# Supplemental Material of You Never Stop Dancing: Non-freezing Dance Generation via Bank-constrained Manifold Projection

## A Freezing Metric Details

**The freezing metrics for each type of music** We provide the freezing metrics for FACT and our approach, as well as Ground-Truth (GT) motions for each music type. We compute $\Delta_{\text{Pose}}$ and $\Delta_{\text{Trans}}$ for each sub-sequence in genre $i$ in the training set and then sort them in ascending order. Then we take $\Delta_{\text{Pose}}$ ranked in 10% percentile as pose threshold $\Delta^{i}_{\text{Pose}}$ and $\Delta_{\text{Trans}}$ ranked in 20% percentile as translation threshold $\Delta^{i}_{\text{Trans}}$ for each dance genre. As shown in Table 1, $\Delta_{\text{Pose}}$ and $\Delta_{\text{Trans}}$ vary in different dance genres (*e.g.*, Break and Pop). This is also the reason we use different thresholds $\Delta^{\text{gt}}$ for each genre. Finally, given each motion sub-sequence with genre $i$, if $\Delta_{\text{Pose}} \leq \Delta^{i}_{\text{Pose}}$ and $\Delta_{\text{Trans}} \leq \Delta^{i}_{\text{Trans}}$, we regard it as a freezing sub-sequence.

As shown in Table 2, our approach performs better than FACT for all dance genres. Note that GT motions from the test set have large freezing rate on two genres of Street Jazz and Lock. We visually check the motions and find that it is because they happen to have many stationary poses during adjacent dance movements.

Table 1: Details about GT motion for freezing metrics.

| Genre | GT (Train & Test) $\Delta_{\text{Pose}} \uparrow$ | $\Delta_{\text{Trans}} \uparrow$ | Freeze $\downarrow$ | GT (Only Test) $\Delta_{\text{Pose}} \uparrow$ | $\Delta_{\text{Trans}} \uparrow$ | Freeze $\downarrow$ |
|---|---|---|---|---|---|---|
| Break | 5.43 | 1.94 | 10.8% | 3.43 | 1.55 | 18.2% |
| House | 4.10 | 1.88 | 10.7% | 3.92 | 1.65 | 35.7% |
| Ballet Jazz | 5.40 | 2.14 | 9.8% | 5.60 | 2.08 | 0.0% |
| Street Jazz | 1.43 | 0.79 | 10.6% | 0.33 | 0.13 | 43.8% * |
| Krump | 3.75 | 1.30 | 10.2% | 2.09 | 0.90 | 11.1% |
| LA style Hip-hop | 3.62 | 1.31 | 9.8% | 4.12 | 1.85 | 0.0% |
| Lock | 3.62 | 1.10 | 11.3% | 0.76 | 0.17 | 77.8% * |
| Middle Hip-hop | 5.19 | 1.90 | 9.8% | 6.11 | 1.73 | 0.0% |
| Pop | 1.91 | 0.69 | 9.8% | 1.94 | 0.76 | 0.0% |
| Waack | 4.61 | 1.03 | 9.7% | 4.50 | 0.79 | 2.3% |
| Total | 3.90 | 1.41 | 10.3% | 3.28 | 1.16 | 18.7% |

**More discussion on freezing determination** In our implementation, we regard a sub-sequence as a freezing sub-sequence when $\Delta_{\text{Pose}} \leq \Delta^{i}_{\text{Pose}}$ and $\Delta_{\text{Trans}} \leq \Delta^{i}_{\text{Trans}}$. We also tried other alternatives including only depending on pose changes or translation changes, respectively. We present results under these two settings in Figure 1 and Figure 2. We can see that using only one threshold (pose or translation) is not suitable to determine freezing situation. So we choose to use both. We also provide freezing rate statics under the setting with 10% percentile pose threshold and different translation thresholds in Figure 3.

## B Supplemental Ablation Results

**Genre-agnostic vs. genre-specific** In our implementation, we perform manifold learning for each dance genre separately (i.e., our manifold bank is constructed in an genre-specific manner). We can

Table 2: Details about generated motion for freezing metrics on the AIST++ test set.

