyRIP [15] and MINI-RGBD [13] respectively with 17 and 24 annotated joints were created benchmark for a standardized evaluation of pose estimation algorithms in infants. AggPose [3] was proposed to train a deep aggregation transformer for infant pose detection. They adopted general movements assessment (GMA) devices to record infant movement videos in supine position. More than 216 hours of videos and 15 million frames were extracted. They randomly sampled 20,748 frames from the videos and let professional clinicians annotate infant 21 keypoints locations. Both MINI-RGBD and AggPose have considerable amounts of data. However, All of these dataset contains image-only data and are not suited for task of infant action recognition. BabyPose [26] contains over 1000 videos of preterm infants captured using a depth-sensing camera along with annotations of 12 limb-joint positions for each frame. However, it only contains the data of newborns with no to limited motions on only supine body pose with one-fold background. The most relevant dataset available for infant action recognition is InfActPrimitive containing small set of 975 video clips from 5 class of action collected from two different sources: Youtube and recruited subjects.

Domain adaptation has been extensively used to bridge the gap between source (real) and target (synthetic) data distributions encountered during the application of Deep Neural Networks [30] [32]. Hatamimajoumerd et al. [12] fine-tuned state-of-the-art skeleton-based action recognition models pre-trained on adult skeleton datasets on InfActPrimitive. Their results indicate a remarkable gap between the action recognition results on infant and adult datasets.

Recent advancements in human pose estimation have led to the development of motion generation models using skeleton data. These models can enrich the training set and tackle the small data problem, specifically in the infant action recognition domain, where data collection is difficult [6, 10, 18, 20, 25]. Tevet et al. [35] developed Motion Diffusion Model (MDM) a transformer-based generative model for human motion, prioritizing sample prediction over noise in diffusion steps while employing geometric losses. Degardin et al. [7] introduced Kinetic-GAN, an architecture blending Generative Adversarial Networks and Graph Convolutional Networks, capable of conditioning up to 120 actions. Compared to GAN-based models, VAEs offer more stable training and interpretable latent spaces, aiding in accurate diversity score assessment. Lucas et al. [24] introduced PoseGPT, an auto-regressive transformer-based approach which internally compresses human motion into quantized latent sequences. Feng et al. [9] leveraged Large Language Models (LLMs) to directly generate 3D human body poses from images or text by embedding SMPL poses within a multi-modal LLM. However, they have not been explored in small data domains like infant motion, where unpredictability and lack of 3D skeleton ground truth present significant challenges.

3. Method

In this section, we introduce infant action generation and classification (InfAGenC) pipeline, crafted to improve action recognition by creating a variety of accurate samples, especially within the challenging context of small data domains, with a particular emphasis on infant action recognition–a prime scenario of data scarcity.

Problem Formulation Consider a skeleton dataset $X = \{x_1, x_2, \ldots, x_n\}$, where each sample x_i encapsulates a sequence of joint locations or axis angle values $x_i \in \mathbb{R}^{K\times R\times T}$. Here, K denotes the number of infant joints, R the dimensionality of the joint representation, and T the sequence length. Each sample x_i is associated with an action class a_i belonging to a set of m actions $A = \{a_1, a_2, \ldots, a_m\}$. Our aim is to devise a generative model $G(X;\theta)$ adept at mirroring the original data distribution $p_{\text{data}}(x)$. The purpose of G is to fabricate synthetic action sequences s, such that $s \sim p_{\text{model}}(x)$, with $p_{\text{model}}(x;\theta)$ closely emulating $p_{\text{data}}(x)$. This entails that the synthetic data should not only exhibit diverse pose variations but also faithfully preserve the intrinsic motion styles and statistical characteristics inherent to the original dataset. Additionally we define an action recognition model, denoted as $M(S)$, which dynamically evaluates each synthetic sample s_i within the set of generated action sequences $S = \{s_1, s_2, \ldots, s_p\}$ to determine if the sample is good enough to be assigned into the correct action class, which means predicted action equals to the given class, noted as $\hat{a}_i = a_i$.To highlight the capabilities of the generative model, we amalgamate the refined, high-quality synthetic action sequences S with the infant real data X, forming a comprehensive training dataset $X_{train} = \{s_1, s_2, \ldots, s_p\} \cup$

$\{x_1, x_2, \dots, x_n\}$, which is employed to further augment the training of generative model G and recognition model M.

3.1. InfAGenC Pipeline

Motion generation, action recognition, and sample selection/filtering are major components of our pipeline. As depicted in Fig. 1, the process begins by feeding real data, which is a set of body keypoints/pose angles, into a pretrained generative model. Subsequently, theresultant samples undergo evaluation by an action recognition model. Utilizing predicted confidence scores and diversity scores, selected samples are transferred to the synthetic data pool for integration into the training data for subsequent iterations. Each component will be thoroughly described below.

Action Generation We adopted a transformer-based conditional VAE model G, and trained it on the training data initially comprising only real infant data X, serving as our baseline generative model. Following this, we embark on the integrated training of our motion generative model on the training set X_{train}. This phase involves iterative training on the set, focusing on the optimization of losses including reconstruction loss $\mathcal{L}_{rec}$, motion/velocity loss $\mathcal{L}_{vel}$, and Kullback–Leibler (KL) divergence loss $\mathcal{L}_{KL}$. The equation representing the loss function of the generative model can be formalized as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{KL}\mathcal{L}_{\text{KL}} + \lambda_{rec}\mathcal{L}_{\text{rec}} + \lambda_{vel}\mathcal{L}_{\text{vel}}, \tag{1}$$

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|^2 \tag{2}$$

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2}\sum_{j=1}^{d}\left(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\right) \tag{3}$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(\mathbf{x}_{i+1} - \mathbf{x}_i) - (\hat{\mathbf{x}}_{i+1} - \hat{\mathbf{x}}_i)\|^2 \tag{4}$$

where λ_{KL}, λ_{rec}, and λ_{vel} represent the weights assigned to each respective loss term.

