Generative Data Augmentation for Skeleton Action Recognition Supplementary File

I. Generation Metrics

We use four complementary metrics to evaluate the quality of our generated skeleton sequences: FID and KID assess the fidelity of generated data by measuring distributional similarity to real motion sequences; Diversity captures the variability among generated motions; and Precision & Recall jointly quantify the realism and coverage of the motion distribution.

a) FID: [2] Fréchet Inception Distance (FID) evaluates the similarity between the real and generated skeleton data distributions in a learned feature space, capturing both the mean and covariance of features. In our task, a lower FID indicates that the synthetic skeleton sequences are closer in structure and temporal coherence to real motion data, reflecting better generation fidelity.

b) KID: [1] Kernel Inception Distance (KID) is an alternative to FID that measures distributional similarity using kernel-based Maximum Mean Discrepancy (MMD). Unlike FID, KID is an unbiased estimator and is more reliable under limited sample sizes, which is particularly relevant for few-shot data augmentation scenarios. A lower KID score suggests a more faithful and realistic generation.

c) Diversity: [3] Diversity quantifies the variability among generated skeleton sequences. In our context, higher diversity indicates the model can produce a broader range of motion styles and patterns, reducing redundancy and helping to prevent overfitting when the synthetic data is used to augment training sets.

d) Precision/Recall: Precision and Recall jointly assess the quality and coverage of generated samples: precision measures how many synthetic motions lie within the real data manifold, reflecting realism, while recall measures how much of the real data distribution is covered by the generator. In our setting, high precision ensures plausibility and action consistency, while high recall reflects the model's capacity to capture a wide spectrum of human actions.

II. Different augmentation ratios experiment

We evaluate the impact of different augmentation ratios (1× and 5× synthetic samples) under varying levels of real-data availability on both HumanAct12 and Refined NTU-RGBD datasets. As shown in Table II and Table I, incorporating our synthetic data consistently improves recognition accuracy across different backbones. For HumanAct12, 5× augmentation often achieves the best performance, particularly when only 75% or 90% of the real data is available. On Refined NTU-RGBD, 1× augmentation already leads to noticeable gains in many settings, while 5× augmentation further enhances performance in low-data scenarios. These findings demonstrate the effectiveness and scalability of our conditional diffusion-based data augmentation approach. However, excessive synthetic data may also overwhelm the model, causing it to overfit to the augmented distribution rather than the true data manifold. This is also our future research direction to develop adaptive strategies to automatically balance these parameters.

III. Visualisation

We present qualitative results in Fig.1 for HumanAct12 and Fig.2 for Refined NTU-RGBD to illustrate the effectiveness of our method.

References

[1] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying mmd gans. In *International Conference on Learning Representations (ICLR)*, 2021.

[2] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pages 2021–2029, 2020.