| Genre | FACT | | | Ours | | |
|---|---|---|---|---|---|---|
| | $\Delta_{\text{Pose}}\uparrow$ | $\Delta_{\text{Trans}}\uparrow$ | Freeze $\downarrow$ | $\Delta_{\text{Pose}}\uparrow$ | $\Delta_{\text{Trans}}\uparrow$ | Freeze $\downarrow$ |
| Break | 0.81 | 0.63 | 85.0% | 1.54 | 1.27 | 68.8% |
| House | 1.08 | 1.03 | 73.8% | 1.21 | 1.18 | 70.0% |
| Ballet Jazz | 3.19 | 1.95 | 2.5% | 3.47 | 2.35 | 0.0% |
| Street Jazz | 1.28 | 1.14 | 0.0% | 1.60 | 1.45 | 0.0% |
| Krump | 1.00 | 1.03 | 42.5% | 1.15 | 1.05 | 37.5% |
| LA style Hip-hop | 1.93 | 1.27 | 30.0% | 2.04 | 1.36 | 22.5% |
| Lock | 0.97 | 0.89 | 27.5% | 1.43 | 1.53 | 1.3% |
| Middle Hip-hop | 1.34 | 1.20 | 77.5% | 1.58 | 1.24 | 71.3% |
| Pop | 0.64 | 0.64 | 30.0% | 1.22 | 1.10 | 0.0% |
| Waack | 1.05 | 0.87 | 27.5% | 1.20 | 1.09 | 18.8% |
| Total | 1.33 | 1.07 | 39.0% | 1.64 | 1.36 | 29.6% |

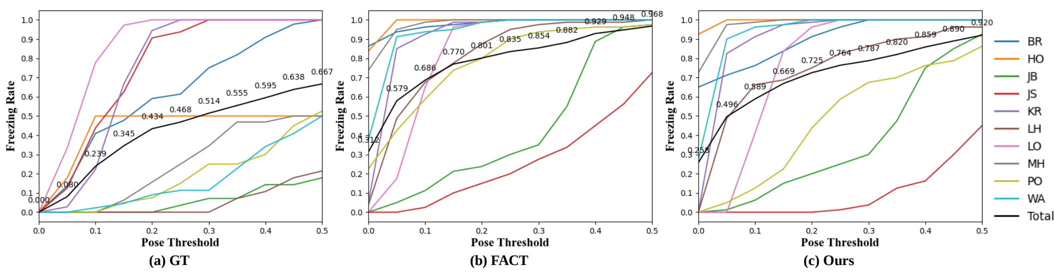


Figure 1: Freezing rate with only pose threshold.

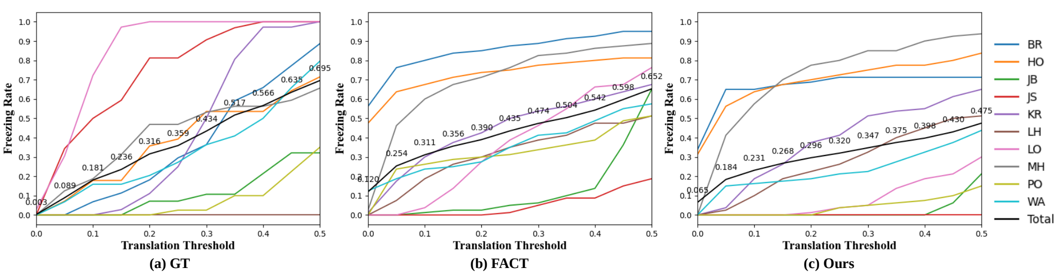


Figure 2: Freezing rate with only translation threshold.

also construct a genre-agnostic manifold bank by training over all the GT motion segments without considering the dance genres. As shown in Table 3, our approach with the genre-agnostic bank can already bring consistent improvement over the baseline. And our proposed genre-specific bank further promotes the performance, which demonstrates the effectiveness of exploring the action-specific context for dance generation.

**Stage-wise training** We train our model in three stages to ensure that the learned manifold bank can accurately reconstruct the GT motions. We can also adopt an end-to-end training strategy and detailed results are presented in the table 4. As shown, stage-wise training strategy has clear advantages in terms of all metrics. This is because bank elements will be learned from the predicted noisy motions during end-to-end training.

## C Licenses of Referenced Assets

We provide the links pointing to the licenses of our referenced assets, including employed models and datasets.

**SMPL model [3]** https://smpl.is.tue.mpg.de/modellicense.html

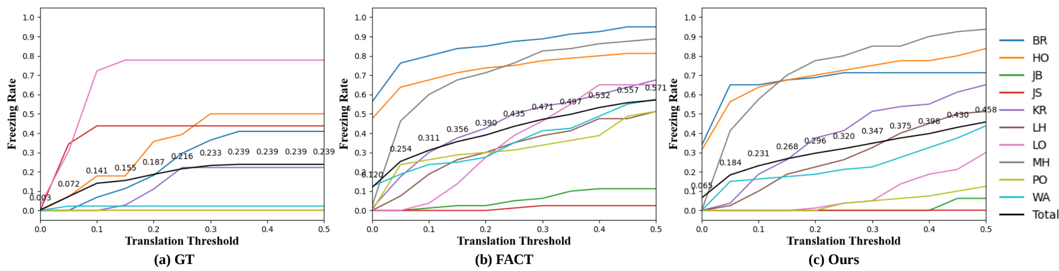


Figure 3: Freezing rate with both pose threshold and translation threshold. We take the pose threshold as 10% percentile for each dance genre.