Action Recognition Parallel to the generative model's development, the action recognition model training plays a pivotal role in the methodology. A well-trained action recognition model M, specifically the ST-GCN in our scenario, is prepared. This model also trained on the original real infant motion data X is crucial for evaluating the generated samples' quality and performing action classification tasks effectively. Cross-entropy loss is the loss function to minimize. Following each generative model G training iteration, it produces samples $\{s_i^{t_k}\}_{i=1}^{N_k}$ and N_k is the number of samples generated at epoch t_k, that are subsequently evaluated by the ST-GCN model to assess their classification accuracy score (CAS). After every 10 epochs, the samples achieving the highest CAS are selected as candidates for the next synthetic data filtering, ensuring a continuous improvement in sample quality. This can be formalized as:

$$t_{\text{best}} = \underset{t_k \in \{t_1, t_2, \dots, t_{10}\}}{\text{argmax}} \frac{1}{N_k}\sum_{i=1}^{N_k} \text{CAS}(s_i^{t_k}). \tag{5}$$

Subsequently, the set of generated samples from the best-performing epoch, t_{best}, is selected as candidate:

$$S_{\text{candidate}} = \{s_i^{t_{\text{best}}}\}_{i=1}^{N_{t_{\text{best}}}}. \tag{6}$$

Notably, with the training set X_{train} augmented by synthetic data, the action recognition model M undergoes 3 epochs of updates to adapt to the new input data. The updated model is then utilized for evaluating the next cycle of generated samples. This update ensures that the model remains attuned to the nuances of both original and newly generated synthetic samples, thereby maintaining high accuracy and efficiency in action classification tasks.

Sample Filtering and Selection Before transferring generated samples to next training iterations, we discard low-quality samples in the set $S_{\text{candidate}}$ based on their classification accuracy and confidence scores. We retain only samples that are correctly classified and exceed a predefined confidence threshold defined in Eq 7, ensuring high-quality motion samples in the synthetic data recycling process. This filtering maintains integrity and quality of the training data.

$$S_{\text{filtered}} = \{s_i \in S_{\text{candidate}} | \text{Conf}(s_i) > \theta \text{ and } \hat{a}_{s_i} = a_{s_i}\}, \tag{7}$$

where S_{filtered} is the set of samples retained after filtering, $\text{Conf}(s_i)$ represents the confidence score of sample s_i, θ is the predefined confidence threshold, and $\text{Acc}(s_i)$ is a boolean indicating whether sample s_i was correctly classified by the action recognition model.Then, given the inherent limitation of VAE models to sample data closely aligned with the distribution of their training dataset, which, in the context of our minimal real training data, restricts the diversity of generated samples, we employ a strategic sample selection method aimed at enhancing the training dataset's variance. This method is quantitatively defined by two essential metrics for each sample s_i in filtered samples S_{filtered}: *Within-Class Distance, d_{within}*, the average of Euclidean distances between the feature vector of synthetic samples and the feature vectors of all real samples within the same class:

$$d_{\text{within}_i} = \frac{1}{N_{a_{s_i}}}\sum_{j=1}^{N_{a_{s_i}}} \|f(s_i) - f(x_j)\|_2\,, \tag{8}$$

Between-Class Distance, $d_{between}$, the average of the Euclidean distances between the feature vector of the synthetic sample and the feature vectors of all real samples from different classes:

$$d_{\text{between}_i} = \frac{1}{N_{a\neq a_{s_i}}} \sum_{j=1}^{N_{a\neq a_{s_i}}} \|f(s_i) - f(x_j)\|_2 \,, \qquad (9)$$

where the $N_{a_{s_i}}$ is the number of real data points x_j with the same action class label of the synthetic sample s_i, while $N_{a\neq a_{s_i}}$ is the number of real data points x_j with the different action class label of the synthetic sample s_i. $f(\cdot)$ is the encoder of action recognition model, which extracts the feature vector of motion sample. To reconcile these objectives, we evaluate the ratio of weighted within-class to between-class distances for each sample, prioritizing samples based on this ratio within each class and ultimately retaining the central 50% of samples to balance diversity and distinction. These selected samples $S_{selected}$ are then used as as part of training data $X_{train} = \{S_{selected} \cup X\}$ for further generative model training and contribute to building our initial infant synthetic motion dataset.

4. Experimental Results

We evaluated our pipeline using two Infant dataset including InfactPrimitive [12] and our collect infant action dataset. We have also used the small portion of NTU [21] to illustrate our pipeline's performance with limited data. With 3D ground truth provided, sourced from a standard skeleton representation, we could effectively control potential noise in estimation, underscoring the robustness of our approach. Detailed implementation our pipeline and experiment hyperparameters are provided in the Supplementary Materials Sec. A.2. We evaluated the models through both generation metrics and action recognition accuracy in Sec. 4.2.