[3] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

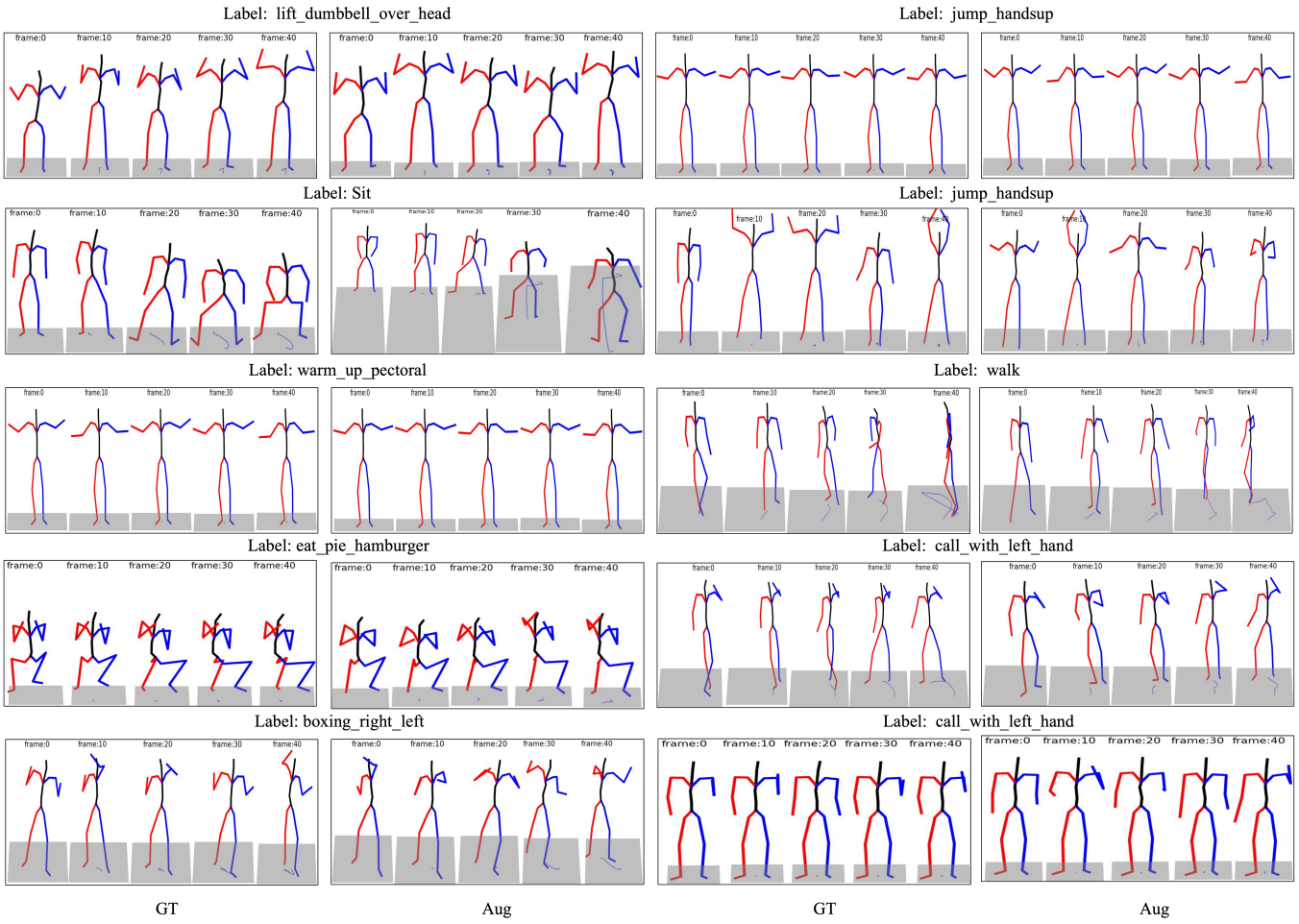

Fig. 1. Visualisation of HumanAct12 dataset.

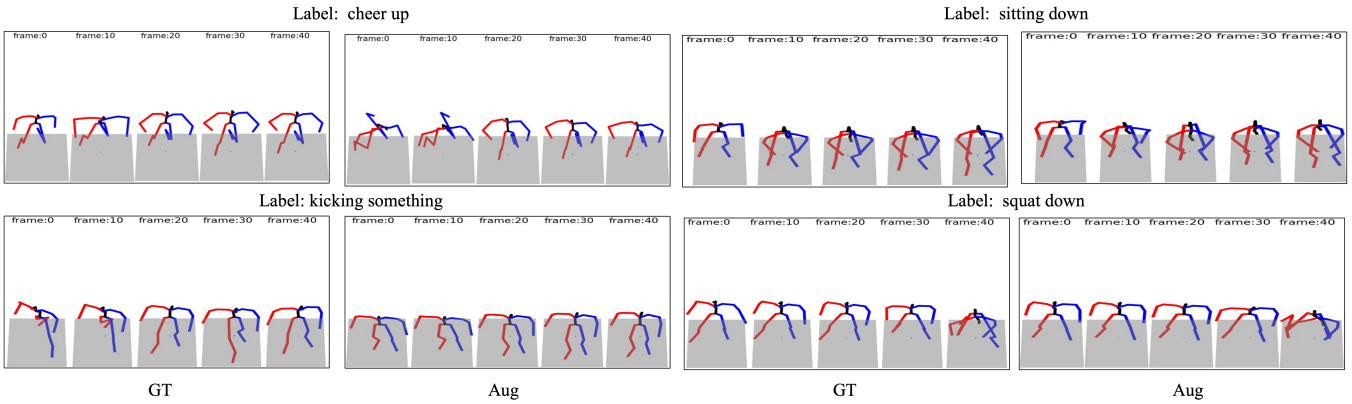

Fig. 2. Visualisation of Refined NTURGB-D dataset.

TABLE I

COMPARISON OF RECOGNITION ACCURACY (%) USING DIFFERENT AUGMENTATION RATIOS ON REFINED NTU-RGBD. "1×" AND "5×" DENOTE 1X AND 5X AUGMENTATION RESPECTIVELY.

Method	**25%**			**20%**			**15%**			**10%**		
	Base	**1×**	**5×**	**Base**	**1×**	**5×**	**Base**	**1×**	**5×**	**Base**	**1×**	**5×**
STGCN++	91.75	92.56	92.72	91.91	93.20	93.04	88.19	91.42	92.39	85.28	84.63	86.25
MSG3D	92.72	92.07	92.88	88.67	91.10	91.10	89.48	89.81	91.26	81.07	82.85	84.95
CTRGCN	91.42	90.13	91.59	91.10	91.10	91.26	85.28	88.19	88.67	77.99	84.47	85.44
BlockGCN	91.10	89.64	91.42	89.64	89.48	89.97	86.56	88.03	85.59	76.05	83.81	84.79

TABLE II

COMPARISON OF RECOGNITION ACCURACY (%) USING DIFFERENT AUGMENTATION RATIOS ON HUMANACT12. "1×" AND "5×" DENOTE 1X AND 5X AUGMENTATION RESPECTIVELY.

Method	**100%**			**95%**			**90%**			**75%**		
	Base	**1×**	**5×**	**Base**	**1×**	**5×**	**Base**	**1×**	**5×**	**Base**	**1×**	**5×**
STGCN++	75.69	77.08	79.86	77.78	79.17	79.86	75.00	79.86	80.56	73.61	76.39	80.56
MSG3D	77.08	79.17	81.94	78.47	79.86	81.94	79.86	82.64	83.33	79.86	81.94	81.25
CTRGCN	75.00	79.17	81.54	77.78	77.78	80.77	75.69	77.08	78.47	75.69	77.08	81.54
BlockGCN	76.39	80.55	79.23	77.08	79.16	81.54	72.22	75.00	78.47	75.00	77.08	77.69