Table 3: Evaluation of our genre-specific bank.

| Method | Quality | | | | | Diversity | | Align |
|---|---|---|---|---|---|---|---|---|
| | $\text{FID}_k \downarrow$ | $\text{FID}_g \downarrow$ | $\Delta_{\text{Pose}} \uparrow$ | $\Delta_{\text{Trans}} \uparrow$ | Freezing $\downarrow$ | $\text{Dist}_k \uparrow$ | $\text{Dist}_g \uparrow$ | BeatAlign $\uparrow$ |
| Baseline | 35.35 | 22.11 | 1.33 | 1.07 | 39.0% | 5.94 | 6.18 | 0.241 |
| Genre-agnostic Bank | 28.57 | 15.92 | 1.58 | 1.33 | 31.9% | 7.29 | 6.42 | 0.247 |
| Genre-specific Bank | **25.96** | **13.42** | **1.64** | **1.36** | **29.6%** | **7.68** | **6.59** | **0.249** |

**FACT model [2]** https://github.com/google-research/mint/blob/main/LICENSE

**AIST++ dataset [2]** https://google.github.io/aistplusplus_dataset/factsfigures.html

**AIST dataset [4]** https://aistdancedb.ongaaccel.jp/terms_of_use/

**Mixamo dataset [1]** https://www.adobe.com/legal/licenses-terms.html

# References

[1] Adobe. Adobe mixamo dataset, 2017.

[2] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021.

[3] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[4] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, Nov. 2019.

Table 4: Evaluation of stage-wise training.

| Method | Quality | | | | | Diversity | | Align |
|---|---|---|---|---|---|---|---|---|
| | $\mathrm{FID}_k \downarrow$ | $\mathrm{FID}_g \downarrow$ | $\Delta_{\mathrm{Pose}} \uparrow$ | $\Delta_{\mathrm{Trans}} \uparrow$ | Freezing $\downarrow$ | $\mathrm{Dist}_k \uparrow$ | $\mathrm{Dist}_g \uparrow$ | BeatAlign $\uparrow$ |
| Without Stage-wise Training | 29.37 | 16.75 | 1.52 | 1.29 | 33.4% | 6.73 | 6.39 | 0.246 |
| With Stage-wise Training | **25.96** | **13.42** | **1.64** | **1.36** | **29.6%** | **7.68** | **6.59** | **0.249** |

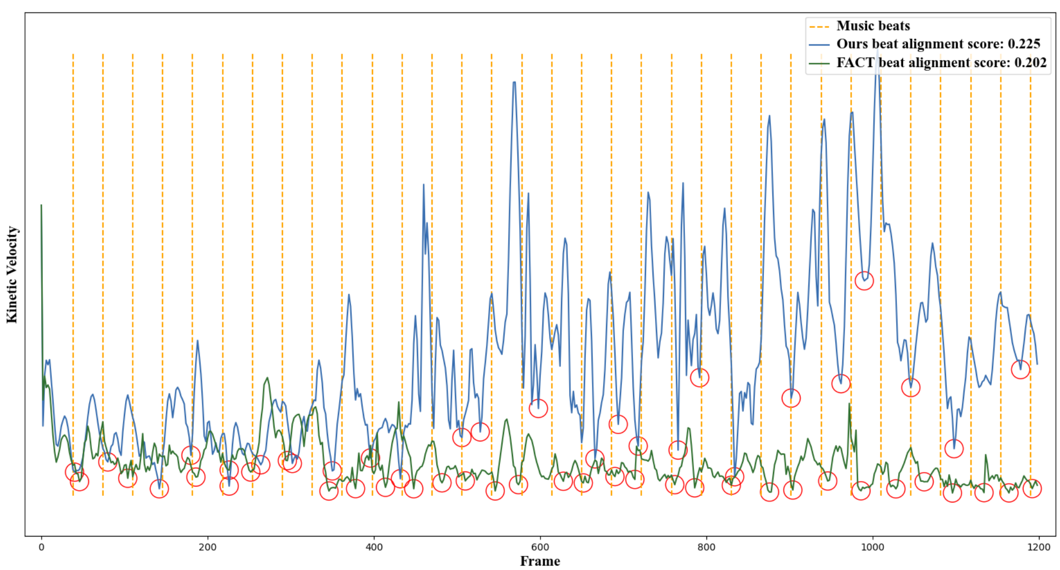


Figure 4: Beats alignment between music and dances generated by FACT and our method for the same music. The red circles are kinematic beats and dash lines denote the musical beats. The kinematic beats are extracted by finding local minima from the kinetic velocity curve. This picture is a supplement to Figure 5 in the main paper.