4.1. Datasets

NTU The NTU RGB+D 120 [21] dataset is a widely-used dataset for the human activity recognition task. We sampled data from the NTU dataset to create a smaller set for training our pipeline. We selected four classes: "Sitting Down", "Standing Up", "Jumping", and "Falling", which are similar to infant action classes. Our training set includes 5 samples per class from 5 subjects, totaling 100 data points. For testing, we created two sets: (1) a small test set with 10 samples per class from another 5 subjects (200 samples in total), comparable to the size of the infant datasets, and (2) a large test set that includes all samples from the remaining subjects for these four classes, totaling 3081 samples, to provide more robust evaluations.

InfActPrimitive The InfActPrimitive [12] dataset comprises video clips from YouTube and real-life scenarios, involving 127 infant subjects. The YouTube portion initially had 400 clips, but after removing unreliable and outlier samples, 310 clips across five classes remained: 46 Supine, 40 Prone, 99 Sitting, 67 Standing, and 58 All-fours. We designated one-third of these clips (103 clips) as a validation set to maintain class balance and avoid subject overlap during training for action recognition and generation.We set the"in-the-wild" segment of InfActPrimitive as our test set, featuring home-based clips of infants aged 3 to 12 months engaged in various activities. The older infants often perform Sitting, Standing, and All-fours actions, while the younger ones primarily showed Supine and Prone actions. This led to a naturally imbalanced distribution of action classes, reported in the Supplementary Materials Tab. S1.

Subject ID	Crawling	Sitting	Standing	Rolling	Total
Inf01	31	41	10	0	82
Inf02	43	1	0	15	59
Inf03	0	1	0	3	4
Inf04	52	0	0	9	61
Inf05	6	24	27	0	57

Table 1. The distribution of action classes of action classes of our created InfantAction dataset.

InfantAction (Ours) The InfActPrimitive dataset contains infant basic actions with limited body movements. To include more complex actions, we created a dataset through recruitment of five infants. They details of our study is discussed in the Supplementary Materials Sec. A.1. Our dataset contains four action classes: "Crawling", "Sitting", "Standing", and "Rolling". However, similar to InfActPrimitive [12] due to age variations among the infants, the distribution of action classes remained imbalanced, as detailed in Tab. 1. We included Inf04 and Inf05 in our test set which cover the whole action classes, and used the remaining 3 subjects in our training resulting a total of 145 and 118 samples in training and test sets.

The collection and usage of infant data under Institutional Review Board (IRB #22-11-32) strictly adhere to the highest ethical standards. All data handling and sharing are in full compliance with applicable data protection regulations, ensuring that the data remains secure and accessible only to authorized personnel for research purposes.

4.2. Evaluation of Infant Action Generation

Tab. 2 reports the performance comparison of our pipeline and the VAE baseline among all three datasets in terms of evaluating Fréchet Inception Distance (FID), action recognition accuracy, overall diversity, as well as diversity and multimodality on a per-action basis. Existing generative models, like MDM and Action2Motion, are primarily designed for adult datasets with a large number of samples. We also adapted and fine-tuned MDM model on the InfActPrimitive training set to benchmark its performance against our model in scenarios with limited data availabil-

Dataset	Model	FID ↓	Accuracy ↑	Diversity ↑	Multimodality ↑
InfActPrimitive [12]	MDM [40]	$17.38^{\pm 5.25}$	$98.3^{\pm 0.42}$	$19.22^{\pm 0.14}$	$6.95^{\pm 0.45}$
	VAE (Baseline)	$\mathbf{5.87^{\pm 1.49}}$	$97.6^{\pm 7.29}$	$18.45^{\pm 0.51}$	$6.78^{\pm 0.36}$
	InfAGenC (Ours)	$15.15^{\pm 4.08}$	$\mathbf{100.00^{\pm 0.09}}$	$\mathbf{26.65^{\pm 1.24}}$	$\mathbf{7.65^{\pm 0.36}}$
InfantAction (Ours)	VAE (Baseline)	$44.31^{\pm 69.39}$	$71.80^{\pm 19.70}$	$13.54^{\pm 0.86}$	$\mathbf{7.45^{\pm 0.70}}$
	InfAGenC (Ours)	$\mathbf{29.00^{\pm 7.45}}$	$\mathbf{98.6^{\pm 2.69}}$	$\mathbf{23.80^{\pm 2.20}}$	$5.32^{\pm 0.78}$
NTU [21]	VAE (Baseline)	$\mathbf{29.35^{\pm 18.81}}$	$72.60^{\pm 7.92}$	$14.02^{\pm 0.46}$	$4.35^{\pm 0.26}$
	InfAGenC (Ours)	$47.61^{\pm 20.68}$	$\mathbf{99.90^{\pm 0.18}}$	$\mathbf{22.84^{\pm 0.98}}$	$\mathbf{4.79^{\pm 0.35}}$

Table 2. Comparative Evaluation of the InfAGenC Model with the Baseline Conditional VAE Model. This table presents the performance metrics of our InfAGenC model versus the baseline model across the NTU, InfActPrimitive, and InfantAction datasets, and also make the contrast with MDM model on InfActPrimitvie dataset. The best results are bold.

