Reframing Music-Driven 2D Dance Pose Generation as Multi-Channel Image Generation

Supplementary Material

1. Mask Construction Details (Reference Conditioning)

In our setup, the VAE uses a temporal stride of 8 (i.e., $T' = T/8$). Hence, one latent time column corresponds to 8 consecutive frames in the original sequence. We define a binary mask

$$M \in \{0,1\}^{8 \times W' \times T'}$$

whose first dimension indexes the *frame-within-chunk* (phase). Each entry $M[\phi, x, t]$ specifies whether the latent at phase ϕ and latent time index t should be treated as *pose-aware* ($M = 1$) or *shape-only* ($M = 0$). Latents with $M = 1$ correspond to replaced reference frames where the model conditions on the actual 2D pose, whereas latents with $M = 0$ only provide shape-only context (e.g., shoulder width, limb lengths).

Given the number of replaced (pose-aware) reference frames $N \in \{0, 1, \dots, T\}$, let

$$q = \lfloor N/8 \rfloor, \qquad r = N \bmod 8.$$

The mask M is then constructed as

(i) No replacement: $N = 0 \;\Rightarrow\; M = \mathbf{0}$ (all shape-only).

(ii) Full latent columns (pose-aware): $M[:, :, 0{:}q] = 1.$

(iii) Partial column (if $r > 0$, pose-aware phases):

$M[0{:}r, :, q] = 1.$

Intuitively, if N frames are replaced by pose-aware references, the first q latent time columns are fully pose-aware, while the remainder r indicates that, in the next column, only the first r phases (corresponding to the first r frames of that 8-frame chunk) are pose-aware and the remaining phases in that column remain shape-only. If $N \geq T$, we clamp q and r to valid ranges (i.e., not exceeding T' and 8, respectively). When $N = 0$, M is all zeros and the model conditions purely on shape-only references across the entire sequence.

2. In-the-Wild-Train Dataset Construction

We collect ∼30K in-the-wild dance videos from the web and apply the following preprocessing pipeline. First, we extract 2D poses for every frame using DW-Pose [3]. Then, we perform *shot-change filtering*: we compute inter-frame pose differences and treat frames whose difference exceeds a threshold as shot boundaries; only contiguous subsequences of at least 256 frames are kept. We also apply a frame-rate sanity check, discarding videos whose native FPS is below 20 or above 60. After these filtering steps, we obtain ≈*600* hours and *240k* training segments.

3. 2D Pose-Space Metric Definitions

We adapt the commonly used FID, DIV, and BAS metrics to operate entirely in the 2D pose space. For FID, we compute the Fréchet Inception Distance [1] between the distribution of kinetic features [2] extracted from generated dances and that of the test-set dances. In the original formulation, these kinetic features are computed from 3D joint trajectories; in our setting, we recompute the same features from 2D keypoints so that FID is evaluated purely in 2D pose space. For DIV, consistent with FACT and Bailando, we measure motion diversity as the average pairwise distance between generated sequences in the kinetic-feature space. Again, the kinetic features are computed from 2D joint trajectories instead of 3D motions, yielding a diversity metric defined in 2D pose space. The Beat Align Score (BAS) is defined as the average temporal distance between each music beat and its closest dancing beat, where dancing beats are detected from the 2D pose sequence.

4. Baseline Adaptation Details

We retrain EDGE, LodGE, and Bailando on our 2D pose datasets using their publicly available code and recommended hyperparameters, making only the minimal changes described below. The original EDGE model predicts 3D motion parameters supervised by a 3D motion loss, a 3D joint-position loss, a 3D joint-velocity loss and a contact-consistency loss. In our 2D setting, we keep the joint-position and joint-velocity objectives but define them on 2D keypoints, replacing the 3D joint-position loss with a 2D keypoint-position loss and the 3D joint-velocity loss with a 2D keypoint-velocity loss. The 3D motion loss and the contact-consistency loss are removed because they cannot be computed from 2D keypoints. LodGE is modified in the same way. We retain the joint- and velocity-based losses and apply them to 2D keypoint coordinates, and we remove the 3D motion loss and the contact loss that explicitly depend on full 3D motion and contact patterns. For Bailando, which follows a multi-stage training pipeline, we modify two stages. In the VQ-VAE stage, we replace the original 3D keypoints with 2D keypoints as the reconstruction target while keeping the architecture unchanged. In the actor–critic learning stage, we omit this component entirely, because its reward function depends on 3D joint angles that are unavailable in the 2D setting. Because none of the original three implementations provides a dedicated mechanism for handling invisible keypoints, we impute missing joints by temporal interpolation from neighboring frames, and discard sequences in which a large fraction of joints remains invisible.

5. Supplementary Videos

We provide additional qualitative results in the attached video folder accompanying this PDF. All videos are generated from either the leakage-free in-the-wild test set or the AIST++2D benchmark using the same settings as in the main paper, and are rendered either as skeleton animations or via the fixed pose-to-video renderer. Please watch the videos **with audio enabled**. The dance quality is best judged together with the music rhythm.

S0 2Dvs3D-#.mp4. Side-by-side comparisons between our 2D in-the-wild model and representative 3D-based music-to-dance methods under the data-regime setting in Table 2. Each clip shows skeleton renderings on the leakage-free test songs, illustrating the generalization gap between 3D-trained models and our 2D-trained generator.

S1 tempo-shift.mp4. An example with an abrupt tempo and energy change in the music. Our model promptly changes motion amplitude and cadence at the change point, whereas competing methods tend to continue low-energy or overly uniform motions, as discussed in Sec. 4.1.

S2 tempo-scaling-fast.mp4 and S2 tempo-scaling-slow.mp4. Covers of the same song at different tempi. These clips show that our generated dances naturally scale step rate and per-beat motion extent with the audio tempo.

S3 diversity-seeds.mp4. Multiple dances generated for the same song by sampling different diffusion noise seeds. The sequences demonstrate stochastic diversity in step patterns, amplitudes, and phrasing, while maintaining beat alignment.

S4 camera-motion.mp4. Examples with camera zoom/push. Apparent subject scale changes in the videos are mirrored smoothly in the generated pose without breaking rhythm.

S5 more-#.mp4. Additional qualitative results on the leakage-free test set, complementing the quantitative comparisons in Table 1.

S6 hand-#.mp4. Qualitative results for the hand-aware variant (Ours-hand). We visualize both full-body and hand skeletons together with the corresponding pose-to-video renders.

S7 ablation-1/2/3.mp4. Ablation videos corresponding to the studies in Sec. 4.2:

- **S7 ablation-1.mp4**: raw 2D coordinates vs. one-hot pose representation (Table 4), highlighting reduced jitter and improved stability for the one-hot variant.
- **S7 ablation-2.mp4**: without–time-shared vs. time-shared positional indexing (Table 5) on a slow→fast music concatenation; our indexing scheme switches choreography at the splice, whereas the baseline persists in the slow style.
- **S7 ablation-3.mp4**: with vs. without reference conditioning (Table 6) on long sequences, showing that removing reference conditioning leads to visible discontinuities at segment boundaries, while our model produces smoother transitions.

References

[1] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017. 1

[2] Kensuke Onuma, Christos Faloutsos, and Jessica K. Hodgins. Fmdistance: A fast and effective distance function for motion capture data. In *Eurographics*, 2008. 1

[3] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 1