Model	InfActPrimitive [12]		InfantAction (Ours)		NTU [21]	
	Real	Real + Syn	Real	Real + Syn	Real	Real + Syn
ST-GCN Acc. (%)	68.71	**88.58**	**80.51**	70.34	94.00	**98.00**
MS-G3D Acc. (%)	58.35	**66.38**	**66.10**	65.30	86.65	**88.28**

Table 3. Performance of Infant Action Recognition Models Using Different Data Configurations. We respectively trained ST-GCN model and MS-G3D model on InfActPrimitive, InfantAction, and NTU dataset under two conditions: (1) Real - utilizing the training set of each dataset and (2) Real+Syn - a hybrid approach combining both real and synthetic data samples. The highest accuracy for each model and dataset are distinguished in bold.

ity in Tab. 2. Our results show substantial improvements in both accuracy and diversity when the InfAGenC model is trained on any of these three datasets. Our model also outperforms the MDM model, confirming the effectiveness of our approach in handling limited and diverse data contexts. Even though it demonstrates higher FID on both InfActPrimitive and NTU datasets compared to the counterparts, this underscores the capability of our generated samples to span a broader distribution, affirming their sufficient variance. Given the constraints of dealing with sparse data and lacking 3D ground truth, our goal is to prevent overfitting to outliers and noise in the dataset while ensuring the diversity of samples remains accurate. This trade-off in reducing fidelity compared to other baseline models may be explained by the inclusion our sampling and selection approach considering both between and within class distance. Additionally, we qualitatively evaluated our models by visualizing the generated samples. Fig. 2 displays snapshots of the skeletal representations of the generated samples based on InfactPrimitive. These visuals demonstrate our model's capability to generate a diverse array of motion samples for each targeted class, significantly enriching the dataset and supporting advancements in research on infant motion. More visualizations of generated motions based on InfantAction and NTU data are presented in the Supplementary Materials Fig. S1 and Fig. S2.

4.3. Evaluation of Infant Action Recognition

We employed two distinct action recognition models. The first model is the ST-GCN model, identical to the one utilized in our methodology, which uses 6D rotations for joints representation. The second model, MS-G3D [23], adopts xyz coordinates for joints representation and features a disentangled aggregation approach coupled with a unified spatial-temporal graph convolution (G3D) operator to enhance action feature learning. We trained each action recognition model trained under two data configurations: (a) Real, using only the real training set and (2) Real+Syn, including both real training set and synthetic data samples generated by our InfAGenC model. The overall performance outcomes of these configurations are compared in Tab. 3. Either the ST-GCN model or the MS-G3D model, the accuracy improves significantly when synthetic data is included, with the notable increase seen in both InfActPrimitive dataset and NTU dataset, indicating that the synthetic data aids in generalizing the models' predictive capabilities. However, for the InfantAction dataset, the performance of both models trained with only the real dataset outperforms those trained with the hybrid dataset. This inconsistency may be attributed to two main factors: (1) Compared to other datasets, the InfantAction dataset is relatively small, with only 145 samples, which may not be sufficient to generate high-quality synthetic data. (2) The imbalance in class distribution could introduce a bias in evaluations, favoring classes with more samples. The addition of balanced syn-

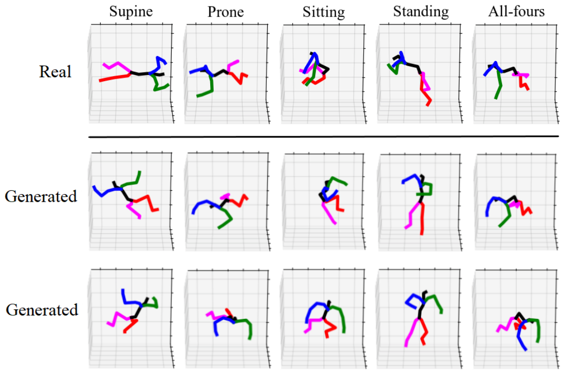
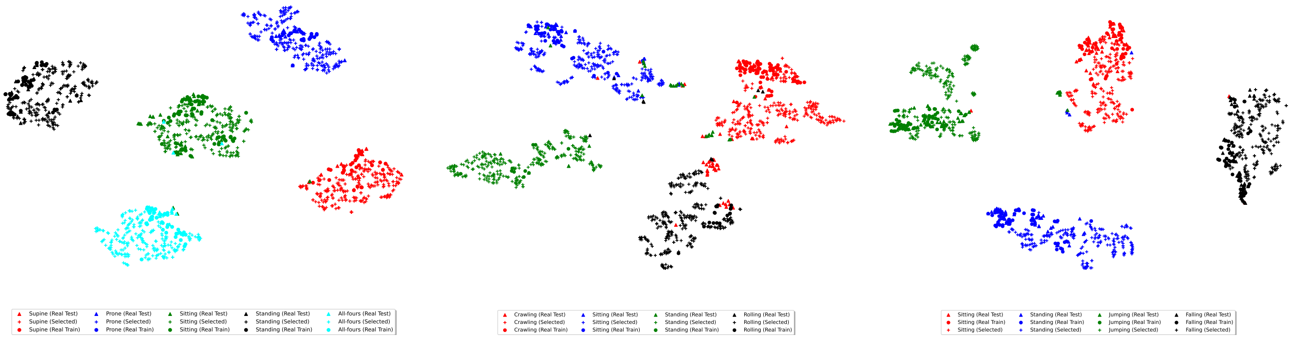


Figure 2. Snapshots of Our Generated Samples. The images are organized by action class, displayed in the following order from left to right: Supine, Prone, Sitting, Standing, and All-fours. First row is the real action samples, the other two rows present distinct motions depicted as skeletons per action class.



(a) InfactPrimitive

(b) InfantAction

(c) NTU

Figure 3. T-SNE Feature Visualization of Real and Synthetic Data. We compared the features of real train data, real test data, and synthetic data (generated during training InfAGenC model) for each experimental dataset (i.e. InfactPrimitive, InfantAction, and NTU). The data points are color-coded to distinguish between various infant actions. And data sources are shape-coded: ● for real part of train data, ▲ for real test data, and while + for selected synthetic samples during training.

		200 Test Samples Acc. (%)					3081 Test Samples Acc.(%)				
Method	Train Set	Sitting	Standing	Jumping	Falling	Overall	Sitting	Standing	Jumping	Falling	Overall
-	Real	82.00	96.00	100.00	98.00	94.00	95.33	91.16	97.78	96.64	95.73
Conditional Transform-based VAE	Syn	90.00	92.00	96.00	96.00	93.50	93.11	96.61	93.54	93.78	94.26
(Baseline)	Real+Syn	92.00	98.00	94.00	100.00	96.00	90.25	98.83	79.97	99.35	92.08
Random Selection InfAGenC	Syn	82.00	90.00	96.00	94.00	90.05	87.65	98.04	84.24	89.25	89.78
(Ours)	Real+Syn	88.00	96.00	98.00	100.00	95.50	93.50	98.43	89.66	98.83	95.10
Distance-based Selection InfAGenC	Syn	86.00	90.00	80.00	94.00	87.50	86.48	81.07	75.19	93.01	83.93
(Ours)	Real+Syn	96.00	98.00	98.00	100.00	**98.00**	95.71	97.65	95.61	99.35	**97.08**

Table 4. Ablation Analysis of ST-GCN Action Recognition Models on NTU Dataset Using Various Synthetic Training Sets. We evaluate the quality of synthetic data produced by three different configurations of generative model training: (1) a baseline conditional transform-based VAE, (2) our InfAGenC network with random sample selection, and (3) our InfAGenC network employing a distance-based sample selection strategy. The action recognition models were trained using different combinations of datasets: (a) Real, which solely utilizes the prepared NTU real training set; (b) Syn, which uses synthetic infant data generated from the corresponding generative model configuration; and (c) Real+Syn, a hybrid approach that combines both real and synthetic data. The highest accuracy for each class is underlined, and the highest overall accuracy is highlighted in bold.

thetic data diminishes the model's preference for specific classes, thereby enhancing performance across all classes, not just those that are overrepresented. To further support our hypotheses, detailed subject-wise and action-wise evaluations will be presented in the upcoming ablation studies. We also visualized each experimental dataset's feature distribution by using T-SNE algorithm in Fig. 3. It is obvious that most classes are clustered very well and those generated samples are largely scattered around the real training data. This expansion of each class's distribution effectively broadens the variety of actions covered within each class, thereby enhancing the performance of the action recognition model.

4.4. Ablation Studies

To ensure an unbiased evaluation of our methods and provide an ablation analysis, we conducted experiments on the subset of NTU data, which features a balanced class distribution, 3D ground truth, and sufficient test data, allowing us to assess our methods effectively.

Synthetic Data Recycling In Tab. 2, we evaluate the performance of our generative model in comparison with a baseline VAE model, specifically examining the quality of synthetic samples generated by both. We assessed the effectiveness of our synthetic data recycling approach by comparing the performance of ST-GCN action recognition models trained on equivalent amounts of synthetic samples produced by (1) the baseline VAE model and (2) our InfAGenC model. These models were tested on the NTU dataset's small test set of 200 samples and a large test set of 3081 samples, with results detailed in Tab. 4. The action recognition model trained with fusion data, combining synthetic samples from our InfAGenC model and real data, shows a performance increase of 4% on the small test set and 1.35% on the large test set compared to models trained only with real data. In contrast, synthetic data from the baseline model, which lacks the recycling component, only improves performance by 2% on the small test set and actually decreases performance by 3.65% on the large test set. These results highlight the significant benefits of integrating our synthetic data recycling module into the training process.

Distance-based Sample Selection We conducted a comparative experiment to assess the efficacy of our distance-based sample selection strategy for synthetic data. In contrast to selectively using high-quality, diverse synthetic samples during InfAGenC model training, we experimented

		ST-GCN Acc. (%)						MS-G3D Acc. (%)					
Subject	Train Set	Supine	Prone	Sitting	Standing	All-fours	Overall	Supine	Prone	Sitting	Standing	All-fours	Overall
D01	Real	73.91	86.67	-	-	-	83.67	**100.00**	**89.33**	-	-	-	**91.84**
	Real+Syn	**78.26**	**90.67**	-	-	-	**87.76**	**100.00**	88.00	-	-	-	90.82
D02	Real	0.00	0.00	**52.94**	28.57	59.09	46.84	**100.00**	**100.00**	48.00	14.28	31.81	35.84
	Real+Syn	**100.00**	0.00	**52.94**	**33.33**	**63.64**	**50.63**	**100.00**	0.00	**60.60**	**47.62**	**45.45**	**52.66**
D03	Real	-	0.00	**86.96**	51.35	60.00	**72.11**	-	**16.67**	68.11	27.02	32.85	51.00
	Real+Syn	-	**16.67**	80.43	**64.86**	**64.29**	**72.11**	-	0.00	**86.23**	**37.83**	**57.14**	**68.92**
D04	Real	-	55.56	-	-	-	55.56	-	**66.67**	-	-	-	**66.67**
	Real+Syn	-	**66.67**	-	-	-	**66.67**	-	22.22	-	-	-	22.22

Table 5. Detailed Performance of Infant Action Recognition Models on the InfActPrimitive Dataset. This table displays the accuracy for each subject and action class, comparing models trained across different data configurations: (1) Real - utilizing the InfActPrimitive dataset and (2) Real+Syn - a hybrid approach that merges both real and synthetic data generated by our InfAGenC model. In the analysis, entries showcasing the highest accuracy for each class and subject are distinguished in bold.

		ST-GCN Acc. (%)					MS-G3D Acc. (%)				
Subject	Train Set	Crawling	Sitting	Standing	Rolling	Overall	Crawling	Sitting	Standing	Rolling	Overall
Inf04	Real	**98.08**	-	-	22.22	**86.89**	75.00	-	-	44.44	67.18
	Real+Syn	75.00	-	-	**55.56**	72.13	**94.23**	-	-	0	**76.56**
Inf05	Real	**100.00**	**100.00**	**44.44**	-	**73.68**	83.33	58.33	16.66	-	42.93
	Real+Syn	83.33	**100.00**	37.04	-	68.42	83.33	58.33	16.66	-	42.93

Table 6. Detailed Performance of Infant Action Recognition Models on the InfantAction Dataset. This table displays the accuracy for each subject and action class, comparing models trained across different data configurations: (1) Real - utilizing the InfantAction dataset and (2) Real+Syn - a hybrid approach that merges both real and synthetic data generated by our InfAGenC model. The highest accuracy for each class and subject are distinguished in bold.

with random synthetic sample selection to augment training data. We evaluated the performance of ST-GCN action recognition models trained with two configurations: (1) using the real NTU training set combined with synthetic samples generated by our InfAGenC model employing a random selection strategy, and (2) using the real NTU training set alongside synthetic samples generated by our InfAGenC model with a distance-based selection strategy. Our findings reveal that the ST-GCN models trained with synthetic data produced via the distance-based selection consistently achieved the highest accuracy, not just on the NTU's small test set but also on the large test set, reaching accuracies of 98% and 97.08%, respectively. This underscores the advantage of the distance-based selection approach in enhancing the effectiveness of training data.

Subject-wise and Action-wise Evaluation Due to the rapid motor development in infants, the range of action shifts over time. Additionally, there is a huge within-class discrepancy in infant actions, as illustrated by [12], even for the same class of action. For better interpretability, we compared the performance outcomes of infant action recognition models for each infant subject and action class across both the InfActPrimitive and InfantAction datasets, as presented in Tab. 5 and Tab. 6, respectively. As shown in Tab. 5, across both ST-GCN and MS-G3D, the highest accuracy for most subjects and action classes is achieved with the hybrid training set. Specifically, older subjects D02 and D03, who perform more advanced actions but exhibit fewer elementary motor skills (their data class distribution is reported in Tab. S1), show significant performance improvements when trained on hybrid data compared to just real data. Similarly, in Tab. 6, although the overall accuracy of models trained solely with real data appears higher, the accuracy distribution across different action classes is more uneven when compared to models trained on hybrid data. This evidence justifies the value of integrating balanced synthetic data to enhance model performance across diverse action classes.

5. Limitation and Conclusion

In this study, we developed an infant motion generation pipeline and introduced a unique dataset with a more complex set of actions. Due to the uncooperativeness of infants, data gathered in home settings without assumptions about camera angles and views vary from infant to infant, resulting in limited samples but substantially different positions. Additionally, due to age variations, even the same action, such as standing, differs among infants (some use support, while others stand by themselves). Another limitation is that acquiring 3D pose ground truth using motion capture is not possible for infants, as they are often surrounded and occluded by toys and other objects, posing significant challenges for 3D pose estimation and tracking in videos. However, our pipeline's results on the NTU dataset, with less noise and reliable ground truth, demonstrate its capability to handle small data effectively. In future work, we plan to increase the dataset size to establish a more robust benchmark for infant action recognition tasks. We hope that the results of this study will encourage others in this field to develop their models, leading to smarter and safer environments for infants.

References

[1] Ghada Alsebayel, Mahsa Nasri, Caleb P Myers, Giovanni M Troiano, Elaheh Hatamimajoumerd, Sarah Ostadabbas, Kristen Allison, and Casper Harteveld. Articumotion: Towards assessing motor speech disorders via gamification. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*, pages 232–247, 2024.

[2] Somaieh Amraee, Bishoy Galoaa, Matthew Goodwin, Elaheh Hatamimajoumerd, and Sarah Ostadabbas. Multiple toddler tracking in indoor videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 11–20, 2024.

[3] Xu Cao, Xiaoye Li, Liya Ma, Yi Huang, Xuan Feng, Zening Chen, Hongwu Zeng, and Jianguo Cao. Aggpose: Deep aggregation vision transformer for infant pose estimation. *In Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22) Special Track on AI for Good*, 2022. 3

[4] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 3

[5] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20186–20196, 2022. 2, 3

[6] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6519–6527, 2020. 3

[7] Bruno Degardin, Joao Neves, Vasco Lopes, Joao Brito, Ehsan Yaghoubi, and Hugo Proença. Generative adversarial graph convolutional networks for human action synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1150–1159, 2022. 3

[8] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 2

[9] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Posegpt: Chatting about 3d human pose. *arXiv preprint arXiv:2311.18836*, 2023. 3

[10] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2021–2029, 2020. 1, 3

[11] Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)*, 39(4):60–1, 2020. 1

[12] Elaheh Hatamimajoumerd, Pooria Daneshvar Kakhaki, Xiaofei Huang, Lingfei Luan, Somaieh Amraee, and Sarah Ostadabbas. Challenges in video-based infant action recognition: A critical examination of the state of the art. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 21–30, 2024. 1, 3, 5, 6, 8

[13] Nikolas Hesse, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Raphael Weinberger, and A Sebastian Schroeder. Computer vision for medical infant motion analysis: State of the art and rgb-d data set. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3

[14] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J Black, Christoph Bodensteiner, Michael Arens, Ulrich G Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, et al. Learning an infant body model from rgb-d data for accurate full body motion analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 792–800. Springer, 2018. 11

[15] Xiaofei Huang, Nihang Fu, Shuangjun Liu, and Sarah Ostadabbas. Invariant representation learning for infant pose estimation with small data. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021. 1, 3

[16] Xiaofei Huang, Lingfei Luan, Elaheh Hatamimajoumerd, Michael Wan, Pooria Daneshvar Kakhaki, Rita Obeid, and Sarah Ostadabbas. Posture-based infant action recognition in the wild with very limited data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4911–4920, 2023. 1

[17] 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, jul 2014. 1

[18] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International Conference on 3D Vision (3DV)*, pages 918–927. IEEE, 2020. 3

[19] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[20] Taeryung Lee, Gyeongsik Moon, and Kyoung Mu Lee. Multiact: Long-term 3d human motion generation from multiple action labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1231–1239, 2023. 3

[21] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 1, 3, 5, 6

[22] Shuangjun Liu, Michael Wan, and Sarah Ostadabbas. Heuristic weakly supervised 3d human pose estimation. *arXiv preprint arXiv:2105.10996*, 2021. 11

[23] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020. 2, 6

[24] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, pages 417–435. Springer, 2022. 3

[25] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017. 3

[26] Lucia Migliorelli, Sara Moccia, Rocco Pietrini, Virgilio Paolo Carnielli, and Emanuele Frontoni. The babypose dataset. *Data in brief*, 33:106329, 2020. 3

[27] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 1

[28] Trevor Powers, Elaheh Hatamimajoumerd, William Chu, Vishakk Rajendran, Rishi Shah, Frank Diabour, Marc Vaillant, Richard Fletcher, and Sarah Ostadabbas. Vision-based treatment localization with limited data: Automated documentation of military emergency medical procedures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1819–1828, 2023. 2

[29] Abhinanda R Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J Black. Babel: Bodies, action and behavior with english labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 722–731, 2021. 3

[30] 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. *arXiv preprint arXiv:2303.10280*, 2023. 3

[31] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004. 1

[32] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021. 3

[33] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[34] Fatma M Talaat and Hanaa ZainEldin. An improved fire detection approach based on yolo-v8 for smart cities. *Neural Computing and Applications*, 35(28):20939–20954, 2023. 11

[35] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 1, 3

[36] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2649–2656, 2014. 1

[37] Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, Weihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, et al. Actformer: A gan-based transformer towards general action-conditioned 3d human motion generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2228–2238, 2023. 2

[38] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1, 2

[39] Xueying Zhan, Huan Liu, Qing Li, and Antoni B Chan. A comparative survey: Benchmarking for pool-based active learning. In *IJCAI*, pages 4679–4686, 2021. 1

[40] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 6

[41] Shaotong Zhu, Amal Mathew, Elaheh Hatamimajoumerd, Michael Wan, Briana Taylor, Rajagopal Venkatesaramani, and Sarah Ostadabbas. Cribnet: Enhancing infant safety in cribs through vision-based hazard detection. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 01–08. IEEE, 2024. 3

[42] Shaotong Zhu, Michael Wan, Sai Kumar Reddy Manne, Elaheh Hatamimajoumerd, Marie J Hayes, Emily Zimmerman, and Sarah Ostadabbas. Subtle signals: Video-based detection of infant non-nutritive sucking as a neurodevelopmental cue. *Computer Vision and Image Understanding*, 247:104081, 2024. 3

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 [34], 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 [22] 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 [14] 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.

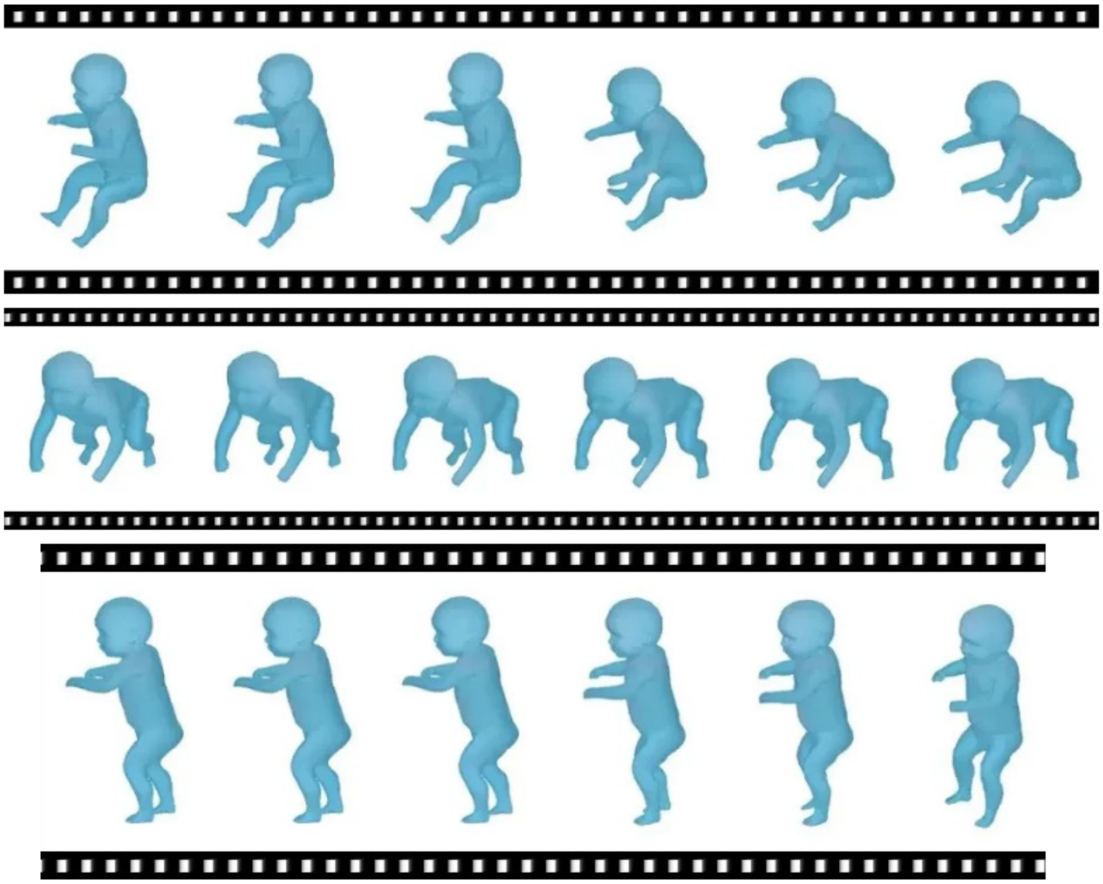


Figure S1. Sample Sequences of Generated Actions. Each sample's frames are extracted from a generated action sequence spanning 3 seconds. The actions, displayed sequentially from top to bottom, are: Sitting, Crawling, Standing.

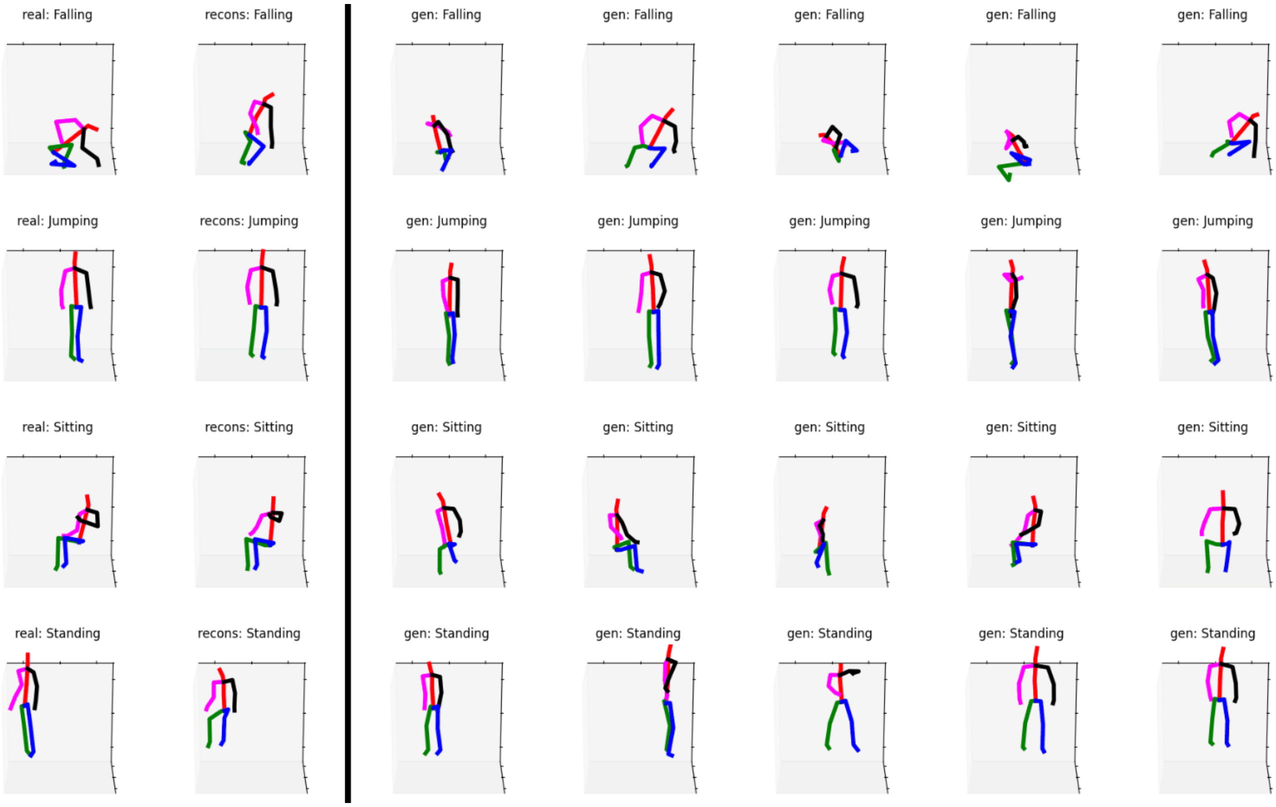


Figure S2. Snapshots of Generated Adult Action Samples. The generated samples are produced by trained our generative model on our prepared small NTU Dataset with four action classes: Falling, Jumping, Sitting, and Standing. Each row shows one action class samples.